Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Primary sampling unit (PSU) masking and variance estimation in complex surveys

by Inho Park

December 2008

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2008
Vol. 34, No. 2, pp. 183-194
Statistics Canada, Catalogue No. 12-001-X

PSU masking and variance estimation in complex surveys

Inho Park [1]

Abstract

The analysis of stratified multistage sample data requires the use of design information such as stratum and primary sampling unit (PSU) identifiers, or associated replicate weights, in variance estimation. In some public release data files, such design information is masked as an effort to avoid their disclosure risk and yet to allow the user to obtain valid variance estimation. For example, in area surveys with a limited number of PSUs, the original PSUs are split or/and recombined to construct pseudo-PSUs with swapped second or subsequent stage sampling units. Such PSU masking methods, however, obviously distort the clustering structure of the sample design, yielding biased variance estimates possibly with certain systematic patterns between two variance estimates from the unmasked and masked PSU identifiers. Some of the previous work observed patterns in the ratio of the masked and unmasked variance estimates when plotted against the unmasked design effect. This paper investigates the effect of PSU masking on variance estimates under cluster sampling regarding various aspects including the clustering structure and the degree of masking. Also, we seek a PSU masking strategy through swapping of subsequent stage sampling units that helps reduce the resulting biases of the variance estimates. For illustration, we used data from the National Health Interview Survey (NHIS) with some artificial modification. The proposed strategy performs very well in reducing the biases of variance estimates. Both theory and empirical results indicate that the effect of PSU masking on variance estimates is modest with minimal swapping of subsequent stage sampling units. The proposed masking strategy has been applied to the 2003-2004 National Health and Nutrition Examination Survey (NHANES) data release.

Key Words: Disclosure control; Stratified multistage sampling; Subsequent stage sampling unit swapping; Design effect; Intracluster correlation coefficient (ICC); Sample mean.

1. Introduction

The analysis of stratified multistage sample data requires the use of design information such as stratum and primary sampling unit (PSU) identifiers, or associated replicate weights, in variance estimation. In large surveys, PSUs often consist of single or multiple counties. Some external sources that are publicly available such as Census data can provide extremely detailed PSU-level demographics. Even with their name suppressed, inclusion of PSU identifiers in public release data files alone can pose an identification risk by allowing their linkage to external sources. Thus, PSU identifiers are often masked as an effort (1) to reduce the risk of data disclosure and (2) yet to allow the user to obtain valid variance estimation. Mayda, Mohl and Tambay (1996) addressed the potential risk of data disclosure that is associated with the inclusion of the original PSU identifiers in the public release data files and considered the stratum-collapsing method by Rust (1986) for balancing out the aforementioned two needs. Due to a potential inconsistency of the variance estimation under the stratum-collapsing method indicated by Valliant (1996), Yung (1997) suggested constructing a set of average bootstrap replicate weights. Lu (2004) demonstrated that supplying replicate weights and giving the stratum and PSU identifiers are practically equivalent in the viewpoint of confidentiality, since one can be easily obtained from the others. Shah (2001) discussed ways to create pseudo-strata and pseudo-PSUs given a set of balanced repeated replication weights. Eltinge (1999) proposed a method similar to the stratum-collapsing methods. Lu, Brick and Sitter (2006) also established conditions for the consistency of the variance estimator under the stratum-collapsing method and also proposed stratum-grouping algorithms yielding efficient and consistent stratum-collapsed variance estimators.

With a limited number of PSUs in the sample, the stratum-collapsing method is not appealing due to insufficient degrees of freedom for variance estimation. Dohrmann, Curtin, Mohadjer, Montaquila and Le (2002), Dohrmann, Lu, Park, Sitter and Curtin (2005) dealt with such situations and considered two PSU masking methods. The first method splits each PSU into two pseudo-PSUs (sets of ultimate sampling units within the PSUs), arbitrarily doubling degrees of freedom for variance estimation. The second method constructs the pseudo-PSUs by swapping second-stage sampling units (SSUs) between the original PSUs, retaining the original degrees of freedom for variance estimation. That is, the PSU and stratum assignments of all ultimate sampling units in one SSU are switched to those in the matched SSU. This method can be generalized so that the original PSUs are divided into one or more splits and are recombined to construct pseudo-PSUs with swapped PSU splits. This approach is different from data swapping (Dalenius and Reiss 1982), which is often used for

1. Inho Park, Statistician, Economic Statistics Department, The Bank of Korea, Namdaemun-Ro 106, Jung-Gu, Seoul 100-794, Korea. E-mail: ipark@bok.or.kr.

protecting confidentiality in a way that values of sensitive survey variables are switched among individual records. Because of the resulting distortion in the clustering structure of the sample design, the two PSU masking methods can result in biased variance estimates possibly with certain systematic patterns between two variance estimates from the unmasked and masked PSU identifiers. Dohrmann *et al.* (2005) observed decreasing funnel-shape curvature patterns in the ratio of the masked and unmasked variance estimates of a sample mean when plotted against the design effect. They explained such patterns based on an approximate relationship of the variance estimate that is monotone in the intracluster correlation coefficient (ICC) of Kish's design effect formula.

This paper focuses on the issues related to the second PSU masking method that swaps subsequent stage sampling units among the original PSUs and discusses its effect on variance estimates. Section 2 deals with the effect of PSU masking on the variance regarding aspects of the clustering structure such as ICC and means and sizes of PSU-splits for swapping under a single-stage cluster sample design. Section 3 investigates how the degree of swapping in PSU masking is related to the bias in the variance utilizing a parametric model for cluster sampling. Section 4 considers a PSU masking strategy through SSU swapping that helps reduce the PSU masking effects on variance estimates and thus the resulting biases under complex surveys. Section 5 briefly reviews the recent work by Dohrmann *et al.* (2002, 2005) and also presents application results of the proposed masking strategy to data from the National Health Interview Survey (NHIS) with some artificial modification. Finally, Section 6 includes some discussions.

2. Effect of distortion in clustering structure on variance of sample mean

Cluster sampling, often used in surveys for its cost and logistic reasons, is a major source of the increase in the variance of an estimator compared with a simple random sample due to the similarity of sampling units within the clusters. Standard sampling texts such as Särndal, Swensson and Wretman (1992, Section 8.7) provide formulae for the variance of a sample mean in terms of the ICC, cluster sizes and means of a survey variable y. It indicates that clustering in the sample design should reveal its impact on the variance through them. In this section, we examine how the distortion in the clustering structure of the sample design affects the variance of a sample mean when the PSUs are masked through swapping their splits between the two PSUs. For our discussion in this section, we consider a single-stage probability-proportional-to-size (PPS) sampling of PSUs. This sampling scheme is rather simple but still complex enough to reveal the effect of PSU masking on the variance in relation to these three components.

