Protecting survey design information by combining strata and accounting for the realized sample selection.

Vladislav Beresovsky[*]

[*]National Center for Health Statistics, 3311 Toledo Rd, Hyattsville, MD 20782

Abstract

Most survey designs contain revealing information about the geographic locality of the sampled items. When combined with other sources, such information could violate the confidentiality of the survey data. We propose the modification of existing methods for masking design information by constructing combined strata variance estimators. In our approach variance estimation error is minimized conditionally on the realized PSU selection, which is often different from the optimal way of selecting PSU with probability proportional to their size (PPS). Advantages of the proposed method for combining strata over other methods are demonstrated by comparing relative deviations between standard errors of the estimates of totals and means calculated using grouped and ungrouped strata. Coverage of the finite population target variables by their estimated confidence intervals is analyzed in a simulation experiment for different ratios between variability of the target variables and PSU sizes. Optimal properties of the proposed method for grouping strata are validated in application to real survey data.

Key Words: confidentiality, stratified clustered sampling, PPS sampling, variance estimation, degrees of freedom.

Introduction

The idea of combining design strata and PSU for approximate variance estimation originated with replication methods for variance estimation, such as Jackknife (JK)and balanced repeated replication (BRR); see McCarthy (1966), Wolter (1985), Lee (1972, 1973), Rust (1984) and Nixon et al. (1998). The main initial goal was to reduce the large number of replicates to accommodate the limited computer power of the day. Usually, a large number of replicates is required when self-representing primary sampling units (PSU) were treated by survey designers as strata and, sometimes, thousands of secondary sampling units (SSU) were considered PSU. In such cases, application of the popular BRR method for variance estimation becomes possible in the form of alternative BRR (ABRR) (Wolter (1985), pp 132-133). This method entails creating multiple *variance strata* by *randomly* pairing PSU from the original strata. Rao and Shao (1996) have proved that ABRR is asymptotically correct. Lu et al. (2006) proposed a consistent methodology for grouping *variance strata* in order to simultaneously maximize degrees of freedom in the specified domains. In most of the cited literature, the main criterion for finding the best method for combining strata and PSU was maximizing the degrees of freedom calculated by a Satterthwaite approximation using combined design information.

Mayda et al. (1996) proposed combining strata and PSU to protect confidentiality of the public use microdata files released by the Canadian National Population Health Survey. They compared coefficients of variations for totals, ratios and regression coefficients estimated for collapsed and uncollapsed survey designs using Jackknife replication method.

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Their conclusion was that collapsing strata does preserve confidentiality without negatively affecting the variance estimate. However, the authors suggested that variance estimated from collapsed design could be unstable for rare characteristics and within small domains. In this paper we further develop approaches for combining strata presented by Lu et al. (2006) and apply them for masking the original survey design information in public use microdata files (PUF) for confidentiality protection.We investigate how combining strata and PSU changes estimated variances and how it affects the coverage of finite population values. To the users of PUF, such information provides confidence in the usability of a masked survey design for variance estimation. This question is closely related to the reduction of degrees of freedom associated with grouping strata. In fact, part of the expression for the expected squared deviation between variances estimated from grouped and original ungrouped designs obtained in this paper was used by Lu et al. (2006) for degrees of freedom calculation.

Variances in strata are presented as a sum of contributions from the variability of the target variable between PSU and the variability of PSU sizes in a given sample. The squared difference between variances of the totals calculated using grouped and ungrouped strata and PSU $d^2\left(\hat{Y}\right)$ is estimated in Section 1 by taking expectation conditional on the PSU sizes. One term of the resulting expression is similar to the result of Lu et al. (2006), but variance per stratum explicitly reflects variability between PSU sizes. Other terms, absent in Lu et al. (2006), depend on the particular way of combining PSU within paired strata. A two-step algorithm for combining strata and PSU is proposed in Section 2 to minimize different terms of the expression for the squared difference between grouped and ungrouped variance estimates of the totals. The first step establishes the optimal way of combining strata and is a variant of algorithms proposed in Lee (1972, 1973) and Lu et al. (2006), simplified for the needs of pairing strata for confidentiality protection. At the second step PSU are optimally paired within combined strata to minimize another term of $d^2\left(\hat{Y}\right)$. We conduct simulations to compare the grouped and ungrouped variances of the means and totals and also to compare the coverage properties of the corresponding confidence intervals. Simulations are described and their results are presented in Section 3. Results of applying the proposed algorithm to real survey data are discussed in Section 4. Conclusions about the possibility of pairing strata and PSU for confidentiality protection and retaining utility of the aggregated design for reliable variance estimation are presented in Section 5.

