Reducing the Risk of Data Disclosure Through Area Masking: Limiting Biases in Variance Estimation

Inho Park[1], Sylvia Dohrmann[1], Jill Montaquila[1], Leyla Mohadjer[1] and Lester R. Curtin[2]
Westat, 1650 Research Blvd., Rockville, MD 20850[1]
National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782[2]

Keywords: Disclosure Control, Segment Swapping, Stratified Multistage Sampling, Weighted Mean, Design Effect, National Health and Nutrition Examination Survey

1. Introduction

The National Health and Nutrition Examination Survey (NHANES) is one of a series of health related programs conducted by the National Center for Health Statistics (NCHS). A unique feature of NHANES is the collection of health data by means of medical examinations carried out for a nationally representative sample of the U.S. population. Beginning in 1999, NHANES has been implemented as a continuous, annual survey. Each single year and any combination of consecutive years comprise a nationally representative sample of the U.S. population. A four-stage sample is selected for NHANES. The first stage of selection is the primary sampling unit (PSU). Within each of the selected PSUs, an average of 24 secondary sampling units (SSUs), or "segments," consisting of census blocks or groups of census blocks are selected. Within each sampled segment, a subsample of the households is selected and screened. Within the screened households, members of particular race/ethnicity-income-sex-age subdomains are selected with pre-specified probabilities.

The practical constraints surrounding the collection of medical data in mobile examination units have limited the NHANES survey to about 15 PSUs in each annual sample. The small number of PSUs in the sample poses a risk for data disclosure. To improve the precision of the published results and to reduce data disclosure risks, NCHS currently prepares public-use files (PUFs) for two-year samples rather than annual samples. The original PSU identifiers are not included on these files.

The first data release of the continuous NHANES survey included the combined 1999-2000 annual samples. The PSU identification in light of the minimal geographic data, other characteristics of the area on the data files, and local publicity campaigns led to concerns about disclosure risks in the release of the NHANES 1999-2000 data file. As a result, NCHS initiated research to examine the disclosure risks of NHANES before the release of these data. The alternative approaches considered for creating variance estimation replicates or pseudo-PSU identifiers that would mask the original PSUs were presented in Dohrmann *et al.* (2002). The decision was to split each of the original PSUs into two pseudo-PSUs in the release of the NHANES 1999-2000 data (Section 2.1).

After the NHANES 1999-2000 data release, there was a need to change the basic methodology used for variance estimation. Beginning in 2002, NHANES is a stratified design, with two-PSUs per stratum for the two-year samples. Given this, and the great number of replicates needed for continual two-year releases, NCHS decided that in future data releases only PSU and stratum identifiers will be released for variance estimation. As a result, a new approach to variance estimation aimed at limiting disclosure risk had to be developed for use with the publicly released data. Investigation of this issue was continued toward improving the methodology for the 2001-2002 data release and the results were presented in Dohrmann *et al.* (2004). Under the adopted method, the pseudo-PSUs are constructed by swapping segments that are similar in a number of characteristics between the original PSUs.

These techniques can reduce the chance of an intruder being able to match PSUs in the sample to PSUs in the population (census or external source files) by blurring the actual composition of the PSUs in the PUF. Let U denote the population and let S and S' denote the sample index sets with the unmasked PSUs and masked PSUs, respectively. For a given characteristic y, the difference between the masked PSU sample mean in the PUF and the PSU mean in the population can be written as

$$\begin{aligned}\bar{y}_{hi|S'} - \bar{y}_{hi|U} &= \left(\bar{y}_{hi|S'} - \bar{y}_{hi|S}\right) + \left(\bar{y}_{hi|S} - \bar{y}_{hi|U}\right) \\ &= \left(\text{masking error}\right) + \left(\text{sampling error}\right),\end{aligned} \quad (1)$$

where $\bar{y}_{hi|S'}$ and $\bar{y}_{hi|S}$ denote the masked and unmasked PSU means of the i th PSU within stratum h in the sample, and $\bar{y}_{hi|U}$ denotes the PSU mean for the corresponding PSU in the population (or census/auxiliary files available to the intruder). Such masking techniques would change PSU means in the sample (that is, $\bar{y}_{hi|S'} \neq \bar{y}_{hi|S}$) for PSUs involved in masking. In addition, the masked PSUs no longer are completely associated with a single real PSU, thus limiting the chance of correctly matching a given individual with the PSU. We should note that the point estimate of the population mean will not change under the PSU masking (that is, $\bar{y}_S = \bar{y}_{S'}$).

