Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Masquage des unités primaires d'échantillonnage (UPE) et estimation de la variance dans les enquêtes complexes

par Inho Park

Décembre 2008

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 203-216
Statistique Canada, N° 12-001-X au catalogue

Masquage des UPE et estimation de la variance dans les enquêtes complexes

Inho Park [1]

Résumé

L'analyse des données recueillies auprès d'un échantillon stratifié à plusieurs degrés requiert de l'information sur le plan de sondage, telle que les identificateurs de strate et d'unité primaire d'échantillonnage (UPE), ou les poids de rééchantillonnage connexes, pour l'estimation de la variance. Dans certains fichiers de données à grande diffusion, l'information sur le plan de sondage est masquée en vue d'éviter le risque de divulgation, tout en permettant à l'utilisateur d'obtenir des estimations valides des variances. Par exemple, dans le cas des enquêtes aréolaires comptant un nombre limité d'UPE, les UPE originales sont divisées et (ou) recombinées pour construire des pseudo-UPE dans lesquelles sont permutées les unités d'échantillonnage de deuxième degré et de degré subséquent. Cependant, ces méthodes de masquage des UPE faussent manifestement la structure de mise en grappes du plan d'échantillonnage, ce qui donne des estimations de variance biaisées pouvant présenter un rapport systématique entre les deux estimations de variance obtenues avec et sans masquage des identificateurs d'UPE. Certains travaux antérieurs ont révélé certaines tendances du ratio des estimations de la variance obtenues avec et sans masquage si on représente ce ratio graphiquement en fonction de l'effet de plan sans masquage. Le présent article traite de l'effet du masquage des UPE sur les estimations de la variance sous échantillonnage en grappes en fonction de divers aspects, dont la structure de mise en grappes et le degré de masquage. En outre, nous tâchons d'établir une stratégie de masquage des UPE par permutation des unités d'échantillonnage du degré subséquent qui réduit le biais résultant des estimations de la variance. En guise d'illustration, nous utilisons des données provenant de la National Health Interview Survey (NHIS) auxquelles nous avons apporté certaines modifications artificielles. La stratégie proposée permet de bien réduire le biais des estimations de la variance. Les résultats tant théoriques qu'empiriques indiquent que l'effet du masquage des UPE sur les estimations de la variance est modeste si la permutation des unités d'échantillonnage de degré subséquent est minimale. Nous avons appliqué la stratégie de masquage proposée aux données diffusées de la National Health and Nutrition Examination Survey (NHANES) de 2003-2004.

Mots clés : Contrôle de la divulgation ; échantillonnage stratifié à plusieurs degrés ; permutation des unités d'échantillonnage de degré subséquent ; effet de plan ; coefficient de corrélation intragrappe (CCI) ; moyenne d'échantillon.

1. Introduction

L'analyse des données recueillies auprès d'un échantillon stratifié à plusieurs degrés requiert l'utilisation d'information sur le plan, telle que les identificateurs de strate et d'unité primaire d'échantillonnage (UPE), ou les poids de rééchantillonnage connexes, pour l'estimation de la variance. Dans les grandes enquêtes, les UPE correspondent souvent à un ou à plusieurs comtés. Certaines sources externes mises à la disposition du public, telles que les données du recensement, peuvent fournir des renseignements démographiques très détaillés au niveau de l'UPE. Même si l'on supprime leur nom, l'inclusion des identificateurs d'UPE dans les fichiers de données à grande diffusion peut à elle seule poser un risque d'identification en permettant leur couplage à des données de source externe. Par conséquent, les identificateurs d'UPE sont souvent masqués en vue de 1) réduire le risque de divulgation des données et 2) permettre néanmoins à l'utilisateur d'obtenir des estimations de variance valides. Mayda, Mohl et Tambay (1996) abordent la question du risque éventuel de divulgation des données associé à l'inclusion des identificateurs originaux d'UPE dans les fichiers de données à grande diffusion et considèrent la méthode de regroupement de strates proposée par Rust (1986) afin de trouver un juste équilibre entre les deux exigences susmentionnées. Étant donné la non-convergence éventuelle de l'estimation de la variance sous la méthode de regroupement de strates signalée par Valliant (1996), Yung (1997) a proposé de construire un ensemble de poids de rééchantillonnage bootstrap moyens. Lu (2004) a démontré que fournir des poids de rééchantillonnage et donner les identificateurs de strate et d'UPE sont des options à peu près équivalentes en ce qui concerne la confidentialité, puisqu'il est facile d'obtenir les uns à partir des autres. Shah (2001) a examiné divers moyens de créer des pseudo-strates et des pseudo-UPE étant donné un ensemble de poids obtenus par la méthode des répliques répétées équilibrées. Eltinge (1999) a proposé une méthode semblable à celle du regroupement des strates. Lu, Brick et Sitter (2006) ont également établi les conditions de convergence de l'estimateur de la variance sous la méthode de regroupement des strates et ont proposé des algorithmes de regroupement de strates produisant des estimateurs de variance efficaces et convergents.

1. Inho Park, statisticien, Département des statistiques économiques, Banque de la Corée, Namdaemun-Ro 106, Jung-Gu, Séoul 100-794, Corée. Courriel : ipark@bok.or.kr.

Quand le nombre d'UPE dans l'échantillon est limité, la méthode de regroupement des strates n'est pas intéressante à cause du nombre insuffisant de degrés de liberté pour l'estimation de la variance. Dohrmann, Curtin, Mohadjer, Montaquila et Le (2002), ainsi que Dohrmann, Lu, Park, Sitter et Curtin (2005) ont abordé ce genre de situation et considéré deux méthodes de masquage des UPE. La première consiste à diviser chaque UPE en deux pseudo-UPE (ensembles d'unités finales d'échantillonnage dans les UPE), en doublant arbitrairement le nombre de degrés de liberté pour l'estimation de la variance. La deuxième méthode consiste à construire les pseudo-UPE par permutation des unités d'échantillonnage de deuxième degré (USE, unités secondaires d'échantillonnage) entre les UPE originales, en gardant les nombres originaux de degrés de liberté pour l'estimation de la variance. Autrement dit, les affectations aux UPE et aux strates de toutes les unités finales d'échantillonnage comprises dans une USE sont remplacées par celles dans l'USE appariée. Cette méthode peut être généralisée de manière que les UPE originales soient divisées en une ou plusieurs fractions qui sont ensuite recombinées pour construire des pseudo-UPE dans lesquelles les fractions d'UPE sont permutées. Cette approche diffère de la permutation des données (Dalenius et Reiss, 1982), qui est souvent utilisée pour empêcher la divulgation grâce à la permutation des valeurs des variables d'enquête de nature délicate entre enregistrements individuels. À cause de la distorsion résultante de la structure de mise en grappes du plan d'échantillonnage, les deux méthodes de masquage des UPE peuvent donner des estimations de la variance biaisées présentant éventuellement un rapport systématique entre les deux estimations de la variance obtenues avec et sans masquage des identificateurs d'UPE. Dohrmann et coll. (2005) ont observé des courbes en forme d'entonnoir décroissantes pour le ratio des estimations des variances avec et sans masquage d'une moyenne d'échantillon, quand ces estimations de variance sont représentées graphiquement en fonction de l'effet de plan. Ils ont expliqué ces courbes en se basant sur une relation approximative de l'estimation de la variance, monotone en le coefficient de corrélation intragrappe (CCI) de la formule de l'effet de plan de Kish.

Dans le présent article, nous nous penchons sur les questions soulevées par la deuxième méthode de masquage des UPE qui consiste à permuter les unités d'échantillonnage de degré subséquent entre les UPE originales et nous examinons les effets de cette méthode sur les estimations de la variance. À la section 2, nous étudions l'effet du masquage des UPE sur la variance en ce qui concerne certains aspects de la structure de mise en grappes, comme le CCI et les moyennes et les tailles des fractions d'UPE pour la permutation sous un plan d'échantillonnage en grappes à un degré. À la section 3, nous examinons la relation entre le degré de permutation pour le masquage des UPE et le biais dans la variance en utilisant un modèle paramétrique pour l'échantillonnage en grappes. À la section 4, nous envisageons une stratégie de masquage des UPE par permutation des USE qui aide à réduire les effets du masquage des UPE sur les estimations de la variance et, donc, les biais résultants dans le cas des enquêtes complexes. À la section 5, nous examinons brièvement les travaux récents de Dohrmann et coll. (2002, 2005) et nous présentons les résultats de l'application de la stratégie de masquage proposée à des données provenant de la National Health Interview Survey (NHIS) en y apportant certaines modifications artificielles. Enfin, à la section 6, nous présentons une discussion.

2. Effet de la distorsion de la structure de mise en grappes sur la variance de la moyenne d'échantillon

L'échantillonnage en grappes, souvent utilisé dans les enquêtes pour des raisons économiques et logistiques, est une source importante d'accroissement de la variance d'un estimateur comparativement à un échantillonnage aléatoire simple qui tient à la similarité des unités d'échantillonnage dans les grappes. Les traités classiques d'échantillonnage, tels que Särndal, Swensson et Wretman (1992, section 8.7) fournissent des formules pour calculer la variance d'une moyenne d'échantillon en fonction du CCI, des tailles de grappe et des moyennes d'une variable étudiée y. Cela signifie que l'effet de la mise en grappes du plan d'échantillonnage sur la variance devrait être révélé par l'entremise de ces éléments. À la présente section, nous examinons comment la distorsion de la structure de mise en grappes du plan d'échantillonnage influe sur la variance d'une moyenne d'échantillon quand les UPE sont masquées par la permutation de leurs fractions entre deux UPE. Pour les besoins de la discussion présentée dans cette section, nous considérons un échantillonnage des UPE à un degré avec probabilité proportionnelle à la taille (PPT). Ce plan d'échantillonnage est relativement simple, mais néanmoins suffisamment complexe pour révéler l'effet du masquage des UPE sur la variance en ce qui a trait aux trois éléments susmentionnés.