2.1 Variance under single-stage PPS cluster sampling

Suppose that a population U of M units is grouped into N PSUs of M_i units each. A random sample of n PSUs is drawn with probabilities p_i ($\sum_{i=1}^{N} p_i = 1$) and every unit in a sampled PSU is included in the sample. For simplicity, we assume the selection of PSUs is with replacement. The weighted sample mean $\hat{\bar{Y}} = (\sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij})^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij} y_{ij}$ is an estimator of the population mean $\bar{Y} = M^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$ of survey variable y, where $w_{ij} = (np_i)^{-1}$ and y_{ij} denote the sampling weight and the value of y for the j^{th} unit of PSU i, respectively. Let $m = \sum_{i=1}^{n} M_i$, $S_y^2 = (M-1)^{-1}$ $\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y})^2$, $\bar{Y}_i = M_i^{-1}\sum_{j=1}^{M_i} y_{ij}$ denote the sample size, the population variance and the PSU means of y, respectively. Assuming N is large so that $N/(N-1) \doteq 1$, its approximate variance can be written as

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

where S denotes the sample index set, $\rho_{yU} = 1 - S_{yw}^2/S_y^2$ is the ICC and $S_{yw}^2 = (M-N)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2$ is the within-PSU mean square deviation. The derivation of (1) is given in the Appendix.

For a common special case of $p_i \propto M_i$, that is, PPS sampling, (1) is simplified as

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

and the ICC is expressed as

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

The second approximation in (2) indicates that PSUs with larger $M_i(\bar{Y}_i - \bar{Y})^2$ contribute more to the variance. The ICC in (3) reveals the precision loss (in a rough sense) of per-cluster relative increase in the variance of $m^{-1}S_y^2$, the variance of the simple sample mean $\bar{y} = m^{-1}\sum_{(ij)\in S} y_{ij}$ that could have been obtained from the same sized with-replacement simple random samples.

A complex survey often involves the above single-stage PPS sampling or other (additional) complex design (*e.g.*, stratification, multi-stage sampling and unequal selection probabilities) or estimation features (*e.g.*, nonresponse

adjustments and calibration adjustments). For example, if p_i had been disproportional to size or further subsampling had been involved to induce unequal weights, then the corresponding complex feature might have come into the picture in variance estimation. The associated impact on variance estimation of complex samples will be discussed in detail in Section 4.

2.2 Means and sizes of PSU-splits

To mask the PSUs, consider that the first two PSUs in the sample are each split into two sets of units, $U_1 = U_{11} \cup U_{12}$ and $U_2 = U_{21} \cup U_{22}$ say, and the two pseudo-PSUs, $U_1^* = U_{11} \cup U_{22}$ and $U_2^* = U_{21} \cup U_{12}$ are constructed by swapping U_{12} and U_{22} between the two PSUs. Let S and S^* denote the unmasked and masked sample index sets, respectively. Let $V(\hat{\bar{Y}} \mid S^*)$ denote the variance of $\hat{\bar{Y}}$ associated with the pseudo-PSUs (also assuming the other non-sampled PSUs in U remain the same). Also, let $\bar{Y}_i^*$, Y_i^*, M_i^* denote the mean and total of y and the size of the i^{th} pseudo-PSU, respectively. Assuming $\bar{Y} = 0$ without loss of generality, the difference between the masked and unmasked variances is written from (2) as

$$V(\hat{\bar{Y}} \mid S^*) - V(\hat{\bar{Y}} \mid S)$$

$$\doteq (mN)^{-1} \sum_{i=1,2} (M_i^* \bar{Y}_i^{*2} - M_i \bar{Y}_i^2),$$

$$\doteq (mN)^{-1} \sum_{i=1,2} \left(\sqrt{M_i^*}\bar{Y}_i^* + \sqrt{M_i}\bar{Y}_i \right) \left(\sqrt{M_i^*}\bar{Y}_i^* - \sqrt{M_i}\bar{Y}_i \right). \quad (4)$$

Expression (4) shows that the difference in variance due to PSU masking depends upon the changes in PSU quantities $\sqrt{M_i}\bar{Y}_i$'s. If Y_{il} and M_{il} denote the total and size of PSU split U_{il}, respectively, for $i, l = 1, 2$, then $Y_i = Y_{i1} + Y_{i2}$, $Y_i^* = Y_{i1} + Y_{i'2}$, $M_i = M_{i1} + M_{i2}$ and $M_i^* = M_{i1} + M_{i'2}$ for $i \neq i' = 1, 2$. It is clear from (4) that the variance will not change under PSU masking if the following condition holds:

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

This result is a bit surprising since by naive intuition one may think that PSU-splits for swapping with $\bar{Y}_i = \bar{Y}_i^*$ will preserve the variance. To better understand (5), consider the following three cases. If $M_{12} = M_{22}$, then (5) implies $Y_{12} = Y_{22}$ or $\bar{Y}_{12} = \bar{Y}_{22}$, where $\bar{Y}_{il} = Y_{il} / M_{il}$ denotes the mean of U_{il}. If $Y_{12} = Y_{22}$, then (5) implies $M_{12} = M_{22}$ or $\bar{Y}_{12} = \bar{Y}_{22}$. If $\bar{Y}_{12} = \bar{Y}_{22}$, then (5) can be written as $M_{i1}\bar{Y}_{i1}(\sqrt{M_i^*/M_i} - 1) + \bar{Y}_{i2}(M_{i2}\sqrt{M_i^*/M_i} - M_{i'2}) = 0$ for $i = 1, 2$, holding when $M_{12} = M_{22}$. It is clearly demonstrated from all of the three cases that the variance will not change if the PSU-splits for swapping are formed equal in both size and mean.

2.3 Change in ICC

The effect of the clustering structure distortion on variance can also be investigated through the ratio of the masked and unmasked variances. Let ρ_{yU}^* denote the masked ICC, that is, the ICC defined with the masked PSU identifiers. From (3), it is clear that the difference between the masked and unmasked ICCs is proportional to the difference between the corresponding variances, that is,

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

The second approximation in (2) indicates that the change in ICC depends upon how the PSU-splits are formed for swapping, that is, the change in $M_i(\bar{Y}_i - \bar{Y})^2$. From (2), the ratio of the masked and unmasked variances is given as

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

See, also, Dohrmann *et al.* (2005, equation 8). Under the relationship of $\rho_{yU}^* = c_y \rho_{yU}$ for any given $c_y > 0$, (6) is monotone in ρ_{yU}. Also, the ratio is very unstable when ρ_{yU} or ρ_{yU}^* is near $-(\bar{M} - 1)^{-1}$, the lower bound of the ICC, because both numerator and denominator with their ICCs being near the lower bound are all close to zero. It indicates that any variable of such kind will be greatly influenced by PSU masking.

In general, surveys collect more than one variable and thus PSU masking based on one variable may not preserve well the ICC and thus not the variance of other variables. To better understand such an aspect, consider situations where the PSU masking results in both fixed and random distortion of the ICC written as $\rho_{yU}^* = c_y \rho_{yU} + e$ for $-0.02 < \rho_y < 0.2$, where $c_y = 0.7, 1.0, 1.3$, $e \sim N(0, 0.05^2)$ and $m = \bar{M} = S_y^2 = 100$. The constant coefficient c_y and the error term e in the model, respectively, allow deterministic and random perturbation in the ICC of the corresponding variable due to masking. Figure 1 displays the resulting ratio of the masked and unmasked standard errors (square-root of variances) against the ICC of the sample design. Three scatter plots in Figure 1 are all similar in their funnel shape with a wide variation for very small ρ_{yU}. However, their generic patterns depend on the magnitude of c_y. For example, $c_y < 1$ produces a decreasing pattern, $c_y > 1$ an increasing pattern, and $c_y = 1$ a non-monotonic pattern, respectively. As will be discussed in Section 3.2, the case of $c_y > 1$ may rarely occur.