1. Error of variance estimation using aggregated survey design

1.1 How grouping strata and PSU changes variance estimators

To simplify our treatment we consider classic survey design, namely two PSU per strata. Grouping of strata results in two or more original strata $h \in 1 \ldots H$ combined in one grouped stratum $g \in 1 \ldots G$ with two PSU. Each aggregated PSU $(gi), i = 1, 2$ includes one of the PSU from every original stratum. Variances estimated assuming grouped and ungrouped survey designs should not be the same. We need to minimize their difference for the given sample by choosing the optimal strategy for grouping strata h and PSU (hi). If PSU were not sampled with probability proportional to size (PPS), the measure of size $N_{hi} = \sum_{j \in (hi)} w_{hij}$ is different between PSU of the same strata. Then, the familiar expression for the variance of the total $\hat{Y} = \sum_s w_{hij} y_{hij}$ estimated from the stratified sample with replacement (WR) with two PSU per strata:

$$v\left(\hat{Y}\right)=\frac{1}{4}\sum_{h=1}^{H}\left(Y_{h1}-Y_{h2}\right)^{2} \qquad (1.1)$$

where $Y_{hi}=\sum\limits_{j\in(hi)}w_{hij}y_{hij}$ is weighted sum in PSU (hi), may be presented as the contribution from two sources of variability. One originates from the stochastic variability between weighted means in PSU $\bar{y}_{hi}=Y_{hi}/N_{hi}$. Another is due to the difference between PSU sizes in strata $\Delta N_h=N_{h1}-N_{h2}$ observed for the given sample. In these notations (1.1) becomes:

$$v\left(\hat{Y}\right)=\frac{1}{4}\sum_{h=1}^{H}N_h^2\left(\bar{y}_{h1}-\bar{y}_{h2}+\Delta N_{h,rel}\bar{y}_h\right)^2=\frac{1}{4}\sum_{h=1}^{H}z_h^2 \qquad (1.2)$$

where $N_h=N_{h1}+N_{h2}$ is estimate of strata size, $\bar{y}_h=\left(\bar{y}_{h1}+\bar{y}_{h2}\right)/2$ is mean of target variable in strata and $\Delta N_{h,rel}=\left(N_{h1}-N_{h2}\right)/\frac{1}{2}N_h$ is relative difference of PSU sizes in strata.

Variance of the total estimated from the grouped design can be derived from (1.1) and (1.2) and expressed using estimates in the original strata:

$$v_g\left(\hat{Y}\right)=\frac{1}{4}\sum_{g=1}^{G}\left(\hat{Y}_{g1}-\hat{Y}_{g2}\right)^2=\frac{1}{4}\sum_{g=1}^{G}\left(\sum_{h\in g}\left(\hat{Y}_{h1}-\hat{Y}_{h2}\right)\right)^2=\frac{1}{4}\sum_{g=1}^{G}\left(\sum_{h\in g}z_h\right)^2=$$
$$\frac{1}{4}\sum_{g=1}^{G}\sum_{h\in g}z_h^2+\frac{1}{4}\sum_{g=1}^{G}\sum_{h\neq h'\in g}z_hz_{h'}=v\left(\hat{Y}\right)+d\left(\hat{Y}\right) \qquad (1.3)$$