2.1 Variance sous échantillonnage en grappes PPT à un degré

Supposons qu'une population U de M unités est groupée en N UPE de M_i unités chacune. Nous tirons un échantillon aléatoire de n UPE avec les probabilités p_i ($\sum_{i=1}^{N} p_i = 1$) et nous incluons dans l'échantillon toutes les unités comprises dans une UPE échantillonnée. Pour

simplifier, nous supposons que la sélection des UPE est faite avec remise. La moyenne d'échantillon pondérée $\hat{\bar{Y}} = (\sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij})^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij}\, y_{ij}$ est un estimateur de la moyenne de population $\bar{Y} = M^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$ de la variable étudiée y, où $w_{ij} = (np_i)^{-1}$ et y_{ij} désignent le poids d'échantillonnage et la valeur de y pour la j^{e} unité de l'UPE i, respectivement. Soit $m = \sum_{i=1}^{n} M_i$, $S_y^2 = (M-1)^{-1}$ $\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y})^2$, $\bar{Y}_i = M_i^{-1}\sum_{j=1}^{M_i} y_{ij}$ la taille d'échantillon, la variance de population et les moyennes de y dans les UPE, respectivement. En supposant que N est grand de sorte que $N/(N-1) \doteq 1$, sa variance approximative peut s'écrire

$$
\begin{aligned}
V(\hat{\bar{Y}} \mid S) &\doteq m^{-1} S_y^2 [1 + (\bar{M} - 1)\rho_{yU}] \\
&\quad + (mN)^{-1}\sum_{i=1}^{N} p_i^{-1} M_i (M_i M^{-1} - p_i)(\bar{Y}_i - \bar{Y}), \\
&\doteq (mMN)^{-1}\sum_{i=1}^{N} p_i^{-1} M_i^2 (\bar{Y}_i - \bar{Y})^2, \qquad (1)
\end{aligned}
$$

où S désigne l'ensemble d'indices dans l'échantillon, $\rho_{yU} = 1 - S_{yw}^2/S_y^2$ est le CCI et $S_{yw}^2 = (M-N)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2$ est l'écart quadratique moyen dans les UPE. La preuve de (1) est donnée en annexe.

Pour un cas particulier commun de $p_i \propto M_i$, c'est-à-dire l'échantillonnage PPT, (1) se simplifie en

$$
\begin{aligned}
V(\hat{\bar{Y}} \mid S) &\doteq m^{-1} S_y^2 [1 + (\bar{M} - 1)\rho_{yU}] \\
&\doteq (mN)^{-1}\sum_{i=1}^{N} M_i (\bar{Y}_i - \bar{Y})^2 \qquad (2)
\end{aligned}
$$

et le CCI est exprimé par

$$
\rho_{yU} \doteq (\bar{M} - 1)^{-1}\left[\frac{V(\hat{\bar{Y}} \mid S) - m^{-1} S_y^2}{m^{-1} S_y^2}\right]. \qquad (3)
$$

Dans (2), la deuxième approximation indique que les UPE pour lesquelles le terme $M_i(\bar{Y}_i - \bar{Y})^2$ est grand contribuent davantage à la variance. Dans (3), le CCI révèle (approximativement) la perte de précision due à l'accroissement relatif par grappe de la variance de $m^{-1}S_y^2$, la variance de la simple moyenne d'échantillon $\bar{y} = m^{-1}\sum_{(ij)\in S} y_{ij}$ qui aurait pu être obtenue à partir des échantillons aléatoires simples avec remise de même taille.

Une enquête complexe comporte souvent l'échantillonnage PPT à un degré susmentionné ou un autre plan complexe (additionnel) (par exemple, stratification, échantillonnage à plusieurs degrés et probabilités de sélection inégales), ou des caractéristiques d'estimation (par exemple, corrections pour la non-réponse et corrections par calage). Par exemple, si p_i n'est pas proportionnelle à la taille ou qu'on a procédé à un sous-échantillonnage poussé qui aurait pour résultat d'obtenir des poids de sondage inégaux, la caractéristique complexe correspondante pourrait joué un rôle dans l'estimation de la variance. L'incidence connexe sur l'estimation de la variance dans le cas d'échantillons complexes sera discutée en détail à la section 4.

2.2 Moyennes et tailles des fractions d'UPE

Afin de masquer les UPE, considérons que les deux premières UPE dans l'échantillon sont divisées chacune en deux ensembles d'unités, $U_1 = U_{11} \cup U_{12}$ et $U_2 = U_{21} \cup U_{22}$, disons, et que les deux pseudo-UPE, $U_1^* = U_{11} \cup U_{22}$ et $U_2^* = U_{21} \cup U_{12}$, sont construites par permutation de U_{12} et U_{22} entre les deux UPE. Soit S et S^* les ensembles d'indices dans l'échantillon sans masquage et avec masquage, respectivement. Soit $V(\hat{\bar{Y}} \mid S^*)$ la variance de $\hat{\bar{Y}}$ associée aux pseudo-UPE (en supposant aussi que les UPE non échantillonnées dans U restent les mêmes). En outre, soit $\bar{Y}_i^*$, Y_i^*, M_i^* la moyenne et le total de y et la taille de la i^{e} pseudo-UPE, respectivement. En supposant que $\bar{Y} = 0$ sans perte de généralité, la différence entre les variances avec et sans masquage s'écrit, en partant de (2), sous la forme

$$
\begin{aligned}
&V(\hat{\bar{Y}} \mid S^*) - V(\hat{\bar{Y}} \mid S) \\
&\doteq (mN)^{-1}\sum_{i=1,2}(M_i^* \bar{Y}_i^{*2} - M_i \bar{Y}_i^2), \\
&\doteq (mN)^{-1}\sum_{i=1,2}\left(\sqrt{M_i^*}\,\bar{Y}_i^* + \sqrt{M_i}\,\bar{Y}_i\right)\left(\sqrt{M_i^*}\,\bar{Y}_i^* - \sqrt{M_i}\,\bar{Y}_i\right). \qquad (4)
\end{aligned}
$$

L'expression (4) montre que la différence de variance due au masquage des UPE dépend des variations des quantités d'UPE $\sqrt{M_i}\bar{Y}_i$. Si Y_{il} et M_{il} désignent le total et la taille de la fraction d'UPE U_{il}, respectivement, pour i, l = 1, 2, alors $Y_i = Y_{i1} + Y_{i2}$, $Y_i^* = Y_{i1} + Y_{i'2}$, $M_i = M_{i1} + M_{i2}$ et $M_i^* = M_{i1} + M_{i'2}$ pour $i \neq i' = 1, 2$. Il découle clairement de (4) que la variance ne variera pas sous masquage des UPE si la condition suivante est vérifiée :

$$
\sqrt{M_i}\bar{Y}_i = \sqrt{M_i^*}\bar{Y}_i^* \quad \text{ou} \quad \frac{Y_{i1} + Y_{i2}}{\sqrt{M_{i1} + M_{i2}}} = \frac{Y_{i1} + Y_{i'2}}{\sqrt{M_{i1} + M_{i'2}}}. \qquad (5)
$$

Ce résultat est un peu surprenant, puisque par intuition naïve, on pourrait penser que les fractions d'UPE pour la permutation avec $\bar{Y}_i = \bar{Y}_i^*$ préserveront la variance. Pour mieux comprendre (5), considérons les trois cas qui suivent. Si $M_{12} = M_{22}$, alors (5) implique que $Y_{12} = Y_{22}$ ou $\bar{Y}_{12} = \bar{Y}_{22}$, où $\bar{Y}_{il} = Y_{il}/M_{il}$ désigne la moyenne de U_{il}. Si $Y_{12} = Y_{22}$, alors (5) implique que $M_{12} = M_{22}$ ou $\bar{Y}_{12} = \bar{Y}_{22}$. Si $\bar{Y}_{12} = \bar{Y}_{22}$, alors (5) peut s'écrire sous la forme $M_{i1}\bar{Y}_{i1}(\sqrt{M_i^*/M_i} - 1) + \bar{Y}_{i2}(M_{i2}\sqrt{M_i^*/M_i} - M_{i'2}) = 0$ pour $i = 1, 2$, expression qui tient quand $M_{12} = M_{22}$. Ces trois cas démontrent clairement que la variance ne varie pas si les fractions d'UPE pour la permutation sont formées de manière que leurs tailles et leurs moyennes soient les unes et les autre égales.

2.3 Variation du CCI

L'effet de la distorsion de la structure de mise en grappes sur la variance peut aussi être étudié à l'aide du ratio des variances avec et sans masquage. Soit ρ^*_{yU} le CCI avec masquage, c'est-à-dire le CCI défini par les identificateurs d'UPE masqués. De (3), il apparaît clairement que la différence entre les CCI avec et sans masquage est proportionnelle à la différence entre les variances correspondantes, c'est-à-dire

$$\rho^*_{yU} - \rho_{yU} \doteq [m^{-1}(\bar{M}-1)S_y^2]^{-1}[V(\hat{\bar{Y}} \mid S^*) - V(\hat{\bar{Y}} \mid S)].$$

Dans (2), la deuxième approximation indique que la variation du CCI dépend de la façon dont les fractions d'UPE sont formées pour la permutation, c'est-à-dire de la variation de $M_i(\bar{Y}_i - \bar{Y})^2$. Partant de (2), le ratio des variances avec et sans masquage est donné par

$$RV(\hat{\bar{Y}} \mid S,\ S^*) \doteq \frac{\rho^*_{yU} + (\bar{M}-1)^{-1}}{\rho_{yU} + (\bar{M}-1)^{-1}}. \qquad (6)$$

Voir, aussi, Dohrmann et coll. (2005, équation 8). Sous la relation $\rho^*_{yU} = c_y \rho_{yU}$ pour tout $c_y > 0$ donné, (6) est monotone en ρ_{yU}. De plus, le ratio est très instable quand ρ_{yU} ou ρ^*_{yU} est proche de $-(\bar{M}-1)^{-1}$, la borne inférieure du CCI, parce que le numérateur et le dénominateur ayant tous deux leur CCI proche de la borne inférieure, ils sont tous deux proches de zéro. Cela signifie que toute variable de ce type sera fortement influencée par le masquage des UPE.