The above discussion may not be extended straightforwardly to other complex survey situations, mainly because surveys often involve complex sample design features

such as stratification, three or higher-stage selection and unequal probability sampling. Under such circumstances, the ICC may not be easily defined and the variance may not be approximated well in the form of (2) (see, *e.g.*, Park (2004) and references cited therein). Nonetheless, the discussion in this section is still helpful to understand the effect of PSU masking on variance estimates in general.

3. Effect of degree of PSU masking on variance of sample mean

The more the clustering structure is distorted, the larger the bias in variance estimation. To study such a relationship, we consider a parametric model used for two-stage sampling. Suppose that two-stage sampling selects n PSUs and m units within each sampled PSU. Following Valliant, Dorfman and Royall (2000, page 248), we assume a sampled value y_{ij} for the j^{th} unit of PSU i is generated from the following model:

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

where μ_{yi}, σ_{yi}^2 and ρ_{yi} are the mean, variance and correlation of units within PSU i, respectively. The variance of a sample mean $\bar{y}_{2st} = (nm)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$ is written as

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

where $\sigma_{yu}^2 = n^{-1}\sum_{i=1}^{n}\sigma_{yi}^2$, $\rho_{yu} = \sum_{i=1}^{n}(\sigma_{yi}^2\rho_{yi})/\sum_{i=1}^{n}\sigma_{yi}^2$ and S denotes the sample index set. Note that ρ_{yu} can be interpreted as the (pooled or σ_{yi}^2 – weighted) ICC under the model ξ.

Let β denote the relative size of PSU splits to be swapped between the PSUs i_1 and i_2. For simplicity, we assume $m\beta$ to be an integer value, which is the number of units in each split for swapping. Let S^* denote the masked sample index set. The variance of $\bar{y}_{2st}$ with S^* can be written as

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

for $\gamma = \beta^2 + (1-\beta)^2$. The proof of (8) is given in the Appendix. Note that $-1 < \gamma - 1 < 0$ for $0 < \beta < 1$. The ratio of the masked and unmasked variances is written as

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

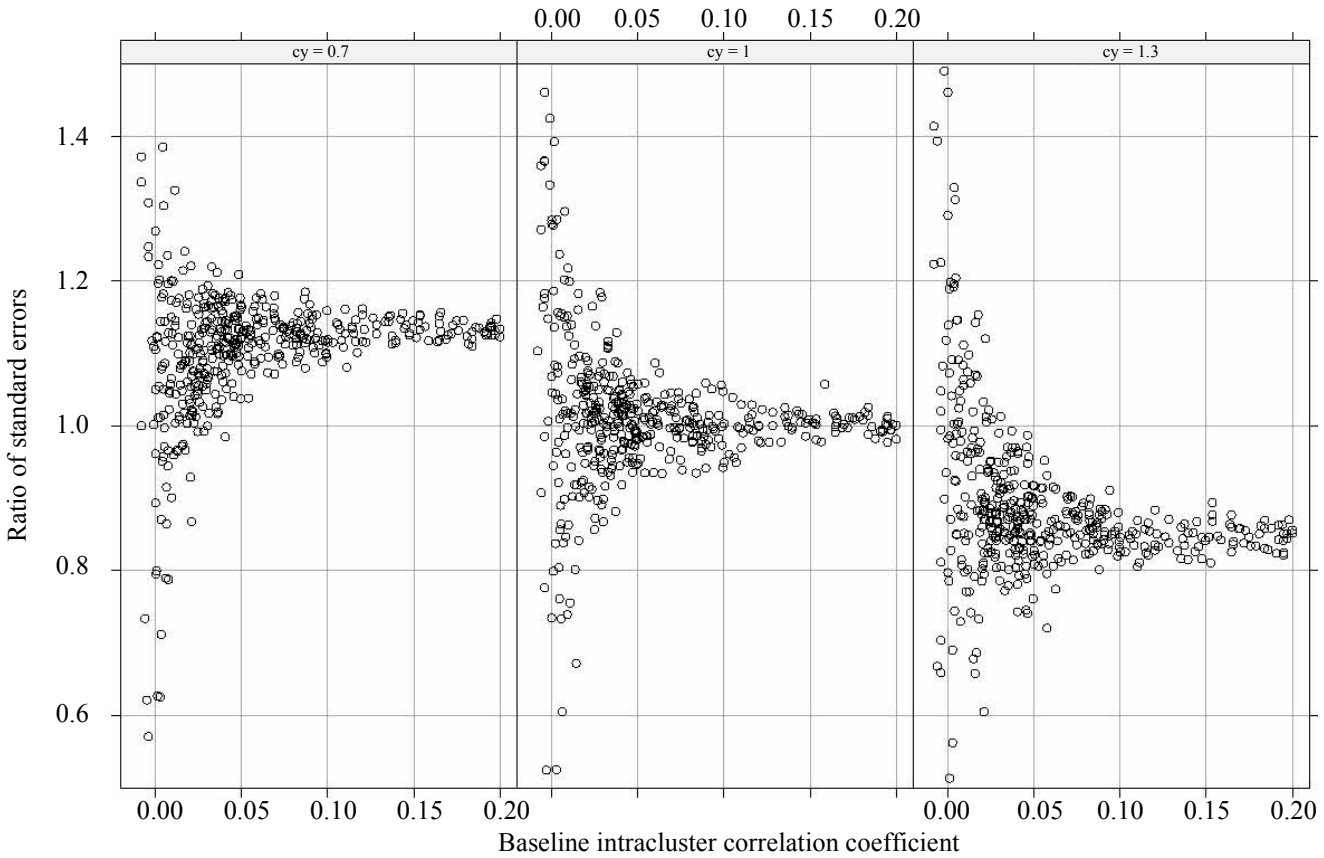

Figure 1 Ratios of the masked and unmasked standard errors against original intracluster correlation coefficient with varying the effect of PSU masking under a model $\rho_{yU}^* = c_y\rho_{yU} + e$ with $e \sim N(0, 0.05^2)$ for $c_y = 0.7, 1.0, 1.3$ and $m = \bar{M} = S_y^2 = 100$

The variance will not change if swapping is done such that $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} = 0$, that is, the correlations within the corresponding PSU being opposite in their direction. Otherwise, the change in variance will be at the rate of $m(\gamma - 1) < 0$ for $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} \neq 0$.