However, one of the challenges with these techniques is that the two associated variance estimates are not equal in general. That is, $v(\bar{y} \mid S) \neq v(\bar{y} \mid S')$, and depending on the masking approach used, we have observed some patterns in the resulting biases when plotting against the (original or unmasked) design effects, where $v(\bar{y} \mid S)$ and $v(\bar{y} \mid S')$ denote, respectively, the variance estimates with the unmasked and masked PSUs.

This paper discusses an improved PSU masking strategy adopted for the recent release of NHANES 2003-2004 data that helps to limit such biases. Section 2 presents a brief overview of the PSU-splitting and recombining methods used for the 1999-2000 and 2001-2002 NHANES data releases, respectively, as well as some challenges of those methods related to the variance estimation. Section 3 describes the new masking strategy and some matching procedures considered for creating alternative sets of PSU and stratum identifiers for variance estimation. A comparison of the variance estimates from each of the strategies considered, and a discussion of how the new method is an improvement over the previous ones adopted for NHANES is presented in Section 4. Section 5 gives concluding remarks.

2. PSU Masking and Bias Issues in Variance Estimation in Previous NHANES Data Releases

2.1 PSU-Splitting in 1999-2000 (Method 1)

As a result of an integrated survey plan adopted by NCHS, the NHANES 1999-2001 design was linked at the PSU level with the National Health Interview Survey (Hunter and Arnett, 1996). Because the sampling frame for 1999-2001 became the already selected NHIS areas, no explicit stratification was used to select the NHANES PSUs. With the first stage design structure, and due to the small number of PSUs in the sample, the initial decision was to use the true PSUs along with a delete-1 jackknife method to create replicates for variance estimation for the analysis of the NHANES 1999-2000 data.

For the purpose of disclosure limitation discussed in Section 1, various PSU splitting methods were considered to split each PSU into two dissimilar pseudo-PSUs, creating a total of 52 pseudo-PSUs. The associated impact on the performance of the resulting jackknife variance estimates and on the disclosure of original PSU indicators was examined (see Dohrmann *et al.*, (2002) for more detail). The final chosen method for the 1999-2000 NHANES release (termed the “cluster-split PSU” alternative in Dohrmann *et al.*, 2002) entailed ordering the segments on minority density and then assigning the first half within a PSU to one pseudo-PSU and the second half to another pseudo-PSU. Due to the ordering on minority density one expects that the resulting pseudo-PSUs formed from this method will not have the same characteristics as the full PSU. In addition, the order of the replicates was then scrambled to further protect confidentiality.

As reported in Dohrmann *et al.* (2002, 2004, 2005), the protection of confidentiality seemed to be adequate but there were concerns about the performance of the resulting variance estimator. For the 70 characteristics investigated, this method resulted in a pattern of bias in the masked jackknife variance estimates when plotted against the design effects. Figure 1(a) gives the side-by-side boxplots of the ratios of the estimated standard errors using the PSU-splitting alternative (method 1) to the estimated standard errors using the unmasked PSUs for four different ranges of the (original) design effects on the x-axis.

Further research revealed the reasons behind the underestimation and its pattern exhibited in Figure 1(a). Dohrmann *et al.* (2004, 2005) showed that the masked variance estimate is approximately equal to half of the unmasked variance estimate plus a non-negative term that is dependent on the between segment variability within the split PSUs. Also, using Kish’s design effect formula, Dohrmann *et al.* (2005) explains how the curvature pattern over the design effect may arise.