En général, les enquêtes ont pour but de recueillir des données sur plus d'une variable et le masquage des UPE fondé sur une seule variable pourrait donc ne pas bien protéger le CCI et, par conséquent, la variance d'autres variables. Pour mieux comprendre cet aspect, considérons les situations où le masquage des UPE produit une distorsion à la fois fixe et aléatoire du CCI exprimée sous la forme $\rho^*_{yU} = c_y \rho_{yU} + e$ pour $-0,02 < \rho_y < 0,2$, où $c_y = 0,7,\ 1,0,\ 1,3$, $e \sim N(0,\ 0,05^2)$ et $m = \bar{M} = S_y^2 = 100$. Dans le modèle, le coefficient constant c_y et le terme d'erreur e tiennent compte, respectivement, de la perturbation déterministe et de la perturbation aléatoire causées par le masquage dans le CCI de la variable correspondante. La figure 1 donne le ratio résultant des erreurs-types (racines carrées des variances) avec et sans masquage en fonction du CCI du plan d'échantillonnage. Les trois diagrammes de dispersion de la figure 1 ont une forme en entonnoir similaire et révèlent une forte variation pour les valeurs très faibles de ρ_{yU}. Cependant, leurs courbes génériques dépendent de la grandeur de c_y. Par exemple, $c_y < 1$ produit une courbe décroissante, $c_y > 1$, une courbe croissante et $c_y = 1$, une courbe non monotone. Comme nous en discuterons à la section 3.2, le cas de $c_y > 1$ pourrait être très rare.

La discussion qui précède ne s'étend pas facilement à d'autres situations d'enquête complexe, principalement parce que les enquêtes comportent souvent des caractéristiques de plan de sondage complexes, telles que la stratification, l'échantillonnage à trois degrés ou plus et l'échantillonnage avec probabilités inégales. Dans de telles circonstances, il pourrait être difficile de définir le CCI et de bien approximer la variance sous la forme de (2) (voir, par exemple, Park (2004) et les références qu'il mentionne). Néanmoins, la discussion de la présente section contribue à la compréhension de l'effet du masquage des UPE sur les estimations de la variance en général.

3. Effet du degré de masquage des UPE sur la variance de la moyenne d'échantillon

Le biais dans l'estimation de la variance est d'autant plus important que la structure de mise en grappes est faussée. Pour étudier cette relation, nous considérons un modèle paramétrique utilisé pour l'échantillonnage à deux degrés. Supposons que l'échantillonnage à deux degrés consiste à sélectionner n UPE et m unités dans chaque UPE échantillonnée. À l'exemple de Valliant, Dorfman et Royall (2000, page 248), nous supposons qu'une valeur échantillonnée y_{ij} pour la j^{e} unité de l'UPE i est produite à l'aide du modèle suivant :

$$\xi: E_\xi(y_{ij}) = \mu_{yi} \ \&\ \mathrm{Cov}_\xi(y_{ij},\ y_{i'j'}) = \begin{cases} \sigma^2_{yi} & \text{si } i = i',\ j = j', \\ \sigma^2_{yi}\rho_{yi} & \text{si } i = i',\ j \neq j', \\ 0 & \text{autrement,} \end{cases}$$

où μ_{yi}, σ^2_{yi} et ρ_{yi} sont la moyenne, la variance et la corrélation des unités dans l'UPE i, respectivement. La variance d'une moyenne d'échantillon $\bar{y}_{2st} = (nm)^{-1}\sum_{i=1}^n \sum_{j=1}^m y_{ij}$ s'écrit

$$V_\xi(\bar{y}_{2st} \mid S) = (nm)^{-1}\sigma^2_{yu}[1 + (m-1)\rho_{yu}], \qquad (7)$$

où $\sigma^2_{yu} = n^{-1}\sum_{i=1}^n \sigma^2_{yi}$, $\rho_{yu} = \sum_{i=1}^n(\sigma^2_{yi}\rho_{yi})/\sum_{i=1}^n \sigma^2_{yi}$ et S désigne l'ensemble d'indices dans l'échantillon. Notons que ρ_{yu} peut être interprété comme étant le CCI (groupé ou pondéré par σ^2_{yi}) sous le modèle ξ.

Soit β la taille relative des fractions d'UPE qui doivent être permutées entre les UPE i_1 et i_2. Pour simplifier, nous supposons que $m\beta$ est une valeur entière, qui est le nombre d'unités dans chaque fraction utilisée pour la permutation. Soit S^* l'ensemble d'indices dans l'échantillon avec masquage. La variance de $\bar{y}_{2st}$ avec S^* peut s'écrire

$$V_\xi(\bar{y}_{2st} \mid S^*) = V_\xi(\bar{y}_{2st} \mid S) + n^{-2}(\gamma - 1)(\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2}) \quad (8)$$

pour $\gamma = \beta^2 + (1-\beta)^2$. La preuve de (8) est donnée en annexe. Notons que $-1 < \gamma - 1 < 0$ pour $0 < \beta < 1$. Le ratio des variances avec et sans masquage s'écrit

$$RV_\xi(\bar{y}_{2st} | S,\ S^*) = 1 + m(\gamma - 1)\frac{(\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2})}{\sigma^2_{yu}[1 + (m-1)\rho_{yu}]}. \quad (9)$$

La variance ne variera pas si la permutation est faite de telle manière que $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} = 0$, c'est-à-dire que les corrélations dans les UPE correspondantes soient de sens opposé. Sinon, la variance variera au taux de $m(\gamma - 1) < 0$ pour $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} \neq 0$.

En général, les unités ont tendance à être plus semblables à l'intérieur d'une UPE qu'entre UPE, la valeur de ρ_{yi} étant faible et positive dans de nombreuses populations (par exemple Valliant et coll. 2000, section 8.2.3). Donc, il est plus probable que $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} > 0$, à moins que $\sigma^2_{yi} \doteq 0$ pour tout $i = i_1,\ i_2$ et la variance avec masquage a tendance à être plus faible que la variance sans masquage, c'est-à-dire $RV_\xi(\bar{y}_{2st} | S,\ S^*) < 1$. La figure 2 illustre la variation de l'erreur-type en fonction du CCI sans masquage (ou de référence) ρ_{yu} lorsque la proportion d'unités permutées entre les deux UPE varie. Elle montre que plus le nombre d'unités permutées augmente, plus la variance est modifiée, ce qui indique qu'il faut effectuer la permutation (c'est-à-dire, le masquage des UPE) minimale afin de ne pas induire de biais important dans la variance. En outre, la figure 2 révèle la courbe décroissante en forme de L du ratio des erreurs-types en fonction du CCI, autrement dit indique une surestimation pour les CCI négatifs, mais une sous-estimation pour les CCI positifs. Par conséquent, sous le masquage des UPE, nous pouvons nous attendre à des courbes de type $c_y = 0,7$ (décroissantes, mais aléatoires) ou $c_y = 1,0$ (purement aléatoires) dans la figure 1, le deuxième type représentant les meilleurs résultats qui peuvent être obtenus par masquage minimal. À la section 4, nous proposons une stratégie de masquage des UPE par permutation des USE qui aide à produire une courbe du deuxième type pour les ratios des variances résultants. À la section 5, nous appliquons la stratégie proposée à des données d'enquête artificielles en faisant varier la proportion d'unités permutées.

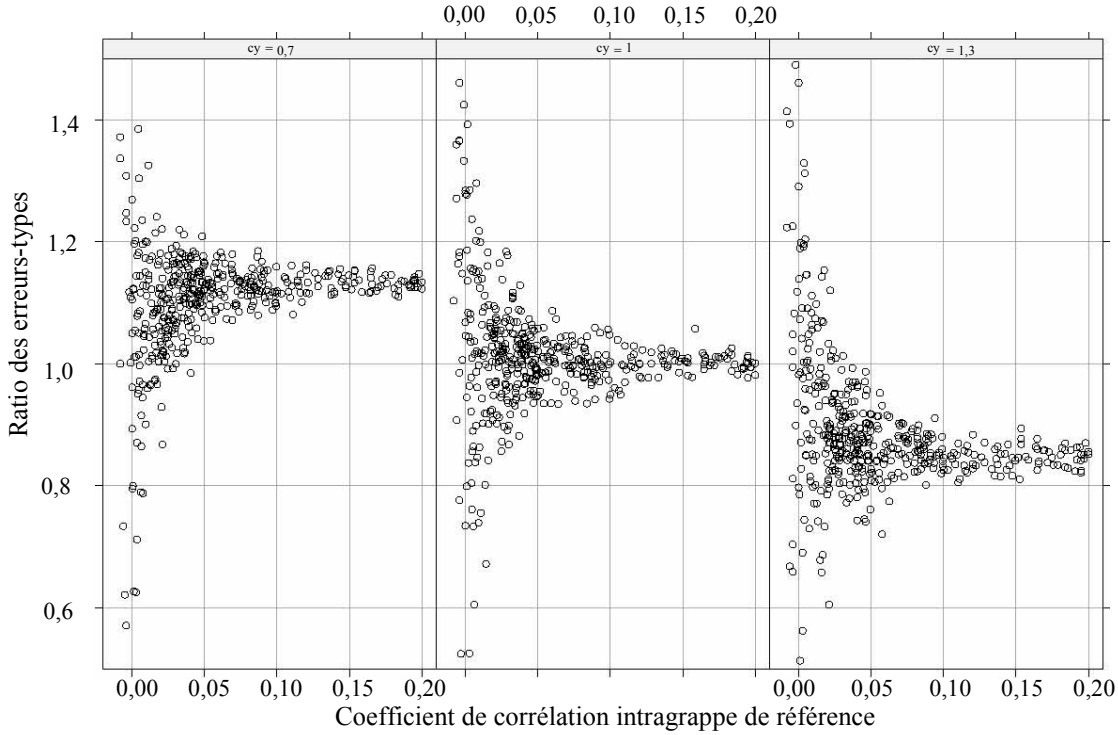

Figure 1 Ratios des erreurs-types avec et sans masquage en fonction du coefficient de corrélation intragrappe original avec variation de l'effet du masquage des UPE sous un modèle $\rho^*_{yU} = c_y\rho_{yU} + e$ avec $e \sim N(0,\ 0,05^2)$ pour $c_y = 0,7,\ 1,0,\ 1,3$ et $m = \bar{M} = S^2_y = 100$