In general, units tend to be more similar within a PSU than across PSUs with ρ_{yi} being small and positive in many populations (*e.g.*, Valliant *et al.* 2000, Section 8.2.3). Thus, it is more likely that $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} > 0$ unless $\sigma^2_{yi} \doteq 0$ for all $i = i_1, i_2$ and the masked variance is prone to be smaller than the unmasked variance, that is, $RV_\xi(\bar{y}_{2st} \mid S, S^*) < 1$. Figure 2 depicts the change in standard error against the unmasked (or baseline) ICC ρ_{yu} with varying the proportion of units to be swapped between the two PSUs. Figure 2 shows that the more units that are swapped, the more the variance is changed, indicating that minimal swapping (*i.e.*, PSU masking) should be done in order to not induce serious bias in the variance. Also, Figure 2 exhibits the L-shape decreasing pattern of the standard error ratio in the ICC, that is, indicating overestimation for negative ICCs but underestimation for positive ICCs. Therefore, under PSU masking, we can expect patterns of either kind $c_y = 0.7$ (decreasing but random) or $c_y = 1.0$ (pure random) in Figure 1, with the latter being the best results attainable with minimal masking. In Section 4, we propose a PSU masking strategy through SSU swapping that helps produce a pattern of the second kind in the resulting variance ratios. In Section 5, we apply the proposed strategy to artificial survey data with varying proportions of swapping.

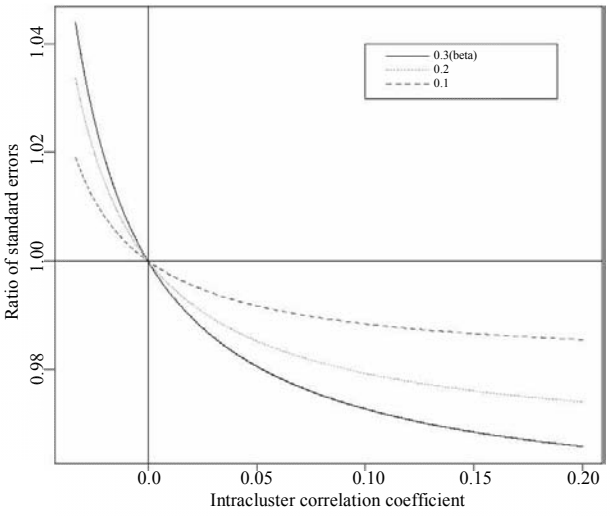

Figure 2 Ratios of the masked and unmasked standard errors $\sqrt{RV_\xi(\bar{y}_{2st} \mid S, S^*)}$ against the ICC with varying the proportion of swapping units from each PSU. $\sigma^2_{yu} \equiv \sigma^2_{yi} = 25, \rho_{yu} \equiv \rho_{yi}$ for all $i, n = 10$, $m = 16$, for $\beta = (0.1, 0.2, 0.3)$ and $-0.5(m-1)^{-1} \leq \rho_{yu} \leq 0.2$

4. PSU masking strategy for limiting biases in variance estimation

Many large-scale surveys involve several stages of sampling with unequal selection probabilities. Under such circumstances, the second stage or subsequent stage sampling units can be a natural choice for swapping to create pseudo-PSUs for operational reasons. For example, in the recent data releases of the National Health and Nutrition Examination Survey (NHANES) (Dohrmann *et al.* 2005) are included the pseudo-PSU identifiers constructed by swapping SSUs between the original PSUs. In this section, we consider SSU swapping for the purpose of PSU masking under stratified multi-stage sampling and their effect on variance estimates. We suggest a SSU swapping strategy based on the contribution of SSUs to variance estimates.

4.1 SSUs in variance estimation under stratified multistage sampling

Suppose that a finite population U of M units is partitioned into N PSUs and similar PSUs in a number of characteristics are grouped to form a total of H strata. Suppose also that each stratum consists of N_h PSUs and each PSU contains N_{hi} SSUs with N_{hij} ultimate sampling units, where $N = \sum_{h=1}^{H} N_h$ and $M = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{N_{hi}} N_{hij}$. Assume that the first stage sampling selects $n_h = 2$ PSUs within each stratum independently across strata and the second stage and subsequent stage sampling select, in turn, n_{hi} SSUs within each sampled PSU (hi) and n_{hij} ultimate units within each sampled SSU (hij), where $h = 1, ..., H$, $i = 1, ..., n_h$ and $j = 1, ..., n_{hi}$. Associated with the sampled ultimate unit $(hijk) \in S$ is the observed value y_{hijk} of survey variable y and the sample weight w_{hijk}, where $k = 1, ..., n_{hij}$ and S denotes the sample index set. The population total $Y = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{N_{hi}}\sum_{k=1}^{N_{hij}} y_{hijk}$ and size M are estimated by $\hat{Y} = \sum_{(hijk)\in S} w_{hijk} y_{hijk}$ and $\hat{M} = \sum_{(hijk)\in S} w_{hijk}$, respectively. Also, the population mean $\bar{Y} = Y/M$ is estimated by $\hat{\bar{Y}} = \hat{Y}/\hat{M}$ and its Taylor series variance estimator (*e.g.*, Shao and Tu 1995) is given by

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

where $z_{hi} = z_{hi}(y) = \sum_{j=1}^{n_{hi}}\sum_{k=1}^{n_{hij}} 2 w_{hijk} z_{hijk}$ are the estimated stratum totals of $z_{hijk} = z_{hijk}(y) = \hat{M}^{-1}(y_{hijk} - \hat{\bar{Y}})$ for PSU (hi).

Writing z_{hi} in (10) in the units of SSUs, we can see SSUs' contribution to the variance estimate, thus helping find better SSU swapping strategies to limit biases in the variance estimates. If w_{hij} and $w_{k|hij}$ denote the SSU sampling weights and the conditional ultimate sampling unit weights, respectively, then $w_{hijk} = w_{hij} \times w_{k|hij}$. Let $\hat{N}_{hij} = \sum_{k=1}^{n_{hij}} w_{k|hij}$ and $\hat{\bar{Y}}_{hij} = \hat{N}_{hij}^{-1}\sum_{k=1}^{n_{hij}} w_{k|hij} y_{hijk}$ denote the estimated

size and sample mean of SSU (hij), respectively. The quantities z_{hi} in (10) can now be written as

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

where

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

It is clear from (10) and (11) that the contribution of the sampled SSUs to the variance estimate is through three components $\{w_{hij}, \hat{N}_{hij}, \hat{\bar{Y}}_{hij}\}$ of SSU (hij). In Section 4.2, we will examine closely the effect of PSU masking on variance estimates through SSU swapping.