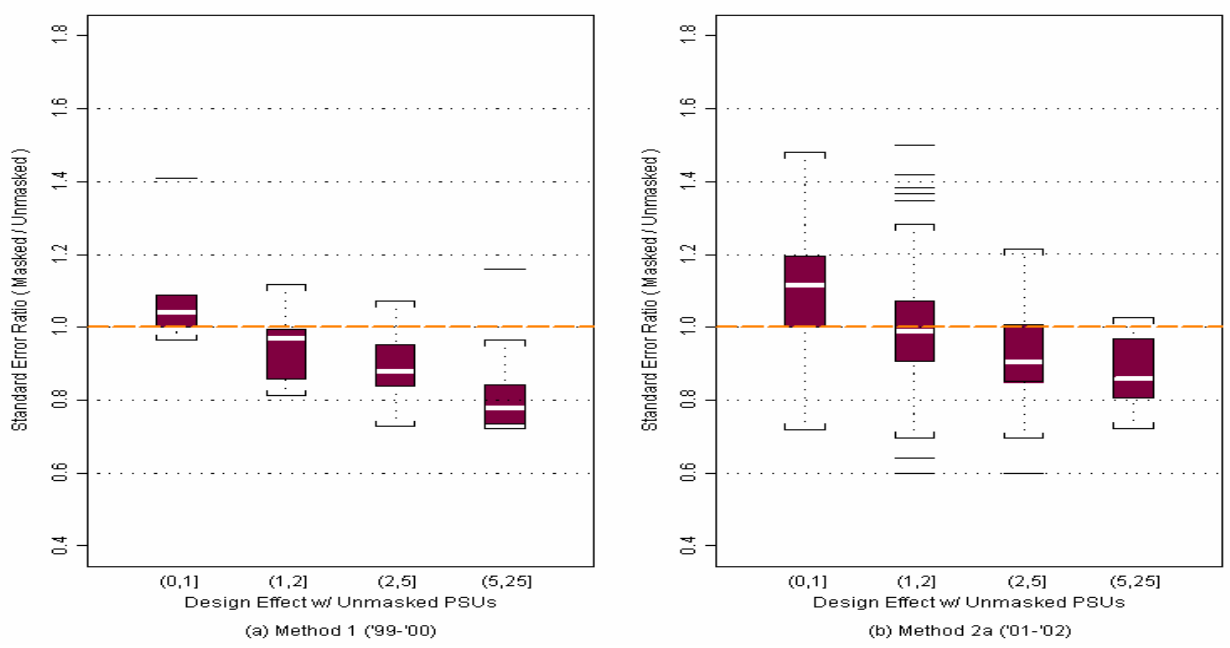

Figure 1. Comparison of standard error ratios against the original design effect for method 1 and method 2

2.2 PSU-Recombining in 2001-2002 (Method 2a)

After the release of the NHANES 1999-2000 data, with the cluster-split jackknife weights, there was a need to change the basic methodology used for variance estimation. Beginning in 2002, NHANES is a stratified design, with two-PSUs per stratum for the two-year samples. Given this, and the great number of replicates needed for continual two-year releases, NCHS decided that in future data releases only PSU and strata indicators will be released for variance estimation. As a result, a new method of variance estimation had to be developed for use with the publicly released data.

One could stay with a PSU-splitting approach in a way to minimize the impact on the variance (Dohrmann *et al.*, 2005, Section 2). However, it still may give a false sense of the number of actual PSU's and therefore an inappropriate view of how many (nominal) degrees of freedom the resulting variance estimator might have (see Lu, Brick and Sitter, 2006 for related discussion). Another obvious alternative is to recombine the splits with splits from other PSUs. This strategy of PSU-splitting and recombining is merely one method of changing the SSU assignment, or swapping segments between PSUs.

Many surveys swap data values between cases for disclosure limitation (*e.g.*, Dalenius and Peiss, 1982). Recall, however, our work is focused on the possibility of revealing PSU identity through the variance estimation method and thus increasing the chance of identifying an individual. Rather than swapping individual values, we decided to swap segments (SSUs). That is, for two similar segments in different PSUs, swapping the PSU and variance stratum identifiers for all sampled cases.

The chosen PSU masking strategy was to apply record linkage techniques (Fellegi and Sunter, 1969) to identify swapping partners similar in a number of demographic, and controls the swapping rate through sampling. This method involves three basic steps: matching, sampling, and bias evaluation. These steps were repeated to adjust the sampling (swapping) rate and the matching method. The process was stopped when we were satisfied that, on average, swapping has negligible effects on key variance estimates. See Dohrmann *et al.* (2004, 2005) for detail.