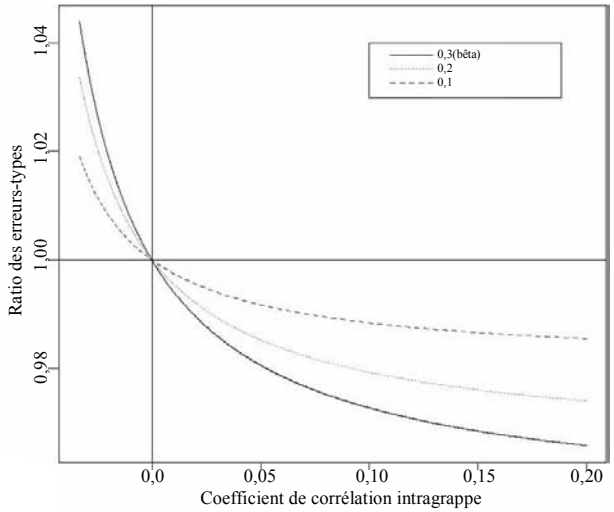

Figure 2 Ratios des erreurs-types avec et sans masquage $\sqrt{RV_{\xi}(\bar{y}_{2st} \mid S, S^*)}$ en fonction du CCI lorsque varie la proportion d'unités permutées provenant de chaque UPE. $\sigma^2_{yu} \equiv 25$, $\rho_{yu} \equiv \rho_{yi}$ pour tout $i, n = 10$, $m = 16$, pour $\beta = (0{,}1,\ 0{,}2,\ 0{,}3)$ et $-0{,}5(m-1)^{-1} \le \rho_{yu} \le 0{,}2$

4. Stratégie de masquage des UPE pour limiter le biais dans l'estimation de la variance

Nombre d'enquêtes à grande échelle comportent plusieurs degrés d'échantillonnage avec probabilités de sélection inégales. Dans de telles circonstances, les unités d'échantillonnage de deuxième degré ou de degré subséquent peuvent être un choix naturel pour la permutation en vue de créer des pseudo-UPE pour des raisons opérationnelles. Par exemple, les diffusions récentes des données de la National Health and Nutrition Examination Survey (NHANES) (Dohrmann et coll. 2005) comprennent les identificateurs de pseudo-UPE construites par permutation des USE entre les UPE originales. À la présente section, nous considérons la permutation des USE en vue de masquer les UPE sous échantillonnage stratifié à plusieurs degrés et son effet sur les estimations de la variance. Nous proposons une stratégie de permutation des USE basée sur la contribution des USE aux estimations des variances.

4.1 USE dans l'estimation de la variance sous échantillonnage stratifié à plusieurs degrés

Supposons qu'une population finie U de M unités est partitionnée en N UPE et que les UPE semblables en ce qui concerne un certain nombre de caractéristiques sont groupées pour former un total de H strates. Supposons aussi que chaque strate est constituée de N_h UPE et que chaque UPE contient N_{hi} USE avec N_{hij} unités finales d'échantillonnage, où $N = \sum_{h=1}^{H} N_h$ et $M = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{N_{hi}} N_{hij}$. Supposons que l'échantillonnage de premier degré consiste à sélectionner $n_h = 2$ UPE dans chaque strate, indépendamment entre les strates, et que le deuxième degré et le degré subséquent d'échantillonnage consistent à sélectionner, tour à tour, n_{hi} USE dans chaque UPE échantillonnée (hi) et n_{hij} unités finales dans chaque USE échantillonnée (hij), où $h = 1, ..., H$, $i = 1, ..., n_h$ et $j = 1, ..., n_{hi}$. L'unité finale échantillonnée $(hijk) \in S$ est associée à la valeur observée y_{hijk} de la variable étudiée y et au poids d'échantillonnage w_{hijk}, où $k = 1, ..., n_{hij}$ et S désigne l'ensemble d'indices dans l'échantillon. Le total de population $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{N_{hi}} \sum_{k=1}^{N_{hij}} y_{hijk}$ et la taille M sont estimés par $\hat{Y} = \sum_{(hijk) \in S} w_{hijk} y_{hijk}$ et $\hat{M} = \sum_{(hijk) \in S} w_{hijk}$, respectivement. En outre, la moyenne de population $\bar{Y} = Y/M$ est estimée par $\hat{\bar{Y}} = \hat{Y}/\hat{M}$ et son estimateur de variance par développement en série de Taylor (par exemple, Shao et Tu 1995) est donné par

$$v(\hat{\bar{Y}} \mid S) = \sum_{h=1}^{H} \left(\frac{z_{h1} - z_{h2}}{2} \right)^2, \qquad (10)$$

où les $z_{hi} = z_{hi}(y) = \sum_{j=1}^{n_{hi}} \sum_{k=1}^{n_{hij}} 2 w_{hijk} z_{hijk}$ sont les totaux estimés de strate de $z_{hijk} = z_{hijk}(y) = \hat{M}^{-1}(y_{hijk} - \hat{\bar{Y}})$ pour l'UPE (hi).

Dans (10), en écrivant z_{hi} en fonction des unités contenues dans les USE, nous pouvons voir la contribution de ces USE à l'estimation de la variance, donc, faciliter la recherche de stratégies de permutation des USE afin de limiter les biais dans les estimations des variances. Si w_{hij} et $w_{k|hij}$ désignent les poids d'échantillonnage des USE et les poids conditionnels des unités finales d'échantillonnage, respectivement, alors $w_{hijk} = w_{hij} \times w_{k|hij}$. Soit $\hat{N}_{hij} = \sum_{k=1}^{n_{hij}} w_{k|hij}$ et $\hat{\bar{Y}}_{hij} = \hat{N}_{hij}^{-1} \sum_{k=1}^{n_{hij}} w_{k|hij} y_{hijk}$ les estimations de la taille et de la moyenne d'échantillon de l'USE (hij), respectivement. Dans (10), les quantités z_{hi} peuvent maintenant s'écrire

$$z_{hi} = \sum_{j=1}^{n_{hi}} 2 w_{hij} z_{hij}, \qquad (11)$$

où

$$z_{hij} = \sum_{k=1}^{n_{hij}} w_{k|hij} z_{hijk} = \hat{M}^{-1} \hat{N}_{hij} (\hat{\bar{Y}}_{hij} - \hat{\bar{Y}}).$$

Il est évident, si l'on examine (10) et (11), que la contribution des USE échantillonnées à l'estimation de la variance a lieu par la voie des trois composantes $\{w_{hij}, \hat{N}_{hij}, \hat{\bar{Y}}_{hij}\}$ de l'USE (hij). À la section 4.2, nous examinerons de près l'effet du masquage des UPE par permutation des USE sur les estimations des variances.

4.2 Effet de la permutation des USE sur les estimations de la variance

Nous supposons maintenant que deux USE $(h_a\, i_a\, j_a)$ et $(h_b\, i_b\, j_b)$ sont permutées entre deux UPE $(h_a\, i_a) \neq (h_b\, i_b)$.

Alors, l'estimation de la variance avec masquage peut, en partant de (10), s'écrire

$$v(\hat{\bar{Y}} \mid S^*) = v(\hat{\bar{Y}} \mid S) + \sum_{h \in \{h_a, h_b\}} \left[\left(\frac{z^*_{h1} - z^*_{h2}}{2} \right)^2 - \left(\frac{z_{h1} - z_{h2}}{2} \right)^2 \right], \quad (12)$$

où z^*_{hi} désigne la quantité z_{hi} dans (11) avec l'ensemble d'indices dans l'échantillon S^* altéré à cause de la permutation. Désignons par i'_a et i_b les autres UPE dans les strates h_a et h_b, respectivement, et définissons $z_{hi(j)} = z_{hi} - z_{hi} - 2w_{hij} z_{hij} = \sum_{l \neq j} 2w_{hil} z_{hil}$. Alors, (12) peut s'écrire

$$v(\hat{\bar{Y}} \mid S^*) = v(\hat{\bar{Y}} \mid S) + e_0(y) g_0(y), \quad (13)$$

où $e_0(y) = 2(w_{h_a i_a j_a} z_{h_a i_a j_a} - w_{h_b i_b j_b} z_{h_b i_b j_b})$ est la différence de quantité $2w_{hij} z_{hij} = 2w_{hij} \hat{M}^{-1} (\hat{\bar{Y}}_{hij} - \hat{\bar{Y}})$ entre les deux USE qui seront permutées et

$$g_0(y) = \begin{cases} [z_{h_a i_a (j_a)} - z_{h_b i_b (j_b)}] & \text{si } h_a = h_b, \\ 2^{-1}[(z_{h_a i'_a} - z_{h_a i_a (j_a)}) - (z_{h_b i'_b} - z_{h_b i_b (j_b)})] & \text{si } h_a \neq h_b, \end{cases}$$

est une fonction de $2w_{hij} z_{hij}$ des USE qui seront gardées dans les UPE originales. Notons que, pour $h_a = h_b$, r_0 peut aussi s'exprimer $g_0 = 2^{-1}[(z_{h_b i_b (j_b)} - z_{h_a i_a (j_a)}) - (z_{h_a i_a (j_a)} - z_{h_b i_b (j_b)})]$ ce qui montre que l'effet de la permutation des USE sur l'estimation de la variance sera négligeable si les deux USE permutées sont appariées de manière que le produit de $e_0(y)$ et $g_0(y)$ soit proche de zéro. Autrement dit, la variation de l'estimation de la variance sous permutation des USE peut être contrôlée si l'on forme une paire de segments tenant compte des trois composantes $\{w_{hij}, \hat{N}_{hij}, \hat{\bar{Y}}_{hij}\}$ de manière à minimiser $e_0(y) \times g_0(y)$ similairement au cas sous échantillonnage en grappes PPT à un degré de la section 2.2.

De plus, en écrivant

$$g_0 = g(e_0) = \begin{cases} (z_{h_a 1} - z_{h_a 2}) - e_0 & \text{si } h_a = h_b, \\ 2^{-1}\left[(z_{h_a 1} - z_{h_a 2}) - (z_{h_b 1} - z_{h_b 2}) - e_0\right] & \text{si } h_a \neq h_b, \end{cases}$$

nous pouvons exprimer (13) comme une fonction quadratique de $e_0(y)$ sachant $\{z_{hi} : h = \{h_a, h_b\}, i = 1, 2\}$. Pour $h_a = h_b$, nous pouvons montrer que $v(\hat{\bar{Y}} \mid S^*) > v(\hat{\bar{Y}} \mid S)$ uniquement pour e_0 compris entre zéro et $(z_{h_a 1} - z_{h_b 2})$. Quand $z_{h_a 1} - z_{h_b 2} \doteq 0$, il pourrait être plus probable que $v(\hat{\bar{Y}} \mid S^*) = v(\hat{\bar{Y}} \mid S)$. Un argument comparable peut être formulé pour $h_a \neq h_b$.