4.2 Effect of SSU swapping on variance estimates

We now assume that two SSUs $(h_a\, i_a\, j_a)$ and $(h_b\, i_b\, j_b)$ are to be swapped between two PSUs $(h_a\, i_a) \neq (h_b\, i_b)$. Then, the masked variance estimate can be written from (10) as

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

where z^*_{hi} denotes the quantity z_{hi} in (11) with the sample index set S^* altered due to swapping. Let i'_a and i_b denote the other PSUs in strata h_a and h_b, respectively, and define $z_{hi(j)} = z_{hi} - 2w_{hij} z_{hij} = \sum_{l \neq j} 2w_{hil} z_{hil}$. Then, (12) can be written as

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

where $e_0(y) = 2(w_{h_a i_a j_a} z_{h_a i_a j_a} - w_{h_b i_b j_b} z_{h_b i_b j_b})$ is the difference in the quantity $2w_{hij} z_{hij} = 2w_{hij} \hat{M}^{-1} (\hat{\bar{Y}}_{hij} - \hat{\bar{Y}})$ of the two SSUs to be swapped and

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

is a function of $2w_{hij} z_{hij}$ of the SSUs to be retained in the original PSUs. Note that, for $h_a = h_b$, r_0 can also be expressed $g_0 = 2^{-1}[(z_{h_b i_b (j_b)} - z_{h_a i_a (j_a)}) - (z_{h_a i_a (j_a)} - z_{h_b i_b (j_b)})]$. It shows that the effect of SSU swapping on the variance estimate will be negligible if the two SSUs for swapping are paired in such a way that the product of $e_0(y)$ and $g_0(y)$ is close to zero. In other words, the change in the variance estimate under SSU swapping can be controlled when a segment pair is formed taking into account all three components $\{w_{hij}, \hat{N}_{hij}, \hat{\bar{Y}}_{hij}\}$ so as to minimize $e_0(y) \times g_0(y)$ as similar to the case under single-stage PPS cluster sampling in Section 2.2.

In addition, by writing

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

(13) can be expressed as a quadratic function of $e_0(y)$ for given $\{z_{hi} : h = \{h_a, h_b\}, i = 1, 2\}$. For $h_a = h_b$, we can show that $v(\hat{\bar{Y}} \mid S^*) > v(\hat{\bar{Y}} \mid S)$ only for e_0 in between zero and $(z_{h_a 1} - z_{h_b 2})$. When $z_{h_a 1} - z_{h_b 2} \doteq 0$, it may be more likely that $v(\hat{\bar{Y}} \mid S^*) = v(\hat{\bar{Y}} \mid S)$. Similar arguments can be made for $h_a \neq h_b$.

4.3 Sequential SSU swapping with multiple matching characteristics

Suppose that there are a total of n_J SSUs in the sample and only R of them are chosen to form pairs for swapping, where $n_J = \sum_{h=1}^{H} \sum_{i=1}^{n_h} n_{hi}$ and $1 \leq R < n_J$. Assume that a fixed number of R SSUs is chosen in accordance with a certain data risk-utility tradeoff consideration. See, for example, Gomatam, Karr and Sanil (2005) for some related discussion concerning data swapping. In addition, assume that their sequential order for the matching process is given as $j_1, j_2, \ldots, j_R$ say. For example, at first, all possible pairs are formed for each of the R SSUs and the best pair is picked based on a certain distance measure such as (12). The order of the R SSUs for the (main) matching process is then determined according to the ascending order of the distances of the R best pairs.

Let S^{r-1} denote the altered sample index set after the $(r-1)^{\text{th}}$ SSU pair has been formed and swapped, where $r = 1, \ldots, R$ and $S^0 \equiv S$. Let $S^{r-1}_{(j)}$ denote the sample index set with SSUs j_r and j being swapped for S^{r-1}. Then, the change in the variance estimate caused by swapping the r^{th} SSU j_r and any other SSU that was not involved in the $(r-1)$ previous match(es) can be written as

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

where $e_{r-1}(y, r)$ and $g_{r-1}(y, r)$ are defined similarly as in (13) but with S^{r-1} and $S^{r-1}_{(j)}$. Clearly, the choice of the best match for the r^{th} SSU depends on the $(r-1)$ previous match(es) and thus the matching process should be viewed as a sequential process. Note that those SSUs that were matched and swapped in the previous match(es) should be excluded in the current search.

In addition, more than one characteristic can be considered for matching, with the hope that they will be related to many other survey variables so as to minimize the bias in the associated variance estimate. Suppose that q

matching characteristics are chosen with care, say $\mathbf{x} = (x_1, x_2, ..., x_q)'$ (*e.g.*, Dohrmann *et al.* 2005, for some related discussion). To measure the distance between SSUs j_r and j, any distance measure of the form

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

or

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

can be considered with any reasonable choice of positive coefficients c_l. For example, $c_l \equiv 1$ simply considers the absolute difference in the variance estimates of $\hat{X}_l$, $c_l = v(\hat{\bar{X}}_l \mid S)^{-1}$ the absolute difference in variance estimates relative to the original variance estimates, $c_l = X_l^{-1}$ the absolute difference in relative variance estimates. The first distance measure (15) considers the change in the variance estimate due to swapping segments of the r^{th} pair. The second distance measure (16) takes into account the cumulative swapping effect of all the r segment pairs.

Matching constraints can be set, for example, to prohibit the pairing of SSUs from the same PSU and to apply a threshold of the proportion of SSUs from each PSU to be swapped (Lu 2004). Let $J_A = \{j_1, ..., j_R\}$ denote the index set of R SSUs that are considered for forming swapping pairs and let J_B denote all possible SSUs that can be matched satisfying a given set of matching constraints. For simplicity, consider that the pairing of SSUs is not allowed within J_A, that is, $J_A \cap J_B = \varnothing$. If $D_r^*(j \mid \mathbf{x})$ denotes the chosen distance measure for SSUs j_r and j, then a sequential SSU swapping algorithm for limiting the biases of variance estimates can be given as follows:

Step 1. Set $r = 1$, $J_A^r = J_A$ and $J_B^r = J_B$;

Step 2. For each of the $(R - r + 1)$ SSUs in J_A^r, compute $D_r^*(\cdot \mid \mathbf{x})$ for all SSUs in J_B^r;

Step 3. Choose the best match with the smallest $D_r^*(\cdot \mid \mathbf{x})$, that is, find j_r' such that $D_r^*(j_r' \mid \mathbf{x}) = \min_{j \in J_B^r} D_r^*(j \mid \mathbf{x})$;

Step 4. Set $r = r + 1$, and drop the chosen pair from the searching pool, that is, set $J_A^r = J_A^{r-1} \setminus \{j_r\}$ and update J_B^r accordingly, where $J_B^r \subseteq J_B^{r-1} \setminus \{j_r'\}$;

Step 5. If $r = R + 1$, then stop; otherwise repeat Steps 2-4.

This SSU matching (or swapping) approach basically searches for the pair at the r^{th} matching that is best in a sense of minimizing the change in variance estimates due to the corresponding SSU swapping. With a large number of SSUs, this method will lead to a scatter plot similar to that of $c_y = 1.0$ in Figure 1 (*i.e.*, a random perturbation with a funnel-shape pattern).

A choice of more sophisticated optimality criterion applied to $\{c_l : l = 1, ..., q\}$ may help improve the above method to reduce the magnitude of such random perturbation in variance estimates. Also, if one uses multivariate techniques such as principal component analysis to develop some kind of scores (*e.g.*, one or more principal component axes) from a larger number of continuous characteristics, the magnitude of such random perturbations in the variance estimates may be further reduced. In Section 5, we give examples regarding SSU swapping.