Figure 1(b) presents a side-by-side boxplot of the estimated standard error ratios drawn for four ranges of the corresponding design effects of the point estimates based on the unmasked PSUs. As compared to Figure 1(a), the curvature underestimation pattern in variance estimates is less severe for the newly adopted PSU-recombining method (Method 2a) although it involves relatively larger variation in the standard error ratio (i.e., larger change in variance estimates) over the entire range of the (original) design effect.

Recall that the PSU-recombining method for the NHANES 2001-2002 adopts a segment matching strategy for swapping that selects swapping partners (segments) nearly identical in the matching characteristics. However, such a matching approach may not be the optimal choice in minimizing the bias of variance estimates. In the next section, we will discuss how we can improve segment matching strategy so as to limit the impact on the variance estimates.

3. Bias Limitation Strategy in Variance Estimation

3.1 Segments in Variance Estimation

Consider that a multi-stage probability sample is chosen from a two-PSU-per-stratum design. 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} segments within each sampled PSU (hi) and n_{hij} ultimate units within each sampled segment. Let $S = \{(hijk) : h = 1,..,H, i = 1,2, \quad j = 1,...,n_{hi}, k = 1,...,n_{hij}\}$ denote the corresponding sample index set. Associated with the sampled ultimate unit $(hijk) \in S$ are the observed value y_{hijk} of characteristic y and the sampling weight w_{hijk}. Then the Taylor series variance estimator of the weighted sample mean $\bar{y} = \sum_S w_{hijk} y_{hijk} / \sum_S w_{hijk}$ is given as

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

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

Writing z_{hi} in (2) in the units of the segments, we can easily see the segments' contribution to the variance estimate, thus helping to find better segment swapping strategies to limit biases in the variance estimates. If w_{hij} and $w_{k|hij}$ denote the segment 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 $\bar{y}_{hij} = \hat{N}_{hij}^{-1} \sum_{k=1}^{n_{hij}} w_{k|hij} y_{hijk}$ denote, respectively, the estimated size and sample mean of segments (hij). The quantities z_{hi} in (2) can be written as

$$\begin{aligned} z_{hi} &= \sum_{j=1}^{n_{hi}} 2 w_{hij} \hat{M}^{-1} \hat{N}_{hij} \left(\bar{y}_{hij} - \bar{y} \right) \\ &= \sum_{j=1}^{n_{hi}} 2 w_{hij} z_{hij} \end{aligned} \tag{3}$$

where $z_{hij} = \hat{M}^{-1} \hat{N}_{hij} \left(\bar{y}_{hij} - \bar{y} \right) = \sum_k w_{k|hijk} z_{hijk}$. It is clear from (2) and (3) that the contribution of the sampled segments to the variance estimate is through three components $\left\{ w_{hij}, \hat{N}_{hij}, \bar{y}_{hij} \right\}$.

Now, to see the effect of segment swapping on the variance estimate, assume that two segments $(h_a i_a j_a)$ and $(h_b i_b j_b)$ are to be swapped between two PSUs $((h_a i_a) \neq (h_b i_b))$. Let i'_a and i'_b denote the other PSUs in strata h_a and h_b , respectively, and define $z_{hi(j)} = z_{hi} - 2 w_{hij} z_{hij} = \sum_{l \neq j} 2 w_{hil} z_{hil}$. The masked variance estimate can be written from (2) as

$$v\left(\bar{y} \mid S^*\right) = v(\bar{y} \mid S) + e_0(y) \times r_0(y), \tag{4}$$

where

$$e_0(y) = 2 w_{h_a i_a j_a} z_{h_a i_a j_a} - 2 w_{h_b i_b j_b} z_{h_b i_b j_b}$$

is the difference in the quantity $2 w_{hij} z_{hij} = 2 w_{hij} \hat{N}_{hij} \left(\bar{y}_{hij} - \bar{y} \right) / \hat{M}$ of the two segments to be swapped and