4.3 Permutation séquentielle d'USE sous caractéristiques d'appariements multiples

Supposons que l'échantillon contient un total de n_J USE et que seulement R d'entre elles sont choisies pour former des paires pour la permutation, où $n_J = \sum_{h=1}^{H} \sum_{i=1}^{n_h} n_{hi}$ et $1 \leq R < n_J$. Supposons qu'un nombre fixe de R USE est choisi conformément à une certaine notion de compromis risque-utilité. Voir, par exemple, Gomatam, Karr et Sanil (2005) pour une discussion apparentée concernant la permutation de données. En outre, supposons que leur ordre séquentiel pour le processus d'appariement est donné par $j_1, j_2, ..., j_R$. Par exemple, pour commencer, toutes les paires possibles sont formées pour chacune des R USE et la meilleure d'entre elles est choisie en se fondant sur une mesure de distance, telle que (12). L'ordre des R USE pour le (principal) processus d'appariement est alors déterminé en fonction de l'ordre croissant des distances des R meilleures paires.

Soit S^{r-1} l'ensemble modifié d'indices dans l'échantillon après que la $(r-1)^{\text{e}}$ paire d'USE a été formée et permutée, où $r = 1, ..., R$ et $S^0 \equiv S$. Soit $S^{r-1}_{(j)}$ l'ensemble d'indices dans l'échantillon quand les USE j_r et j sont permutées pour S^{r-1}. Alors, la variation de l'estimation de la variance causée par la permutation de la r^{e} USE j_r et toute autre USE non comprise dans les $(r-1)$ appariements précédents peut s'écrire sous la forme

$$\begin{aligned} \delta_r(y, j) &= v(\hat{\bar{Y}} \mid S^{r-1}_{(j)}) - v(\hat{\bar{Y}} \mid S^{r-1}) \\ &= e_{r-1}(y, r)\, g_{r-1}(y, r), \end{aligned} \quad (14)$$

où $e_{r-1}(y, r)$ et $g_{r-1}(y, r)$ sont définis de la même façon que dans (13), mais avec S^{r-1} et $S^{r-1}_{(j)}$. Manifestement, le choix de la meilleure paire pour la r^{e} USE dépend des $(r-1)$ appariements précédents et, donc, le processus d'appariement doit être considéré comme étant séquentiel. Notons que les USE qui ont été appariées et permutées au moment des appariements précédents doivent être exclues de la recherche courante.

En outre, plusieurs caractéristiques peut être prises en considération pour l'appariement, dans l'espoir qu'elles seront reliées à de nombreuses autres variables étudiées, de façon à minimiser le biais dans l'estimation connexe de la variance. Supposons que q caractéristiques d'appariement sont choisies avec soin, disons $\mathbf{x} = (x_1, x_2, ..., x_q)'$ (par exemple, Dohrmann et coll. 2005, pour une discussion à ce sujet). Pour déterminer la distance entre les USE j_r et j, nous pouvons considérer toute mesure de distance de la forme

$$D_r(j \mid \mathbf{x}) = \sum_{l=1}^{q} c_l \left| v[\hat{\bar{X}}_l \mid S^{r-1}_{(j)}] - v[\hat{\bar{X}}_l \mid S^{r-1}] \right| \quad (15)$$

ou

$$\Delta_r(j \mid \mathbf{x}) = \sum_{l=1}^{q} c_l \left| v[\hat{\bar{X}}_l \mid S^{r-1}_{(j)}] - v[\hat{\bar{X}}_l \mid S] \right| \quad (16)$$

moyennant tout choix raisonnable de coefficients positifs c_l. Par exemple, $c_l \equiv 1$ tient compte simplement de l'écart

absolu entre les estimations de la variance de $\hat{X}_l$, $c_l = v(\hat{\bar{X}}_l \mid S)^{-1}$, de l'écart absolu entre les estimations de la variance comparativement aux estimations originales de la variance et $c_l = X_l^{-1}$, de l'écart absolu entre les estimations de la variance relative. La première mesure de distance (15) rend compte de la variation de l'estimation de la variance due à la permutation de segments de la r^{e} paire. La deuxième mesure de distance (16) tient compte de l'effet cumulatif des permutations pour les r paires de segments.

Nous pouvons imposer des contraintes d'appariement, par exemple, pour empêcher l'appariement d'USE provenant de la même UPE et pour appliquer un seuil de la proportion d'USE provenant de chaque UPE qui seront permutées (Lu 2004). Soit $J_A = \{j_1, \ldots, j_R\}$ l'ensemble d'indices des R USE prises en considération pour former les paires à permuter et soit J_B toutes les USE possibles qui peuvent être appariées en satisfaisant un ensemble donné de contraintes d'appariement. Pour simplifier, considérons que l'appariement des USE n'est pas permis à l'intérieur de J_A, c'est-à-dire $J_A \cap J_B = \varnothing$. Si $D_r^*(j \mid \mathbf{x})$ désigne la mesure de distance choisie pour les USE j_r et j, un algorithme de permutation séquentielle des USE en vue de limiter les biais dans les estimations de la variance peut prendre la forme suivante :

Étape 1. Poser que $r = 1$, $J_A^r = J_A$ et $J_B^r = J_B$;

Étape 2. Pour chacune des $(R - r + 1)$ USE dans J_A^r, calculer $D_r^*(\cdot \mid \mathbf{x})$ pour toutes les USE dans J_B^r ;

Étape 3. Choisir le meilleur appariement ayant la distance $D_r^*(\cdot \mid \mathbf{x})$ la plus faible, autrement dit, trouver j_r' telle que $D_r^*(j_r' \mid \mathbf{x}) = \min_{j \in J_B^r} D_r^*(j \mid \mathbf{x})$;

Étape 4. Poser que $r = r + 1$ et retirer la paire choisie du groupe d'unités sur lequel porte la recherche, autrement dit, poser que $J_A^r = J_A^{r-1} \setminus \{j_r\}$ et mettre à jour J_B^r en conséquence, où $J_B^r \subseteq J_B^{r-1} \setminus \{j_r'\}$;

Étape 5. Si $r = R + 1$, arrêter ; sinon, répéter les étapes 2 à 4.

Cette méthode d'appariement (ou de permutation) des USE consiste fondamentalement à rechercher la paire qui, à l'étape du r^{e} appariement, est la meilleure au sens de la minimisation de la variation des estimations de la variance due à la permutation des USE correspondantes. Si le nombre d'USE est grand, cette méthode produit un diagramme de dispersion semblable à celui illustré pour $c_y = 1{,}0$ à la figure 1 (c'est-à-dire, une perturbation aléatoire avec une courbe en forme d'entonnoir).

Le choix de critères d'optimalité plus perfectionnés en ce qui concerne $\{c_l : l = 1, \ldots, q\}$ pourrait aider à améliorer la méthode susmentionnée en vue de réduire l'ampleur de ce genre de perturbation aléatoire dans les estimations de la variance. En outre, en utilisant des techniques multivariées, telles que l'analyse en composantes principales, pour établir certains types de scores (par exemple, un ou plusieurs axes factoriels (composantes principales)) à partir d'un grand nombre de caractéristiques continues, il serait peut-être possible de réduire davantage l'importance de cette perturbation aléatoire dans les estimations de la variance. À la section 5, nous donnons des exemples en ce qui concerne la permutation des USE.

5. Exemples

5.1 Travaux antérieurs

Dans le cas d'un plan d'échantillonnage sans stratification, mais comportant un petit nombre d'UPE, Dohrmann et coll. (2002) ont examiné diverses méthodes de partitionnement des UPE en pseudo-UPE afin d'appliquer la méthode du jackknife avec suppression d'une unité pour estimer la variance. Leur idée fondamentale était de doubler le nombre d'UPE masquées en gardant les UPE fractionnées comme des UPE masquées distinctes, espérant ainsi réduire le risque de divulgation des données grâce à la rupture du lien entre les UPE réelles et masquées. Leur étude empirique a révélé, dans les estimations résultantes de la variance pour les variables présentant des effets de plan importants, des courbes de sous-estimation appréciable qui ressemblent au graphique de $c_y = 0{,}7$ à la figure 1. Soit S et $S^{\dagger}$ les ensembles d'indices dans l'échantillon sans masquage et avec masquage, respectivement. Soit w_{ij} le poids d'échantillonnage et soit y_{ij} la valeur observée de y pour la j^{e} unité échantillonnée dans l'UPE i. Pour expliquer les courbes de sous-estimation observées, Dohrmann et coll. (2005) ont établi la relation suivante

$$v(\hat{\bar{Y}} \mid S^{\dagger}) = \frac{n-1}{2n-1} v(\hat{\bar{Y}} \mid S) + \frac{1}{n(2n-1)} \sum_{i=1}^{n} (z_{1,i} - z_{2,i})^2,$$

où les $z_{g,i} = \sum_{j \in S_{g,i}} 2 w_{ij} z_{ij}$ sont les totaux dans les fractions d'UPE de $z_{ij} = (\sum_{ij} w_{ij})^{-1} (y_{ij} - \sum_{ij} w_{ij} y_{ij} / \sum_{ij} w_{ij})$ et $S_{u,i}$ sont les ensembles d'indices de la u^{e} fraction de l'UPE i pour $i = 1, \ldots, n$ et $u = 1, 2$. Cette expression indique que l'estimation de la variance résultante est égale à environ la moitié de celle obtenue sans masquage à laquelle s'ajoute une valeur positive reflétant les totaux entre les fractions d'UPE des z_{ij} dans les UPE. Si les ensembles $S_{u,i}$ sont formés de manière que $z_{1,i} \doteq z_{2,i}$, cette méthode de partitionnement des UPE donne lieu à environ la moitié de l'estimation de la variance sans masquage et l'estimation de la variance avec masquage doit donc être doublée pour s'approcher de la valeur sans masquage.