5. Examples

5.1 Previous work

For a sample design with no stratification but a small number of PSUs, Dohrmann *et al.* (2002) considered various methods of splitting PSUs into pseudo-PSUs in order to use the delete-one jackknife variance estimation method. Their basic idea is to double the number of masked PSUs by keeping the split PSUs as separate masked PSUs, thus hoping to reduce data disclosure risk as a result of the broken linkage between the true and masked PSUs. In their empirical study, noticeable underestimation patterns were present for the resulting variance estimates for variables with large design effects, which resemble the plot of $c_y = 0.7$ in Figure 1. Let S and $S^\dagger$ denote the unmasked and masked sample index sets respectively. Let w_{ij} denote the sample weight and let y_{ij} denote the observed value of y for the j^{th} sampled unit in PSU i. To explain the observed underestimation patterns, Dohrmann *et al.* (2005) derived the following relationship

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

where $z_{g,i} = \sum_{j \in S_{g,i}} 2 w_{ij} z_{ij}$ are the PSU-split totals of $z_{ij} = (\sum_{ij} w_{ij})^{-1} (y_{ij} - \sum_{ij} w_{ij} y_{ij} / \sum_{ij} w_{ij})$ and $S_{u,i}$ are the index sets of the u^{th} split of PSU i for $i = 1, ..., n$ and $u = 1, 2$. It indicates that the resulting variance estimate is about a half of the unmasked one plus a positive value reflecting the between PSU-split totals of z_{ij} within the PSUs. If $S_{u,i}$ are formed such that $z_{1,i} \doteq z_{2,i}$, this PSU-splitting method leads to about a half of the unmasked variance estimate and thus the masked variance estimate could be doubled to get close to the unmasked value.

For the two-PSU-per-stratum design, Dohrmann *et al.* (2005) considered an alternative approach under which the pseudo-PSUs are constructed by swapping SSUs between the PSUs. As discussed in Section 1, this approach can be viewed as dividing the PSUs into one or more splits and recombining them to construct pseudo-PSUs with swapped PSU splits. For simplicity, we assume that each PSU is divided into two splits $S_{1,hi}$ and $S_{2,hi}$. If it is done so with

$S_{h1}^{\dagger} = S_{1,h1} \cup S_{1,h2}$ and $S_{h2}^{\dagger} = S_{2,h1} \cup S_{2,h2}$, then the masked variance estimate can be written as

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

where $e_h(y) = z_{1,h2} - z_{2,h1}$ is the difference between the PSU-split totals of $S_{1,h2}$ and $S_{2,h1}$ to be swapped and $g_h(y) = z_{1,h1} - z_{2,h2}$ is the difference between the PSU-split totals of $S_{1,h1}$ and $S_{2,h2}$ to be retained in the original PSUs. The proof of (17) is given in the Appendix. Equation (17) indicates that a similar strategy for splitting PSUs would help preserve the magnitude of the original variance estimate. Dohrmann *et al.* (2005) adopted a probability-based record linkage technique (Fellegi and Sunter 1969) to form pairs of SSUs for swapping that are similar in their means $\hat{\bar{Y}}_{hij}$ for several characteristics, with the hope that the terms e_h in (17) are all close to zero. Dohrmann *et al.* (2005) demonstrated that the PSU-recombination method can help reduce the biases of variance estimates and the resulting underestimation patterns to some degree as compared to the PSU-split method used in Dohrmann *et al.* (2002). To increase speed and flexibility, Lu (2004) developed SSU-swapping algorithms based on sequentially evaluating distance measures between SSU means without directly considering the impact of successive swapping on the bias of the variance estimate. As discussed in Section 4.2, the effect of SSU swapping on variance estimates can be further reduced by direct consideration of SSU's contribution to the variance estimates. In the next section, we apply both strategies, one by Dohrmann *et al.* (2005) and the other proposed in Section 4.2, to artificial data from a complex survey.

5.2 Data example

To illustrate the effect of PSU masking on variance estimates of sample means, we used real survey data from the 1993 National Health Interview Survey (NHIS) Year 2000 Health objectives Public Use Data File (PUF) with some artificial modification. The NHIS is an annual household health interview survey of the civilian non-institutionalized population of the United States. The NHIS involves a typical multistage, stratified sample design, with the first stage PSUs consisting of counties or metropolitan areas and the second stage SSUs consisting of segments (that is, a small number of households in a small geographic area) within sampled PSUs. This specific Year 2000 topic questionnaire was administered to one adult sample person per family only in the last half of 1993. The NHIS data used here and its documentation are available from National Center for Health Statistics (1994) or the United States Centers for Disease Control National Center for Health Statistics website (http://www.cdc.gov/nchs/about/major/nhis/quest_data_related_doc.htm).

This PUF contains the stratum and PSU identifiers, and sample person's final weights for the purpose of variance estimation. For our example, we used only ten strata but limited their number of PSUs to two per stratum. Two of the selected strata, 110 and 520, were restricted to their two largest PSUs, (181,410) and (048,233), respectively, and the other eight strata, 102, 142, 192, 211, 261, 300, 561 and 571 contain only two PSUs. The PUF also includes the SSU identifiers but not their sample weights. To generate SSU sample weights w_{hij}, we employed a two-way nested random effects model to fit $\log w_{hijk} = \log w_{hij} + \log w_{k|hij}$ such that $\log w_{hij} = \mu + \alpha_{hi} + \beta_{j(hi)}$ and $\log w_{k|hij} = \varepsilon_{k(hij)}$, where μ is a common value, α_{hi} is the random effect of PSU (hi), $\beta_{j(hi)}$ is the random effect of SSU j nested within PSU (hi) and $\varepsilon_{k(hij)}$ is the random effect of sampled person k within SSU (hij). We restricted our study to include only those SSUs with five or more sampled persons, giving a total of 293 SSUs in the analysis. The resulting weight decomposition $(w_{hij}, w_{k|hij})$ may involve possible model misspecification but it suffices our need for the illustration, since both w_{hij} and $w_{k|hij}$ are all positive under the model. To obtain SSU pairs for swapping, we used six socio-demographic variables, denoted as $x_1, x_2, \ldots, x_6$. They are listed in Table 1 with their description, definition, overall sample mean and squared root of design effect (*i.e.*, design factor or *deft* in short). The sample means of these variables range from 0.05 to 0.63 and the design factors from 1.285 to 8.511.