$$r_0(y) = \begin{cases} \left[z_{h_a i_a (j_a)} - z_{h_b i_b (j_b)} \right] & \text{if } h_a = h_b, \\ \dfrac{1}{2} \left\{ \begin{array}{l} \left(z_{h_a i'_a} - z_{h_a i_a (j_a)} \right) \\ - \left(z_{h_b i'_b} - z_{h_b i_b (j_b)} \right) \end{array} \right\} & \text{if } h_a \neq h_b, \end{cases}$$

is a function of $2 w_{hij} z_{hij}$ of the segments to be retained in the original PSUs. It shows that the effect of segment swapping on the variance estimate will be negligible if the two segments for swapping are paired in such a way that the product of $e_0(y)$ and $r_0(y)$ is close to zero. In other words, the change in the variance estimate under segment swapping can be controlled when a segment pair is formed taking into account all three components $\left\{ w_{hij}, \hat{N}_{hij}, \bar{y}_{hij} \right\}$ so as to minimize $e_0(y) \times r_0(y)$. See Park (2006) for the proof of (4).

3.2 Sequential Segment Swapping with Multiple Matching Characteristics (Method 2b)

Suppose that a total of R segments are chosen to form their pairs for swapping. Let $j_1,\ldots,j_R$ denote their labels listed in the sequential order for the segment matching process. Let S^{r-1} denote the sample index set after the $(r-1)$ segments in the list, where $r=1,\ldots,R$ and $S^0 \equiv S$. The change in the variance estimate caused by swapping the r th segment (j_r) and any other segment that were not involved in the previous match(es) can be written as

$$\begin{aligned}\delta_r(y) &= v\left(\bar{y} \mid S^{r-1}_{(j_r,j)}\right) - v\left(\bar{y} \mid S^{r-1}\right)\\ &= e_{r-1}(y) \times r_{r-1}(y),\end{aligned} \tag{5}$$

where $e_{r-1}(y)$ and $r_{r-1}(y)$ are defined similarly as in (4) but are based on the sample index set S^{r-1} and $S^{r-1}_{(j_r,j)}$ denotes the sample index set with segments j_r and j being swapped. Clearly, the choice of the best match for the r th segment depends on the previous match(es) and thus the matching process should be viewed as a sequential process. Note that those segments that were matched and swapped in the previous matches should be excluded in the current search.

In addition, more than one characteristic can be considered for segment 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 $x=\left(x_1,\ldots,x_q\right)'$ (see Dohrmann *et al*., 2005, for some related discussion). To measure the distance between the two segments j_r and j, any distance measure of the forms

$$D_{2b,r}(j \mid x) = \sum_{l=1}^{q} c_l \left| v\left(\bar{x}_l \mid S^{r-1}_{(j_r,j)}\right) - v\left(\bar{x}_l \mid S^{r-1}\right)\right| \tag{6a}$$

or

$$\Delta_{2b,r}(j \mid x) = \sum_{l=1}^{q} c_l \left| v\left(\bar{x}_l \mid S^{r-1}_{(j_r,j)}\right) - v\left(\bar{x}_l \mid S\right)\right| \tag{6b}$$

can be considered with any reasonable choice of positive coefficients c_l. For example, $c_l \equiv 1$ simply considers the absolute difference in variance estimates and $c_l = v\left(\bar{x}_l \mid S\right)^{-1}$ the absolute difference in variance estimates relative to the original variance estimates. The first distance measure (6a) considers the change in the variance estimate due to swapping segments of the r th pair. The second distance measure (6b) takes into account the cumulative swapping effects of all the r segment pairs.

Matching constraints can be set, for example, to prohibit the pairing of segments from the same PSU and to apply a threshold of the proportion of segments from each PSU to be swapped (Lu, 2004). More sophisticated choices of $\left\{c_l : l=1,\ldots,q\right\}$ may help to further minimize variation in the change of the variance estimates. Also, if one uses multivariate techniques such as a principal component analysis to develop scores (e.g., one or more principal component axes) from a larger number of continuous characteristics, variation in the change of the variance estimates may be further reduced.

4. Application and Evaluation on NHANES 2003-2004

Two segment matching strategies were investigated using NHANES 2003-2004 data. Each method swapped segments between PSUs. Once segments were swapped, SUDAAN was used to calculate variance estimates via Taylor series using the resulting pseudo-PSUs.