Variances estimated using grouped and ungrouped designs differ by $d\left(\hat{Y}\right)$. Conditions for the optimal way of grouping strata will follow from the minimization of the squared value of this term:

$$d^2\left(\hat{Y}\right)=\frac{1}{16}\sum_{g,g'=1}^{G}\sum_{\substack{h\neq h'\in g\\ k\neq k'\in g'}}\Big[(A_hA_{h'}+D_hD_{h'})(A_kA_{k'}+D_kD_{k'})+$$
$$(A_hD_{h'}+A_{h'}D_h)(A_kD_{k'}+A_{k'}D_k)+$$
$$(A_hA_{h'}+D_hD_{h'})(A_kD_{k'}+A_{k'}D_k)+(A_hD_{h'}+A_{h'}D_h)(A_kA_{k'}+D_kD_{k'})\Big] \qquad (1.4)$$

where notations $A_h=N_h\left(\bar{y}_{h1}-\bar{y}_{h2}\right)$ and $D_h=N_h(\Delta N_{h,rel})\bar{y}_h$ are used for brevity.

1.2 Expectation conditional on PSU sizes

Variance estimates (1.2) - (1.3) are expressed as statistics depending on two random variables defined at the PSU level: weighted mean of the target variable $\bar{y}_{hi}$ and weighted PSU size N_{hi}. In this paper we are interested in minimizing the difference between estimates of the variance using grouped and ungrouped PSU and strata *conditionally on the given sample*. We consider such an approach more relevant to practical situations when PSU are combined for confidentiality protection *after* the particular sample was selected.

Taking expectation over the distribution of target variable $\bar{y}_{hi}$ we assume the following:

$$E\left(\bar{y}_{hi}\right) = \bar{y}_h \quad (1.5a)$$

$$E\left[\left(\bar{y}_{h1} - \bar{y}_{h2}\right)\left(\bar{y}_{h'1} - \bar{y}_{h'2}\right)\right]/2 = \delta_{hh'}\sigma^2_{yh} \quad (1.5b)$$

where $\delta_{hh'} = 1$ if $h = h'$ and 0 otherwise. Then, conditional expectation of ungrouped variance estimator(1.2) over the distribution of $\bar{y}_{hi}$ is:

$$E_{y_{hi}}\left[v\left(\hat{\hat{Y}}\right)|N_{hi}\right] = \frac{1}{4}\sum_{h=1}^{H} 2N_h^2\left[\sigma^2_{yh} + \left(\Delta N^2_{h,rel}/2\right)\bar{y}_h^2\right] = \frac{1}{4}\sum_{h=1}^{H} 2\tilde{\sigma}^2_h \quad (1.6)$$

where the contribution from each stratum h to the variance estimator includes the sum of coefficients of variation of target variable and PSU sizes $2\tilde{\sigma}^2_h = 2N_h^2\bar{y}_h^2\left(CV^2(\bar{y}_{hi}) + \widehat{CV}^2(N_{hi})\right)$. $CV(x) = StdErr(x)/E(x)$ denotes coefficient of variation of random variable x. Conditional expectations of the individual terms contributing to the squared difference between ungrouped and grouped variance estimators (1.4) can be expressed using moments of the distribution of $\bar{y}_{hi}$ (1.5a-b) :

$$E_{y_{hi}}\left[A_hA_{h'}A_kA_{k'}|N_{hi}\right] = 4N_hN_{h'}N_kN_{k'}\sigma^2_{yh}\sigma^2_{yk}\delta_{gg'}\left[\delta_{hk}\delta_{h'k'} + \delta_{h'k}\delta_{hk'}\right] \quad (1.7a)$$

$$E_{y_{hi}}\left[A_hD_{h'}A_kD_{k'}|N_{hi}\right] = 2N_hN_{h'}N_kN_{k'}\sigma^2_{yh}(\Delta N_{h',rel}\bar{y}_{h'})(\Delta N_{k',rel}\bar{y}_{k'})\delta_{gg'}\delta_{hk} \quad (1.7b)$$

$$E_{y_{hi}}\left[D_hD_{h'}D_kD_{k'}|N_{hi}\right] = N_hN_{h'}N_kN_{k'}\Delta N_{h,rel}\bar{y}_h\Delta N_{h',rel}\bar{y}_{h'}\Delta N_{k,rel}\bar{y}_k\Delta N_{k',rel}\bar{y}_{k'} \quad (1.7c)$$