Table 1
Variables used for matching

Variable	Description	Definition	Sample mean	Design factor
x_1	Male	SEX=1	0.49	1.285
x_2	Hispanicity	HISPANIC	0.14	8.511
		= 00, 01, ..., 08	0.14	8.511
x_3	Married couple	MARSTAT = 1,2	0.63	3.209
x_4	College or higher education	EDUCR = 4, 5, 6	0.45	2.902
x_5	High family income of $50k or higher	INCFAMR = 8	0.23	3.558
x_6	Has household air been tested for Radon?	TESTRDN = 1	0.05	2.191

We applied the two SSU matching strategies discussed in Section 4 and Section 5.1, respectively. The first strategy employs a distance measure (15) for any SSU pair (r_a, r_b) with $c_l \equiv 1$ for all $l = 1, \ldots, 6$. Let S^{r-1} and S^{r} denote the two sample index sets after the $(r-1)^{\text{th}}$ and r^{th} swapping, respectively. Then the distance of the r^{th} matching pair of the first strategy (variance-matching) is written as

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

where $v(\hat{\bar{X}}_l|S^r)$ and $v(\hat{\bar{X}}_l|S^{r-1})$ represent the variance estimates of $\hat{\bar{X}}_l$ for the l^{th} matching characteristic with S^r and S^{r-1} respectively. The smaller the distance is, the smaller the biases of variance estimates arises from swapping the r^{th} matching pair. The second strategy by Dohrmann *et al.* (2005) is to pair SSUs that are similar in their sample means of the six matching characteristics. This strategy (mean-matching) defines the distance of the r^{th} matching pair as:

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

where $\hat{\bar{X}}_{l,r_i}$ represents the SSU sample mean of SSU r_i $(i = a,\ b)$ for matching characteristic x_l.

Table 2 lists standard error ratios of the six matching characteristics at each matching in the sequential order for each strategy with 18 swapping pairs (representing about 12% of the SSUs in the study). The first strategy, shown in the left panel of the table, gave a moderate but slightly increasing range of variations in standard error ratios over the sequence of the 18 swapping pairs. The second strategy, shown in the right panel of the table, produced a rather wider range of variation in standard error ratios over the sequence with its dramatic changes from the thirteenth and higher pairs in the swapping sequence. Although both strategies tend to lose their control over the biases in the variance estimates for higher orders of the swapping sequence, the first strategy was quite successful in controlling the biases of the variance estimates for a relatively large number of swapping pairs.

Figure 3 plots the standard error ratios against the design factors for the two strategies varying the number of SSUs swapped. These three sets included 6 (4%), 12 (8%) and 18 (12%) SSU pairs (percentage of SSUs involved in swapping), respectively. Each plot includes two sets of characteristics, 6 matching characteristics marked with the corresponding numbers as listed in Table 1 and 92 characteristics marked with × that are not used in matching. For the scenario with only 4% of the SSUs swapped, the difference between the two strategies is negligible for both sets of characteristics. However, as the percentage of SSUs swapped increases, the perturbation in the variance estimates becomes greater for both strategies and both sets of characteristics. This result indicates that a small percentage of swapping should occur, reinforcing the findings of Section 3. In addition, the standard error ratios are clustered more closely to the line of one (*i.e.*, small biases of the masked variance estimates) with the first strategy than with the second strategy. The second strategy produced a rather steeply decreasing pattern over the design factor even for the six matching characteristics. That is, the mean-matching strategy is seen more poignant for the variables used for matching.

Table 2
Standard error ratios by swapping sequence: Comparison of the two matching criteria with 12% (18 pairs) SSU swapping

	Variance-matching						Mean-matching					
Swapping Sequence	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

See Table 1 for the description of the six matching characteristics (x_1, ..., x_6).

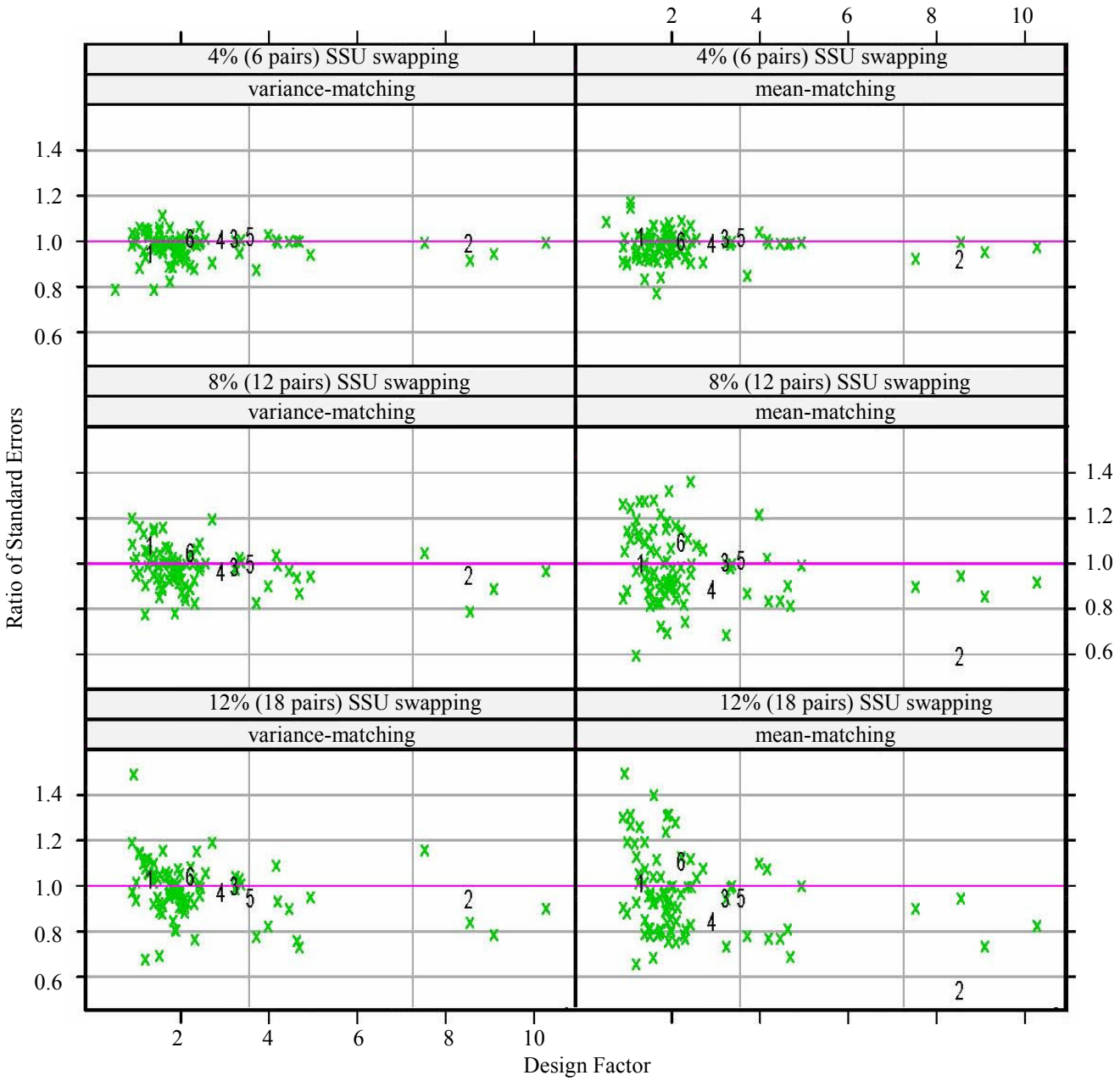

Figure 3 Ratio of Standard Errors *vs.* Baseline Design Factors. Six numbers represent the points of the corresponding matching characteristics and × marks represent those of 92 characteristics not used in matching

6. Discussion

In this paper, we investigated the effect of PSU masking on variance estimates in complex surveys. Obviously, PSU masking distorts the clustering structure of the original sample design, possibly yielding systematic biases in the analysis of the resulting data as seen in Sections 2, 3 and 5.2. The proposed PSU masking strategy in Section 4 can help reduce such biases but still leave a random perturbation in the variance estimation and thus a loss of inferential efficiency. Research on the effect of PSU masking would be interesting on other types of complex data analyses such as regression and multivariate analyses. Although PSU masking can provide disclosure control, the degree of masking should be minimal to limit the resulting biases of variance estimates as discussed in Sections 3 and 5.2.