The new method, referred to as Method 2b, formed segment pairs for swapping based on the following distance measure:

$$D^*_{2b,r}(j \mid x) = \sum_{l=1}^{q} \left| \frac{v\left(\bar{x}_l \mid S^{r-1}_{(j_r,j)}\right) - v\left(\bar{x}_l \mid S^{r-1}\right)}{v\left(\bar{x}_l \mid S\right)} \right|,$$

which results in a search for the segment pair that minimizes the relative change in the variance estimates for q matching characteristics.

In its application, we took the following steps:

Step 1. For each segment, find the (initial) partner from a different PSU with the smallest $D^*_{2b,1}(j \mid x)$;

Step 2. Sort the segments in ascending order of $D^*_{2b,1}(j \mid x)$ within each PSU, and pick a number of segments from the top for swapping;

Step 3. Sort those chosen segments in ascending order of $D^*_{2b,1}(j \mid x)$ across PSUs;

Step 4. Find (or update) each segment's swapping partner sequentially so as to minimize $D^{*}_{2b,r}(j\,|\,x)$.

The previous method, referred to as Method 2a, formed segment pairs for swapping using the following distance measure:

$$D^{*}_{2a,r}(j)=\sum_{l=1}^{q}\left|\bar{x}_{l,j_r}-\bar{x}_{l,j}\right|.$$

It is equivalent to the one used for NHANES 2001-2002 data since the above distance measure basically searches for the segment pair that are the most similar in matching characteristics. Its application steps are the same as those for Method 2b with using the above distance measure.

Table 1 shows some descriptive statistics for the matching methods examined for the NHANES 2003-2004 data. The results of two methods are also presented in Figure 2 using side-by-side boxplots of the standard error ratios for four distinct ranges of the (original) design effects. As expected from the matching criteria, Method 2b reduced a decreasing curvature trend over the design effects. The standard error ratio distributions in most of the design effect ranges (i.e., 0 to 5) are more balanced around the point of one for Method 2b. For the extreme design effects (larger than 5), Method 2b produced less variation in the standard error ratios with their center being closer to the reference value of one. Based on the results of the above analyses, the pseudo-PSU and stratum indicators resulting from Method 2b were released for the 2003-2004 sample.

Table 1. Comparisons of distribution of standard error ratios by original design effects

Standard error ratio		Baseline design effect				
	Statistics	(.1,1.0]	(1.0,2.0]	(2.0,5.0]	(5.0,25.0]	Overall
Matching method	Number of characteristics	143	282	232	44	701
Method 2a	mean	1.030	1.005	0.964	0.890	0.989
	100%	1.796	1.423	1.353	1.028	1.796
	90%	1.174	1.131	1.094	0.997	1.128
	75%	1.051	1.054	1.018	0.953	1.034
	50%	1.000	1.003	0.982	0.907	0.998
	25%	0.987	0.966	0.886	0.823	0.931
	10%	0.909	0.874	0.826	0.750	0.847
	0%	0.647	0.604	0.631	0.668	0.604
	IQR	0.064	0.088	0.132	0.130	0.103
	Range	1.149	0.819	0.722	0.360	1.192
Method 2b	mean	1.038	0.996	0.977	0.923	0.994
	100%	1.524	1.291	1.237	1.032	1.524
	90%	1.170	1.095	1.065	0.999	1.096
	75%	1.071	1.055	1.029	0.968	1.036
	50%	1.007	1.000	0.987	0.928	0.997
	25%	0.983	0.948	0.936	0.880	0.938
	10%	0.906	0.889	0.874	0.825	0.878
	0%	0.763	0.672	0.761	0.806	0.672
	IQR	0.088	0.107	0.093	0.088	0.098
	Range	0.761	0.619	0.476	0.226	0.852