Expectation of the rest of the terms in (1.4) is 0 because they include either an odd number of multipliers A_h or these multipliers clearly belong to different strata of the same group $A_hA_{h'}, h \neq h'$. Using (1.7a-c), and after some straightforward algebra, expectation of the squared difference between grouped and ungrouped variance estimators (1.4) could be written as:

$$\begin{aligned}
&E_{y_{hi}}\left[d^2\left(\hat{Y}\right)|N_{hi}\right] = \\
&\frac{1}{2}\left[\sum_{g=1}^{G}\left(\sum_{h\in g}\tilde{\sigma}^2_h - \frac{1}{G}\left(\sum_{g=1}^{G}\sum_{h\in g}\tilde{\sigma}^2_h\right)\right)^2 + \frac{1}{G}\left(\sum_{g=1}^{G}\sum_{h\in g}\tilde{\sigma}^2_h\right)^2 - \sum_{g=1}^{G}\sum_{h\in g}\tilde{\sigma}^4_h\right] + \\
&\qquad \frac{1}{2}\sum_{g=1}^{G}\sum_{h\neq h'\neq k\in g} N_h^2N_{h'}N_k\sigma^2_h\Delta N_{h',rel}\Delta N_{k,rel}\bar{y}_{h'}\bar{y}_k + \\
&\qquad \frac{1}{16}\left(\sum_{g=1}^{G}\sum_{h\neq h'\in g} N_hN_{h'}\Delta N_{h,rel}\Delta N_{h',rel}\bar{y}_h\bar{y}_{h'}\right)^2 - \\
&\qquad\qquad -\frac{1}{8}\sum_{g=1}^{G}\sum_{h\neq h'\in g} N_h^2N_{h'}^2\Delta N^2_{h,rel}\Delta N^2_{h',rel}\bar{y}_h^2\bar{y}_{h'}^2 \quad (1.8)
\end{aligned}$$

2. Optimal algorithm for combining strata and PSU

The basic features of the proposed algorithm for combining strata and PSU can be easily described in application to the simplest survey design assuming an even number of strata H with 2 PSU in each stratum. Pairing of original strata will result in $G = H/2$ combined strata each having 2 combined PSU.

The optimal algorithm for combining strata and PSU should minimize expectation of the difference between grouped and ungrouped variance estimators (1.8) conditional on the sampled PSU. The 1st term of this expression depends only on the grouping of strata. Minimizing this term is identical to maximizing degrees of freedom of variance estimators discussed in Lu et al. (2006). However, in this approach, variances in strata $\tilde{\sigma}_h^2$ explicitly depend on the variability of PSU sizes (1.6). Minimization is achieved if the original design strata h are grouped to insure approximate equality of the combined variance in every group $\sum_{h \in g} \tilde{\sigma}_h^2$. Application of the semi-ascending order arrangement (SAOA) algorithm proposed by Lee (1972, 1973) will allow one to achieve this goal. It is particularly simple in the case of grouping only two design strata for confidentiality protection:

1. Sort the H design strata in descending order of $\tilde{\sigma}_h^2$.
2. Rearrange the last $H/2$ strata in ascending order of $\tilde{\sigma}_h^2$.
3. Combine each stratum $i \in 1 \ldots H/2$ with stratum $(i + H/2)$ to form $G = H/2$ grouped strata.

After the grouping of strata is defined, some other terms of (1.8) can be minimized by the proper grouping of PSU within combined strata. Minimizing these terms is simplified when *only two* design strata are combined within each group, which is often sufficient for confidentiality protection in practice. In such case the 2nd term disappears because it results from grouping 3 design strata. The 3rd term includes squared sum of contributions from paired design strata over all groups. Below we will assume that the means of the target variable $\bar{y}_h$ are equal between strata. The sign of the contribution to this sum from a group $(h, h' \in g)$ depends on the relative sizes of the combined PSU. Combining larger (or smaller) PSU together produces positive contribution, but combining larger PSU from one stratum to smaller PSU from another produces negative contribution. To be specific, suppose that for both paired strata the first PSU is larger than the second $N_{(h,h')1} > N_{(h,h')2}$. Under these assumptions we propose the following algorithm for grouping PSU to minimize the 3rd term of (1.8).