Dans le cas du plan à deux UPE par strate, Dohrmann et coll. (2005) ont envisagé une autre approche dans laquelle les pseudo-UPE sont construites par permutation des USE

entre les UPE. Comme nous le mentionnons à la section 1, cette approche peut être considérée comme la division des UPE en une ou plusieurs fractions et la recombinaison de ces fractions pour construire des pseudo-UPE dans lesquelles les fractions d'UPE sont permutées. Pour simplifier, nous supposons que chaque UPE est divisée en deux fractions $S_{1,hi}$ et $S_{2,hi}$. Si la partition est telle que $S_{h1}^{\dagger} = S_{1,h1} \cup S_{1,h2}$ et $S_{h2}^{\dagger} = S_{2,h1} \cup S_{2,h2}$, l'estimation de la variance avec masquage peut s'écrire

$$v(\hat{\bar{Y}} \mid S^{\dagger}) = v(\hat{\bar{Y}} \mid S) + \sum_{h=1}^{H} e_h(y)\, g_h(y), \qquad (17)$$

où $e_h(y) = z_{1,h2} - z_{2,h1}$ est la différence entre les totaux dans les fractions d'UPE $S_{1,h2}$ et $S_{2,h1}$ à permuter et $g_h(y) = z_{1,h1} - z_{2,h2}$ est la différence entre les totaux dans les fractions d'UPE $S_{1,h1}$ et $S_{2,h2}$ à garder dans les UPE originales. La preuve de (17) est donnée en annexe. L'équation (17) indique qu'une stratégie semblable pour le partitionnement des UPE aiderait à préserver la grandeur de l'estimation originale de la variance. Dohrmann et coll. (2005) ont adopté une méthode probabiliste de couplage d'enregistrements (Fellegi et Sunter 1969) pour former des paires d'USE pour la permutation dont les moyennes $\hat{\bar{Y}}_{hij}$ sont semblables pour plusieurs caractéristiques, dans l'espoir que, dans (17), les termes e_h soient tous proches de zéro. Dohrmann et coll. (2005) ont démontré que la méthode de recombinaison des UPE permet dans une certaine mesure de réduire les biais dans les estimations de la variance et les courbes de sous-estimation résultantes comparativement à la méthode de partitionnement des UPE appliquée dans Dohrmann et coll. (2002). Afin de rendre la méthode plus rapide et plus souple, Lu (2004) a élaboré des algorithmes de permutation des USE basés sur l'évaluation séquentielle de mesures de distance entre les moyennes d'USE, sans tenir compte directement de l'incidence des permutations successives sur le biais de l'estimation de la variance. Comme nous l'avons exposé à la section 4.2, l'effet de la permutation des USE sur les estimations de la variance peut être réduit davantage en tenant compte directement de la contribution des USE à ces estimations. À la section suivante, nous appliquons deux stratégies, celle décrite par Dohrmann et coll. (2005) et celle proposée à la section 4.2, à des données artificielles provenant d'une enquête complexe.

5.2 Exemple de données

Afin d'illustrer l'effet du masquage des UPE sur les estimations de la variance des moyennes d'échantillon, nous avons utilisé des données d'enquête réelles en apportant certaines modifications artificielles au fichier à grande diffusion (FGD) de données sur les objectifs en matière de santé pour l'an 2000 de la National Health Interview Survey (NHIS) de 1993. La NHIS est une enquête-ménages annuelle avec interview sur la santé menée auprès de la population civile non placée en établissement des États-Unis. La NHIS est réalisée selon un plan d'échantillonnage à plusieurs degrés stratifié typique dans lequel les UPE de premier degré correspondent aux comtés ou aux régions métropolitaines et les USE de deuxième degré correspondent à des segments (c'est-à-dire, un petit nombre de ménages dans une petite région géographique) à l'intérieur des UPE échantillonnées. Le questionnaire particulier sur le thème de l'an 2000 a été administré à un seul adulte échantillonné par famille durant la deuxième moitié de 1993. Les données de la NHIS utilisées ici et la documentation les concernant peuvent être consultées dans National Center for Health Statistics (1994) ou sur le site Web du National Center for Health Statistics des Centers for Disease Control des États-Unis (http://www.cdc.gov/nchs/about/major/nhis/quest_data_related_doc.htm).

Ce FGD contient les identificateurs de strate et d'UPE, ainsi que les poids finaux des personnes échantillonnées pour l'estimation de la variance. Dans notre exemple, nous n'avons utilisé que dix strates, mais avons limité le nombre d'UPE à deux par strate. Deux des strates sélectionnées, 110 et 520, ont été limitées à leurs deux plus grandes UPE, (181, 410) et (048, 233) respectivement, et les huit autres strates, 102, 142, 192, 211, 261, 300, 561 et 571, ne contiennent que deux UPE. Le FGD contient aussi les identificateurs des USE, mais non leurs poids d'échantillonnage. Pour produire les poids d'échantillonnage des USE, w_{hij}, nous nous sommes servis d'un modèle à deux facteurs avec effets aléatoires emboîtés pour ajuster $\log w_{hijk} = \log w_{hij} + \log w_{k|hij}$ de manière que $\log w_{hij} = \mu + \alpha_{hi} + \beta_{j(hi)}$ et $\log w_{k|hij} = \varepsilon_{k(hij)}$, où μ est une valeur commune, α_{hi} est l'effet aléatoire de l'UPE (hi), $\beta_{j(hi)}$ est l'effet aléatoire de l'USE j emboîtée dans l'UPE (hi) et $\varepsilon_{k(hij)}$ est l'effet aléatoire de la personne échantillonnée k dans l'USE (hij). Nous avons restreint notre étude seulement aux USE comptant cinq personnes échantillonnées ou plus, ce qui nous a donné un total de 293 USE pour l'analyse. La décomposition résultante des poids $(w_{hij}, w_{k|hij})$ peut comprendre d'éventuelles spécifications incorrectes du modèle, mais elle suffit pour les besoins de l'illustration, puisque les w_{hij} ainsi que les $w_{k|hij}$ sont tous positifs sous le modèle. Pour obtenir des paires d'USE pour la permutation, nous avons utilisé six variables sociodémographiques, désignées par $x_1, x_2, \ldots, x_6$. Elles sont énumérées au tableau 1 avec leur description, leur définition, leur moyenne globale d'échantillon et la racine carrée de leur effet de plan (« design factor » ou deft en abrégé). Les moyennes d'échantillon de ces variables varient de 0,05 à 0,63 et les valeurs deft, de 1,285 à 8,511.

Nous avons appliqué les deux stratégies d'appariement des USE discutées aux sections 4 et 5.1, respectivement. La

première consiste à employer une mesure de distance (15) pour toute paire d'USE (r_a, r_b) avec $c_l \equiv 1$ pour tout $l = 1, ..., 6$. Soit S^{r-1} et S^r les deux ensembles d'indices dans l'échantillon après les $(r-1)^{\text{e}}$ et r^{e} permutations, respectivement. Alors, la distance de la r^{e} paire appariée de la première stratégie (variance-appariement) s'écrit

$$D_r(v|\mathbf{x}) = \sum_{l=1}^{6} \left| v(\hat{\bar{X}}_l|S^r) - v(\hat{\bar{X}}_l|S^{r-1}) \right|,$$

où $v(\hat{\bar{X}}_l|S^r)$ et $v(\hat{\bar{X}}_l|S^{r-1})$ représentent les estimations de la variance de $\hat{\bar{X}}_l$ pour la l^{e} caractéristique d'appariement avec S^r et S^{r-1}, respectivement. Le biais dans les estimations de la variance qui résulte de la permutation de la r^{e} paire appariée est d'autant plus faible que la distance est courte. La deuxième stratégie, proposée par Dohrmann et coll. (2005), consiste à apparier les USE dont les moyennes d'échantillon sont semblables pour les six caractéristiques d'appariement. Cette stratégie (moyenne-appariement) définit la distance de la r^{e} paire appariée comme étant :

$$d_r(\mu \mid \mathbf{x}) = \sum_{l=1}^{6} \left| \hat{\bar{X}}_{l,r_a} - \hat{\bar{X}}_{l,r_b} \right|,$$

où $\hat{\bar{X}}_{l,r_i}$ représente la moyenne d'échantillon de l'USE r_i $(i = a, b)$ pour la caractéristique d'appariement x_l.

Tableau 1
Variables utilisées pour l'appariement

Variable	Description	Définition	Moyenne d'échantillon	Valeur deft
x_1	Sexe masculin	SEX=1	0,49	1,285
x_2	Hispanicité	HISPANIC	0,14	8,511
		= 00, 01, ..., 08	0,14	8,511
x_3	Couple marié	MARSTAT = 1,2	0,63	3,209
x_4	Études collégiales ou de niveau plus élevé	EDUCR = 4, 5, 6	0,45	2,902
x_5	Revenu familial élevé (50 000 $ ou plus)	INCFAMR = 8	0,23	3,558
x_6	L'air ambiant dans le ménage a-t-il été testé pour la présence de radon ?	TESTRDN = 1	0,05	2,191

Le tableau 2 donne les ratios des erreurs-types pour les six caractéristiques d'appariement à chaque étape d'appariement dans l'ordre séquentiel pour chaque stratégie avec 18 paires permutées (environ 12 % de la totalité des USE dans les données). La première stratégie, présentée dans la partie de gauche du tableau, donne un intervalle modéré, mais légèrement croissant, de variation des ratios des erreurs-types sur la série des 18 paires sélectionnées pour la permutation. La deuxième stratégie, présentée dans la partie de droite du tableau, produit un intervalle assez vaste de variation des ratios des erreurs-types sur la série, les changements étant très grands pour la treizième paire et les suivantes de la série de permutations. Bien que ces stratégies tendent toutes deux à perdre le contrôle sur le biais dans les estimations de la variance pour les ordres élevés de la série de permutations, la première permet de relativement bien maîtriser le biais dans les estimations de la variance pour un nombre assez grand de paires à permuter.

À la figure 3, les ratios des erreurs-types sont tracés en fonction des valeurs deft pour les deux stratégies en faisant varier le nombre d'USE permutées. Les trois ensembles comprennent 6 (4 %), 12 (8 %) et 18 (12 %) paires d'USE (pourcentage d'USE intervenant dans les permutations), respectivement. Chaque graphique comprend deux ensembles de caractéristiques, six caractéristiques d'appariement marquées par les chiffres correspondant à ceux énumérés au tableau 1 et 92 caractéristiques marquées d'un × qui ne sont pas utilisées dans l'appariement. Dans le scénario où 4 % seulement des USE sont permutées, la différence entre les deux stratégies est négligeable pour les deux ensembles de caractéristiques. Cependant, à mesure que le pourcentage d'USE permutées augmente, la perturbation dans les estimations de la variance s'accroît pour les deux stratégies et les deux ensembles de caractéristiques. Ce résultat indique que le pourcentage de permutations effectuées devrait être faible, ce qui renforce les résultats de la section 3. En outre, les ratios des erreurs-types sont regroupés plus près de la ligne de valeur un (c'est-à-dire petit biais dans les estimations de la variance avec masquage) dans la première stratégie que dans la seconde. Cette dernière produit une courbe assez fortement décroissante en fonction des valeurs deft, même pour les six caractéristiques d'appariement. Autrement dit, la stratégie moyenne-appariement semble avoir un effet plus prononcé pour les variables utilisées pour l'appariement.