In addition, the reduction of the identification risk incurred by SSU masking may be better understood by writing the distance between the masked sample PSU mean and the PSU mean in the population as follows:

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

where $\hat{\bar{Y}}_{hi|S^\dagger}$ and $\hat{\bar{Y}}_{hi|S}$ denote the masked and unmasked PSU means in the sample, respectively, and $\bar{Y}_{hi|U}$ denote the PSU mean in the population that may be available to an intruder (*i.e.*, a malicious data user) from external sources such as Census data. One can show easily that the first term in the right-hand side of (18) is not equal to zero, in general, with PSU masking. $\hat{\bar{Y}}_{hi|S^\dagger}$ and $\bar{Y}_{hi|U}$, together with non-neglible sample variation of the second term in the right-hand side of (18), are never equal except by rare chance. Dohrmann *et al.* (2005) compare $\{\hat{\bar{Y}}_{hi|S^\dagger}\}$ of the sample to $\{\bar{Y}_{hi|U}\}$ of the population by a stylish stem-and-leaf diagram to demonstrate how hard it would be for an intruder to identify a sampled PSU in the public release data files in association with two aspects: 1) few pairs of $(\hat{\bar{Y}}_{hi|S^\dagger}, \bar{Y}_{hi|U})$ being close to each other; and 2) many unsampled PSU's with population values similar to $\{\hat{\bar{Y}}_{hi|S^\dagger}\}$ or $\{\bar{Y}_{hi|U}\}$ of the sampled PSUs. Some forms of probabilistic measurements may be interesting to evaluate identification risk reduction (*e.g.*, Eltinge 1999) but are beyond the scope of this paper. The proposed masking strategy has been applied to the 2003-2004 National Health and Nutrition Examination Survey (NHANES) release (Park, Dohrmann, Montaquila, Mohadjer and Curtin 2006).

Appendix

Proofs

Proof of equation (1)

From Park and Lee (2004, Section 4.2),

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

where $\text{Deft}^2(\hat{\bar{Y}} \mid S)$ represents the design effect of $\hat{\bar{Y}}$ for a given S, the second and the last approximations follow from $(N-1)/N \doteq 1$ and the third equation from $S_y^2[1 + (\bar{M} - 1)\rho_{yU}] \doteq (N-1)^{-1}\sum_{i=1}^N M_i(\bar{Y}_i - \bar{Y})^2$, which completes the proof.

Proof of equation (8)

By definition, the variance of the sample PSU total $y_i = \sum_j y_{ij}$ is $V_\xi(y_i \mid S) = m\sigma_{yi}^2 + m(m-1)\sigma_{yi}^2\rho_{yi}$ for $i = 1, ..., n$. Suppose that $\{y_{ij}: j = m(\beta-1)+1, ..., n\}$ for the two PSUs $i = a$ and b are to be switched between the PSUs. Then these two PSUs have their variance changed to $V_\xi(y_a \mid S^*) = m(1-\beta)\sigma_a^2 + m(1-\beta)[m(1-\beta)-1]\sigma_a^2\rho_a + m\beta\sigma_b^2 + m\beta(m\beta-1)\sigma_b^2\rho_b$ and to $V_\xi(y_b \mid S^*)$ being the same with switching the indices a and b. Since $m(1-\beta)[m(1-\beta)-1] + m\beta(m\beta-1) = m(m\gamma-1)$, the proof is completed from observing

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

Proof of equation (13)

Suppose that two PSUs $(h1)$ and $(k1)$ from two different strata $h \neq k$ are to be reconstructed by swapping each of their SSUs, $(h1j_a)$ and $(k1j_b)$. Let $e_{hk} = 2(w_{h1j_a}y_{h1j_a} - w_{k1j_b}y_{k1j_b})$ denote the difference between the contributions of the two SSUs to z_{hi} in (11). Let $z_{h1(j_a)} = \sum_{j\neq j_a} 2w_{h1j}z_{h1j}$ and $z_{k1(j_b)} = \sum_{j\neq j_b} 2w_{k1j}z_{k1j}$ denote, respectively, z_{hi} excluding the contributions from the SSUs to be swapped. By noting that $z_{h1}^* = z_{h1} - e_{hk}$, $z_{h2}^* = z_{h2}$, $z_{k1}^* = z_{k1} + e_{hk}$, it follows from (12) that

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

and thus, (13) holds with $g_{hk} = 2^{-1}\{[z_{h2} - z_{h1(j_a)}] - [z_{k2} - z_{k1(j_b)}]\}$. The proofs for the other three cases are similar. When $h = k$, we have $e_{hh} = 2(w_{h1j_a}z_{h1j_a} - w_{h1j_b}z_{h1j_b})$, $z_{h1}^* = z_{h1} - e_{hh}$ and $z_{h2}^* = z_{h2} - e_{hh}$. The proof is completed by letting $g_{hh} = 2^{-1}\{[z_{h2(j_b)} - z_{h1(j_a)}] - [z_{h1(j_a)} - z_{h2(j_b)}]\} = z_{h1(j_b)} - z_{h2(j_b)}$.

Proof of equation (17)

By definition, we have $z_{hi} = z_{1,hi} + z_{2,hi}$ and $z_{yi}^\dagger = z_{i,hi} + z_{i,hi}$ for any h and i. Thus, observing $z_{h1}^\dagger - z_{h2}^\dagger = z_{h1} - z_{h2} + 2(z_{1,h2-z_{2,h1}})$ and $z_{h1} - z_{h2} + z_{1,h2} - z_{2,h1} = z_{1,h1} - z_{2,h2}$, we have

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

which completes the proof.

Acknowledgements

This work was done while the author was at Westat, Inc., U.S.A. The author thank Leyla Mohadjer, Sylvia Dohrmann, Jill Montaquila and Lexter R. Curtin for their support to this research. The author is also grateful to Barry Graubard, the associate editor and two anonymous reviewers for their helpful comments, which helped improve the paper.

References

Dalenius, T., and 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. and 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. and 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. Submitted to a journal.

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

Gomatam, S., Karr, A.F. and 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. Unpublished Ph.D. dissertation, Simon Fraser University, Department of Statistics and Actuarial Science.

Lu, W., Brick, M.J. and 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. and 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 (machine readable data file and documentation), National Center for Health Statistics, Hyattsville, Maryland.

Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L. and 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., and Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology*, 30, 183-193.

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. and 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. Unpublished manuscript, Research Triangle Institute.

Shao, J., and 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. and 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.