4. Order all combined strata in descending order of their absolute contribution to the sum of the 3rd term $\left| N_h N_{h'} \Delta N_{h,rel} \Delta N_{h',rel} \right|$.
5. For the first combined stratum, pair PSU from both design strata with the same numbers $(h1) \leftrightarrow (h'1)$ and $(h2) \leftrightarrow (h'2)$, producing a positive contribution to the sum.
6. For the second strata, combine PSU with different numbers $(h1) \leftrightarrow (h'2)$ and $(h2) \leftrightarrow (h'1)$, producing a negative contribution to the sum.
7. Keep progressing over the ordered combined strata while checking the running sum at each step. If this sum is positive, combine PSU with different numbers producing a negative contribution, and vice versa. Then move to the next step.

Because of the descending ordering of the combined strata at Step 4, the absolute values of contributions from the consecutive steps are decreasing. Such a process is guaranteed to minimize the 3rd term of (1.8) as much as possible for the given collection of combined strata. The 4th term includes $\left(\Delta N_{h,rel} \Delta N_{h',rel}\right)^2$ and does not depend on the particular grouping of PSU within combined strata.

3. Variance estimation errors and coverage of the finite population totals and means: simulation experiment

The extent of the deviation between standard errors estimated using grouped strata $SE^{(g)}$ and ungrouped design strata $SE^{(d)}$ can be assessed by their relative difference:

$$R^{(g)} = \left(SE^{(g)} - SE^{(d)}\right) \Big/ SE^{(d)} \tag{3.1}$$

The algorithm described in Section 2 is supposed to minimize this deviation for a given sample. However, specific details of the survey design and the need to satisfy practical requirements for confidentiality protection may require some relaxations from the optimal algorithm. A conducted simulation experiment illustrates the effect of these deviations on the estimated standard errors of the estimates of means and totals and on the coverage by the associated confidence intervals.

3.1 Simulating synthetic finite population and sample selection

For the synthesized finite population we assumed a two-level clustered design with PSU stratified in $n = 36$ strata. According to the mathematical formalism developed in Section 1, expectations of variance (1.6) and squared difference between grouped and ungrouped variance estimators (1.8) over the distribution of target variable y_{hi} depend on survey design through the dependence on the strata sizes in population N_h and the difference of PSU sizes ΔN_h within strata.

Characteristics of the synthetic population and sampling were modeled after real survey data: sample size in strata M_h and average PSU sample size (considering 2 PSU per strata) $m_h = M_h/2$, variance of PSU sample sizes m_{hi} relative to the mean sampled PSU size $m_{(s)} = 1/(2n) \sum_{h,i} m_{hi}$ over all strata $v_{(s)}^2 = 1/(2n) \sum_{h,i} \left(m_{hi}/m_{(s)} - 1\right)^2$. For each strata we postulated weights of selecting PSU at the first level $w_h^{(1)}$ and units within PSU at the second level $w_h^{(2)}$ to ensure that the weighted size of the strata in the simulated population was 10% of the strata size in real survey to keep in check the amount of computing time. PSU selection weight $w_h^{(1)}$ was assigned values no less than 5 to ensure sufficient variability of simulated samples.

Sample sizes in PSU were randomly generated for each strata from the normal distribution with mean value in that strata m_h and relative variances $v_{(s)}^2$ estimated for the whole sample $m_{hi}^{(sim)} \sim N\left(m_h, (m_h v_{(s)})^2\right)$. Extreme deviations from the mean were trimmed by the lower $0.4 m_h$ and upper $1.6 m_h$ thresholds. The number of simulated population PSU in strata h was $2 w_h^{(1)}$ and the number of population units in every PSU was $w_h^{(2)} m_{hi}^{(sim)}$.