Tableau 2
Ratios des erreurs-types selon la série de permutations : comparaison de deux critères d'appariement avec permutation de 12 % (18 paires) d'USE

Série de permutations	Variance-appariement						Moyenne-appariement					
	x_1	x_2	x_3	x_4	x_5	x_6	x_1	x_2	x_3	x_4	x_5	x_6
1	0,999	1,000	1,000	0,998	1,000	0,998	0,998	1,000	1,000	1,002	1,002	1,001
2	1,002	1,000	1,000	1,000	1,000	0,996	0,999	1,000	1,000	1,002	1,001	1,001
3	1,004	1,000	1,000	1,001	1,000	0,996	0,998	1,000	0,999	0,997	0,994	1,001
4	1,012	1,001	0,999	1,000	1,000	0,989	1,025	1,000	0,999	1,001	0,994	1,016
5	1,009	1,001	1,000	0,998	1,001	0,988	1,021	1,004	0,964	0,968	0,951	1,013
6	1,007	1,000	1,000	1,000	1,003	0,988	1,020	1,004	0,964	0,968	0,955	1,015
7	1,011	1,000	1,000	1,002	1,003	1,008	1,020	1,004	0,964	0,970	0,954	1,017
8	1,016	1,000	0,997	1,002	1,003	1,026	1,022	0,998	0,957	0,963	0,964	1,005
9	1,009	0,998	1,000	1,002	1,003	1,020	1,021	0,997	0,955	0,965	0,982	1,034
10	1,007	0,996	1,001	1,010	1,006	1,014	1,019	0,997	0,931	0,960	0,972	1,033
11	1,014	0,994	1,005	1,010	1,001	1,012	1,020	0,995	0,946	0,953	0,989	1,034
12	1,029	0,995	1,003	1,013	1,011	1,036	1,021	0,991	0,946	0,953	0,987	1,035
13	1,064	0,992	1,003	1,001	1,008	1,047	1,035	0,990	0,946	0,932	0,967	1,114
14	1,008	0,991	1,000	1,007	1,022	1,044	1,031	0,955	0,946	0,929	0,952	1,103
15	1,042	0,988	0,984	1,017	1,015	1,044	1,052	0,955	0,946	0,922	0,952	1,124
16	1,012	0,982	0,986	1,024	1,041	1,042	1,107	0,939	0,936	0,920	0,942	1,128
17	0,987	0,978	1,000	1,016	1,009	1,021	1,107	0,878	0,936	0,927	0,935	1,123
18	1,029	0,943	1,000	0,970	0,947	1,042	1,014	0,538	0,946	0,841	0,945	1,106

Voir le tableau 1 pour la description des six caractéristiques d'appariement (x_1, ..., x_6).

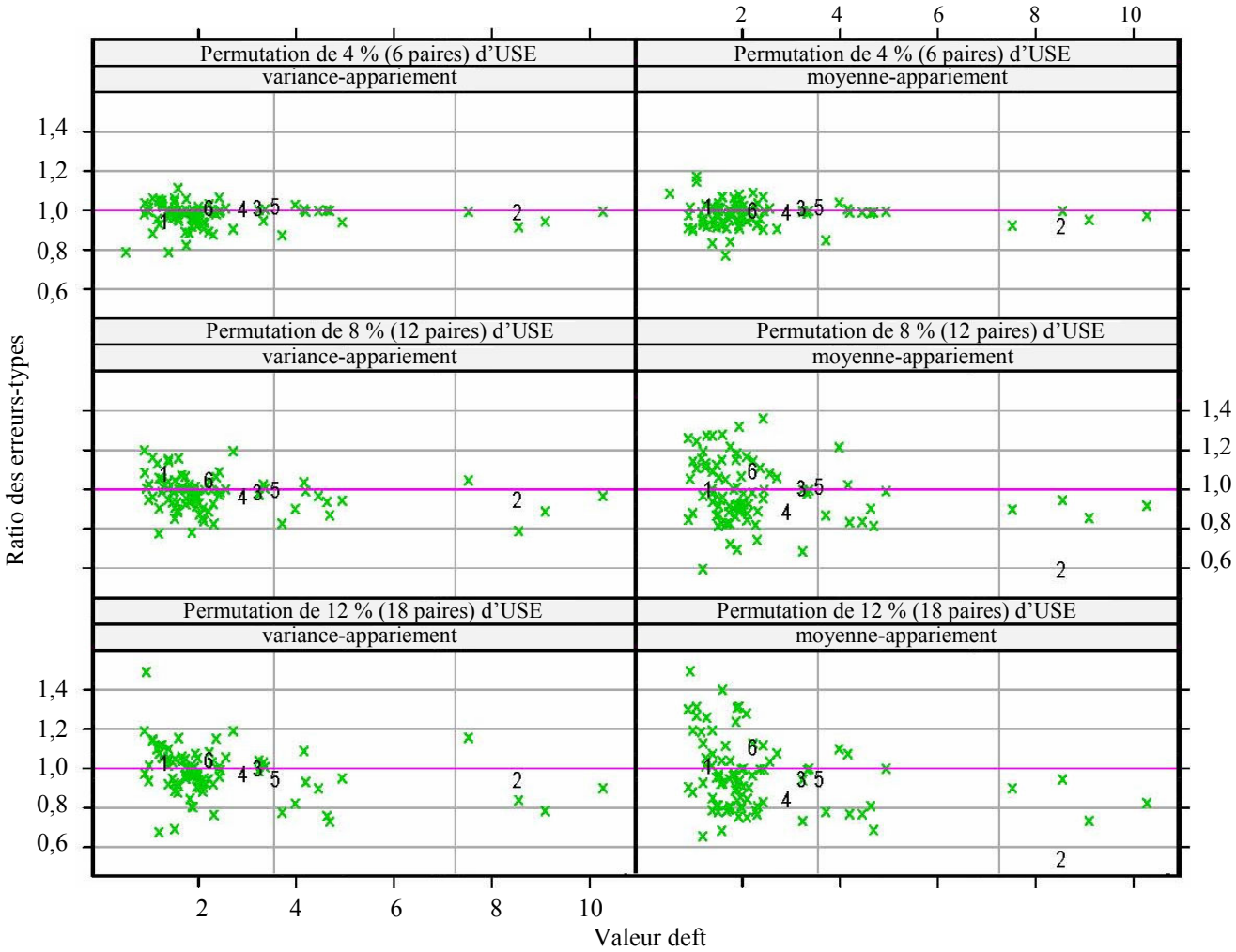

Figure 3 Rapport des erreurs-types en fonction des valeurs deft de référence. Six chiffres représentent les points pour les caractéristiques d'appariement correspondantes et les × représentent ceux des 92 caractéristiques qui n'ont pas été utilisées pour l'appariement

6. Discussion

Dans le présent article, nous avons étudié l'effet du masquage des UPE sur les estimations de la variance dans les enquêtes complexes. Manifestement, ce masquage fausse la structure de mise en grappes du plan d'échantillonnage original, ce qui peut induire des biais systématiques dans l'analyse des données résultantes, comme l'illustrent les sections 2, 3 et 5.2. La stratégie de masquage des UPE proposée à la section 4 peut aider à réduire ces biais, mais laisse néanmoins une perturbation aléatoire dans l'estimation de la variance et, donc, une perte d'efficacité inférentielle. Il serait intéressant d'étudier l'effet du masquage des UPE sur d'autres types d'analyse de données complexes, comme les analyses par régression et les analyses multivariées. Bien que le masquage des UPE puisse offrir un moyen de contrôler la divulgation, le degré de masquage doit être minimal afin de limiter les biais résultants dans les estimations de la variance, comme il est discuté aux sections 3 et 5.2.

En outre, la réduction du risque d'identification attribuable au masquage des USE pourrait être mieux comprise en exprimant la distance entre la moyenne des UPE échantillonnées masquées et la moyenne des UPE dans la population de la façon suivante :

$$\hat{\bar{Y}}_{hi|S^\dagger} - \bar{Y}_{hi|U} = (\hat{\bar{Y}}_{hi|S^\dagger} - \hat{\bar{Y}}_{hi|S}) + (\hat{\bar{Y}}_{hi|S} - \bar{Y}_{hi|U}), \qquad (18)$$

où $\hat{\bar{Y}}_{hi|S^\dagger}$ et $\hat{\bar{Y}}_{hi|S}$ désignent les moyennes des UPE masquées et non masquées dans l'échantillon, respectivement, et $\bar{Y}_{hi|U}$ désigne la moyenne des UPE dans la population auxquelles pourrait avoir accès un intrus (c'est-à-dire un utilisateur de données malveillant) à partir de sources externes, telles que les données de recensement. Il est facile de montrer qu'en général, le premier terme du deuxième membre de (18) n'est pas nul en cas de masquage des UPE. Les termes $\hat{\bar{Y}}_{hi|S^\dagger}$ et $\bar{Y}_{hi|U}$, conjugués à une variation dans l'échantillon non négligeable du deuxième terme du membre de (18), ne sont jamais égaux, sauf dans des cas très rares. Dohrmann et coll. (2005) comparent $\{\hat{\bar{Y}}_{hi|S^\dagger}\}$ de l'échantillon et $\{\bar{Y}_{hi|U}\}$ de la population au moyen d'un diagramme arborescent élégant pour démontrer à quel point un intrus aurait de la difficulté à identifier une UPE échantillonnée dans les fichiers de données à grande diffusion compte tenu de deux aspects, à savoir que 1) peu de paires $(\hat{\bar{Y}}_{hi|S^\dagger}, \bar{Y}_{hi|U})$ sont proches l'une de l'autre et 2) de nombreuses UPE non échantillonnées ont des valeurs de population semblables à $\{\hat{\bar{Y}}_{hi|S^\dagger}\}$ ou à $\{\bar{Y}_{hi|U}\}$ des UPE échantillonnées. Certaines formes de mesure probabiliste pourraient être intéressantes pour évaluer la réduction du risque d'identification (par exemple, Eltinge 1999), mais elles dépassent le cadre du présent article. La stratégie de masquage proposée a été appliquée aux données diffusées de la National Health and Nutrition Examination Survey (NHANES) de 2003-2004 (Park, Dohrmann, Montaquila, Mohadjer et Curtin 2006).