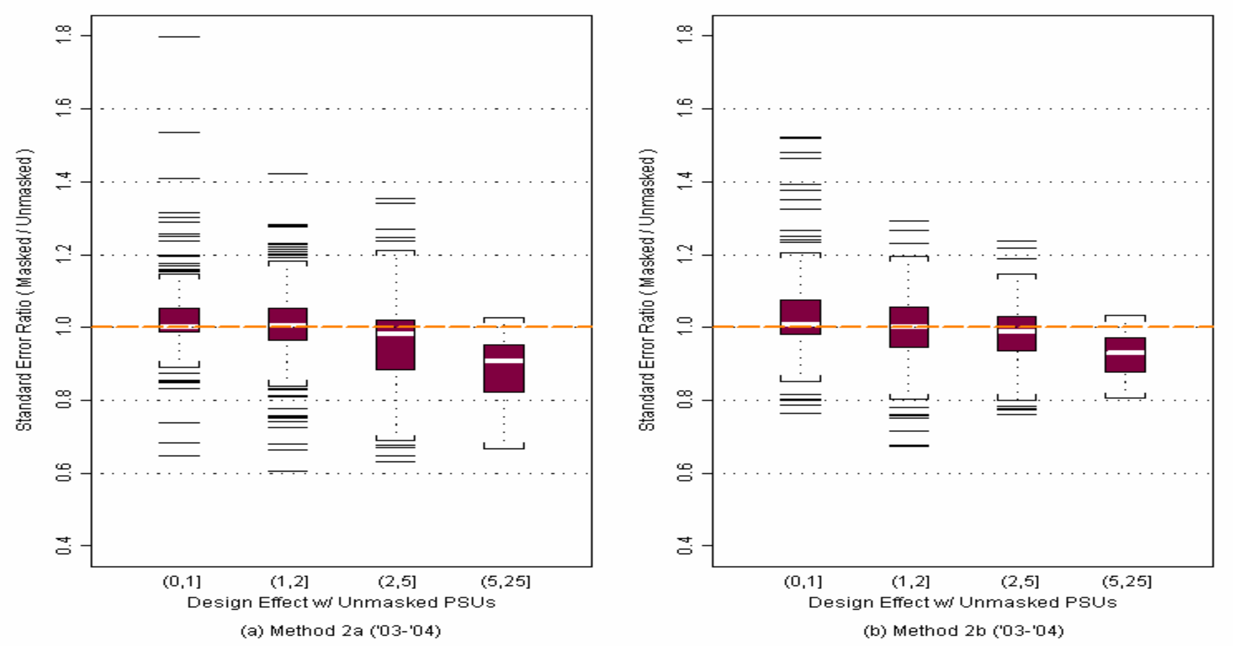

Figure 2. Comparisons of ratios of standard errors against original design effects

5. Concluding Remarks

This research was conducted as part of our ongoing effort to improve data utility while keeping the NHANES PSUs confidential. In general, PSU masking can distort the clustering structure in the original sample design, possibly yielding systematic biases in the variance estimation (Section 2). However, the proposed PSU masking strategy (Section 3) can help reduce such biases to a larger extent as seen in our application to NHANES 2003-2004 data (Section 4). Research on the effects of PSU masking would be interesting for other types of complex data analysis such as regression and multivariate analyses.

Acknowledgement

The authors wish to thank Lee Harding for his assistance with the analysis.

References

Dalenius, T., Peiss, S.P. (1982), Data-swapping: A Technique for Disclosure Control. *Journal of Statistical Planning and Inferences*, 6, pp. 73-85.

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 Surveys. *Submitted to Journal of Official Statistics.*

Dohrmann, S., Mohadjer, L., Montaquila, J., Sitter, R. Lu, W., and Curtin, L.R. (2004). Limiting the Risk of Data Disclosure by Using Swapping Techniques in Variance Estimation. *Proceedings of the Survey Research Methods*, American Statistical Association, pp. 3429-3434.

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,* American Statistical Association, pp. 807-812.

Felligi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Hunter, E.L. and Arnett, R. (1996). Survey “Reinventing” at Health and Human Services. Chance 9, 54-57.

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*, accepted.

Lu, W. (2004). *Confidentiality and Variance Estimation in Complex Surveys.* Unpublished doctoral dissertation, Simon Fraser University, Burnaby, Canada.

Park, I. (2006). PSU Masking and Variance Estimation in Complex Surveys. Working paper.