We simulated binomial target variable y_{ij} for unit j in PSU i using the following normal-binomial two-level distribution model:

$$\begin{aligned} \text{logit}\left(p_i\right) = \text{logit}\left(p_0\right) + v_i;\ v_i \sim N\left(0, \sigma_0^2\right) \\ y_{ij} = Bin(p_i) \end{aligned} \tag{3.2}$$

Note that the simulated target variable was independent of stratum h in the population.

Random variability of the target variable between PSU significantly affects both expectations of variance (1.6) and squared deviation of the grouped variance estimator (1.8). Vari-

ability of the target variable between PSU (1.5b) could be manipulated by changing the standard deviation σ_0 in (3.2). Three populations were simulated having target variables with the same mean binomial probability $p_0 = 0.25$ and three different standard deviations $\sigma_0 = 1$; 0.4 and 0.1, corresponding to the following coefficients of variation $CV(\bar{y}_{hi}) =$ 0.62; 0.27 and 0.07. Variability between PSU sizes was the same for all populations with coefficient of variation $CV(N_{hi}) = 0.42$.
From the simulated population, stratified clustered samples were drawn in two stages. At the first stage, 2 PSU were selected *with replacement* in all 36 strata out of $2w_h^{(1)}$ population PSU. At the second stage, there were $m_{hi}^{(sim)}$ units selected with equal probability *without replacement* in every sampled PSU hi. Such sample selection provides for non-zero contribution from variability between PSU sizes $\Delta N_h = \left(m_{h1}^{(sim)} - m_{h2}^{(sim)}\right) w_h^{(1)} w_h^{(2)}$ to the estimated variance (1.2).

3.2 The difference between grouped and ungrouped variance estimates

In the course of simulations, $N_{sim} = 1000$ samples were drawn from one synthetic population. Population means and totals and their standard errors SE_{mean} and SE_{tot} were estimated for every sample using the original design strata and the combined strata obtained by different grouping methods. To investigate the importance of the different aspects of the proposed algorithm for grouping strata, the following grouping methods were considered:

1. *Grouping 1.* Optimum algorithm proposed in Section 2.

2. *Grouping 2.* Strata were ordered by $N_h^2 \sigma_{yh}^2$ instead of $\tilde{\sigma}_h^2$, ignoring the difference between PSU sizes. PSU in the combined strata were grouped at random.

3. *Grouping 3.* Both design strata and PSU within strata were grouped at random.

4. *Grouping 4.* Strata were ordered by $N_h^2 \sigma_{yh}^2$. Within combined strata smaller PSU of one design strata were preferably (with probability 0.8) grouped with larger PSU from another design strata.

For all simulated samples, relative deviations between grouped and ungrouped estimates of standard errors of the means $R_{mean}^{(g)}$ and totals $R_{tot}^{(g)}$ were calculated (3.1). Boxplot diagrams of the distribution of these values over all samples are presented in Figure 1.
Relative deviations of the grouped-strata estimates of the standard errors of the means $R_{mean}^{(g)}$ were distributed very uniformly for all methods of grouping strata and variabilities of the target variable (3.2). The proposed method (*Grouping 1*) did not offer much advantage over less optimal methods. In all cases the medians of the distributions $R_{mean}^{(g)}$ were located around 0, the interquartile range was found to be within $\sim \pm 10\%$, and most of the observed relative deviations were localized within $\sim \pm 20\%$.
However, estimates of the standard errors of the total $R_{tot}^{(g)}$ demonstrated substantial sensitivity to the method of grouping depending upon the relation between variability of the target variable y_{hi} and PSU sizes $\Delta N_{h,rel}$. Three different cases were considered.
(a) Large variability of the target variable $CV(\bar{y}_{hi}) \geq CV(N_{hi})$. Variability of $R_{tot}^{(g)}$ shows small dependence on the method of grouping strata: *Grouping 3* results in a wider interquartile range and *Grouping 4* produces noticeable negative shift of the median. Optimal *Grouping 1* is only marginally better than other methods.
(b) Intermediate variability of the target variable $CV(\bar{y}_{hi}) \sim CV(N_{hi})$. $R_{tot}^{(g)}$ corresponding to optimal *Grouping 1* appears more tightly concentrated around 0 than for other groupings. Negative shift of the median for *Grouping 4* becomes obvious.