Annexe

Preuves

Preuve de l'équation (1)

D'après Park et Lee (2004, section 4.2),

$$\begin{aligned} V(\hat{\bar{Y}} \mid S) &= \frac{1}{m} S_y^2 \times \mathrm{Deft}^2(\hat{\bar{Y}} \mid S) \\ &\doteq \frac{1}{m} S_y^2 [1 + (\bar{M} - 1)\rho_{yU}] \\ &\quad + \frac{1}{mN} \sum_{i=1}^N \frac{M_i}{p_i}\left(\frac{M_i}{p_i} - p_i\right)(\bar{Y}_i - \bar{Y})^2 \\ &\doteq \frac{1}{m(N-1)} \sum_{i=1}^N M_i (\bar{Y}_i - \bar{Y})^2 \\ &\quad + \frac{1}{mN} \sum_{i=1}^N \frac{M_i}{p_i}\left(\frac{M_i}{M} - p_i\right)(\bar{Y}_i - \bar{Y})^2, \\ &\doteq \frac{1}{mMN} \sum_{i=1}^N \frac{M_i^2}{p_i} (\bar{Y} - \bar{Y})^2, \end{aligned}$$

où $\mathrm{Deft}^2(\hat{\bar{Y}} \mid S)$ représente l'effet de plan de $\hat{\bar{Y}}$ pour un S donné, les deuxième et dernière approximations découlent de $(N-1)/N \doteq 1$ et la troisième équation, de $S_y^2[1 + (\bar{M} - 1)\rho_{yU}] \doteq (N-1)^{-1}\sum_{i=1}^N M_i(\bar{Y}_i - \bar{Y})^2$, ce qui complète la preuve.

Preuve de l'équation (8)

Par définition, la variance du total des UPE échantillonnées $y_i = \sum_j y_{ij}$ est $V_\xi(y_i \mid S) = m\sigma_{yi}^2 + m(m-1)\sigma_{yi}^2\rho_{yi}$ pour $i = 1, ..., n$. Supposons que les $\{y_{ij} : j = m(\beta - 1) + 1, ..., n\}$ pour les deux UPE $i = a$ et b sont permutées. Alors, la variance de ces deux UPE devient $V_\xi(y_a \mid S^*) = m(1-\beta)\sigma_a^2 + m(1-\beta)[m(1-\beta)-1]\sigma_a^2\rho_a + m\beta\sigma_b^2 + m\beta(m\beta - 1)\sigma_b^2\rho_b$ et $V_\xi(y_b \mid S^*)$ donnée par la même expression en permutant les indices a et b. Puisque $m(1-\beta)[m(1-\beta)-1] + m\beta(m\beta - 1) = m(m\gamma - 1)$, la preuve est complétée en observant que

$$\begin{aligned} (nm)^2 V_\xi(\bar{y}_{2st} \mid S^*) &= \sum_{i \neq a,b}^n [m\sigma_{yi}^2 + m(m-1)\sigma_{yi}^2\rho_{yi}] \\ &\quad + \sum_{i=a,b} [m\sigma_{yi}^2 + m(m\gamma - 1)\sigma_{yi}^2\rho_{yi}] \\ &= (nm)^2 V_\xi(\bar{y}_{2st} \mid S) \\ &\quad + \sum_{i=a,b} m^2(\gamma - 1)\sigma_{yi}^2\rho_{yi}. \end{aligned}$$

Preuve de l'équation (13)

Supposons que deux UPE (h1) et (k1) provenant de deux strates distinctes $h \neq k$ sont reconstruites en

permutant chacune de leur USE, $(h1j_a)$ et $(k1j_b)$. Soit $e_{hk} = 2(w_{h1j_a}y_{h1j_a} - w_{k1j_b}y_{k1j_b})$ la différence entre les contributions des deux USE à z_{hi} dans (11). Posons que $z_{h1(j_a)} = \sum_{j\neq j_a} 2w_{h1j}z_{h1j}$ et $z_{k1(j_b)} = \sum_{j\neq j_b} 2w_{k1j}z_{k1j}$ représentent, respectivement, z_{hi} en excluant les contributions des USE qui seront permutées. En notant que $z^*_{h1} = z_{h1} - e_{hk}$, $z^*_{h2} = z_{h2}$, $z^*_{k1} = z_{k1} + e_{hk}$, il découle de (12) que

$$\begin{aligned} 4[v(\hat{\bar{Y}} \mid S^*) - v(\hat{\bar{Y}} \mid S)] &= (z^*_{h1} - z^*_{h2})^2 - (z_{h1} - z_{h2})^2 \\ &\quad + (z^*_{k1} - z^*_{k2})^2 - (z_{k1} - z_{k2})^2 \\ &= 2e_{hk}[e_{hk} - (z_{h1} - z_{h2}) + (z_{k1} - z_{k2})] \end{aligned}$$

et, donc, l'équation (13) est vérifiée avec $g_{hk} = 2^{-1}\{[z_{h2} - z_{h1(j_a)}] - [z_{k2} - z_{k1(j_b)}]\}$. Les preuves pour les trois autres cas sont semblables. Quand $h = k$, nous avons $e_{hh} = 2(w_{h1j_a}z_{h1j_a} - w_{h1j_b}z_{h1j_b})$, $z^*_{h1} = z_{h1} - e_{hh}$ et $z^*_{h2} = z_{h2} - e_{hh}$. La preuve est achevée en posant $g_{hh} = 2^{-1}\{[z_{h2(j_b)} - z_{h1(j_a)}] - [z_{h1(j_a)} - z_{h2(j_b)}]\} = z_{h1(j_b)} - z_{h2(j_b)}$.

Preuve de l'équation (17)

Par définition, nous avons $z_{hi} = z_{1,hi} + z_{2,hi}$ pour tout h et i. Donc, en observant que $z^\dagger_{h1} - z^\dagger_{h2} = z_{h1} - z_{h2} + 2(z_{1,h2-z_{2,h1}})$ et $z_{h1} - z_{h2} + z_{1,h2} - z_{2,h1} = z_{1,h1} - z_{2,h2}$, nous avons

$$\begin{aligned} v(\hat{\bar{Y}} \mid S^\dagger) &= \sum_{h=1}^{H}\left(\frac{z^\dagger_{h1} - z^\dagger_{h2}}{2}\right)^2 \\ &= \sum_{h=1}^{H}\left[\left(\frac{z_{h1} - z_{h2}}{2}\right) + (z_{1,h2} - z_{2,h1})\right]^2 \\ &= \sum_{h=1}^{H}\left(\frac{z^\dagger_{h1} - z^\dagger_{h2}}{2}\right)^2 \\ &\quad + \sum_{h=1}^{H}(z_{1,h2} - z_{2,h1})(z_{h1} - z_{h2} + z_{1,h2} - z_{2,h1}) \\ &= v(\hat{\bar{Y}} \mid S) + \sum_{h=1}^{H}(z_{1,h1} - z_{2,h2})(z_{1,h2} - z_{2,h1}), \end{aligned}$$

ce qui complète la preuve.

Remerciements

Ces travaux ont été réalisés pendant que l'auteur était à Westat, Inc., aux États-Unis. L'auteur remercie Leyla Mohadjer, Sylvia Dohrmann, Jill Montaquila et Lexter R. Curtin de leur appui pour cette étude. Il remercie aussi Barry Graubard, le rédacteur associé, et deux examinateurs anonymes de leurs commentaires constructifs qui l'ont aidé à améliorer l'article.

Bibliographie

Dalenius, T., et Peiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inferences*, 6, 73-85.

Dohrmann, S., Curtin, L.R., Mohadjer, L., Montaquila, J. et Le, T. (2002). National Health and Nutrition Examination Survey limiting the risk of data disclosure using replication techniques in variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 807-812.

Dohrmann, S., Lu, W., Park, I., Sitter, R. et Curtin, L.R. (2005). Variance estimation and data disclosure issues in the National Health and Nutrition Examination Survey limiting the risk of data disclosure using replication techniques in variance estimation. Soumis dans une revue.

Eltinge, J.L. (1999). Use of stratum mixing to reduce primary-unit-level identification risk in public-use survey datasets. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference*.

Fellegi, I.P., et Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Gomatam, S., Karr, A.F. et Sanil, A.P. (2005). Data swapping as a decision problem. *Journal of Official Statistics*, 21, 635-655.

Lu, W. (2004). Confidentiality and Variance Estimation in Complex Surveys. Thèse de doctorat non publiée, Simon Fraser University, Department of Statistics and Actuarial Science.

Lu, W., Brick, M.J. et Sitter, R.R. (2006). Algorithms for constructing combined strata grouped jackknife and balanced repeated replications with domains. *Journal of the American Statistical Association*, 101, 1680-1692.

Mayda, J.E., Mohl, C. et Tambay, J.-L. (1996). Variance estimation and confidentiality: They are related! *Proceedings of the Survey Methods Section*, *Statistical Society of Canada*, 135-141.

National Center for Health Statistics (1994). Data file document, National Health Interview Survey of Topics Related to the Year 2000 Health Objectives, 1993 (Fichier de données exploitable à la machine et documentation), National Center for Health Statistics, Hyattsville, Maryland.

Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L. et Curtin, L.R. (2006) Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1761-1767.

Park, I. (2004). Assessing complex sample designs via design effect decompositions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4135-4142.

Park, I., et Lee, H. (2004). Effets de plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage complexe. *Techniques d'enquête*, 30, 205-216.

Rust, K.F. (1986). Efficient replicated variance estimation. In the *Proceedings of the Survey Research Methodos Section*, American Statistical Association, 81-87.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Shah, B. (2001). A Method to Create Pseudo Strata and PSU's based on BRR weights. Manuscript non-publié, Research Triangle Institute.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer-Verlag.

Valliant, R. (1996). Limitations of balanced half-sampling. *Journal of Official Statistics*, 12, 225-240.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Predication Approach*. New York : John Wiley & Sons, Inc.

Yung, W. (1997). Variance estimation for public use files under confidentiality constraints. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 434-439.