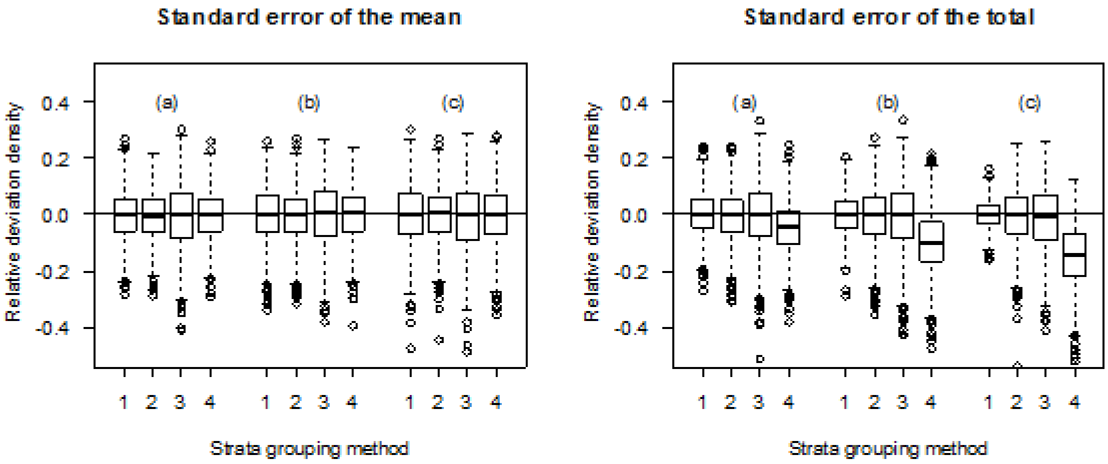

Figure 1: Distribution of the relative deviations between standard errors of the means $R^{(g)}_{mean}$ and totals $R^{(g)}_{tot}$(3.1) estimated using different methods of grouping strata and the estimates using design strata depending on the variability of the target variable y_{hi} between PSU (3.2): (a) $\sigma_0 = 1$; (b) $\sigma_0 = 0.4$; (c) $\sigma_0 = 0.1$.

(c) Small variability of the target variable $CV(\bar{y}_{hi}) \leq CV(N_{hi})$. *Grouping 1* clearly dominates other methods. Negative shift of the median for *Grouping 4* becomes even more pronounced. Random grouping of strata (*Grouping 3*) results in the largest spread of the distribution of relative deviations. Ignoring the difference between PSU sizes (*Grouping 2*) considerably increases the spread of relative deviations compared to *Grouping 1*.

3.3 Finite population properties of grouped and ungrouped variance estimators

Relative deviations $R^{(g)}$ (3.1) quantify differences between standard errors estimated using grouped and ungrouped strata. This is important for validating standard errors estimated from public use micro data files with masked design information. Conducted simulations also allow for investigation of finite population properties of different variance estimators: deviation from the root mean squared error (RMSE) estimated over all simulations and coverage of the finite population means and totals by the confidence intervals estimated using grouped and ungrouped strata. Relative deviation between estimates of the standard errors SE_i for every simulated sample and $RMSE = \sqrt{1/N_{sim} \sum_{i=1}^{i=N_{sim}} \left(\hat{Y}_i - Y\right)^2}$ can be defined similarly to deviation between different variance estimates (3.1):

$$R_i^{RMSE} = (SE_i - RMSE)/RMSE \tag{3.3}$$

Figure 2 presents results for the standard errors of the means R^{RMSE}_{mean} and totals R^{RMSE}_{tot} estimated using grouped and ungrouped strata. Overall, the small negative shift may be attributed to using a simplified variance estimator which neglected variance contribution from the second stage of sample selection.

Standard errors estimated using original design strata (0) deviate somewhat less from RMSE than all estimates using grouped strata. This is expected because grouping of strata results in reduction of degrees of freedom which is a measure of precision of a variance estimator. The proposed method (1) for grouping strata does not offer much advantage over other grouping methods (2-4) for estimating standard errors of the totals, as was the case in Fig-

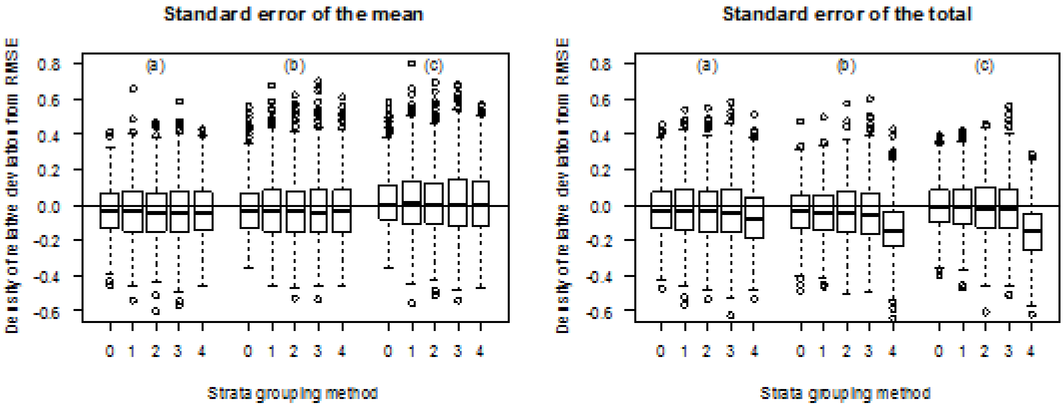

Figure 2: Distribution of the relative deviations between standard errors of the means R^{RMSE}_{mean} and totals R^{RMSE}_{tot} and RMSE over all simulations (3.3). Standard errors were estimated using original design strata (0) and strata grouped by different methods(1-4) depending on the variability of the target variable y_{hi} between PSU (3.2): (a) $\sigma_0 = 1$; (b) $\sigma_0 = 0.4$; (c) $\sigma_0 = 0.1$.

ure 1. This is also expected, because, strictly speaking, the proposed method is optimal for minimizing the difference between variance estimators (1.8) and not the error of the variance estimator itself. One common feature of Figures 1 and 2 is the negative bias of the standard error of the total for grouping method (4), when larger PSU from one strata are preferably grouped with smaller PSU from another. This bias should be positive in the opposite case of preferential grouping together larger (smaller) PSU from different strata. These results closely correspond to the calculated coverage probability of the finite population means and totals by the estimated confidence intervals presented in Table 1. Confidence intervals were estimated from the *t*-distribution with degrees of freedom defined as $DF_d = N_{PSU} - N_{STRATA}$, which in our simulations equals 36 for the original design and 18 for the design with paired strata and PSU.

Table 1: Coverage probabilities of finite population means and totals by the confidence intervals estimated using original design strata and strata grouped by different methods. Results presented for different variabilities of the simulated target variables y_{hi} (3.2) between PSU.

	Strata grouping method				
Variability of target variable	*Ungrouped*	*Grp 1*	*Grp 2*	*Grp 3*	*Grp 4*
Coverage of the finite population means					
Large variability ($\sigma_0 = 1$)	0.948	0.952	0.948	0.941	0.950
Intermediate variability ($\sigma_0 = 0.4$)	0.941	0.951	0.948	0.948	0.949
Small variability ($\sigma_0 = 0.1$)	0.942	0.954	0.944	0.941	0.949
Coverage of the finite population totals					
Large variability ($\sigma_0 = 1$)	0.935	0.947	0.935	0.944	0.931
Intermediate variability ($\sigma_0 = 0.4$)	0.940	0.942	0.950	0.942	0.926
Small variability ($\sigma_0 = 0.1$)	0.946	0.955	0.953	0.945	0.909

Coverage probabilities of both means and totals appear to be close to nominal 0.95 in all cases, except for the coverage of totals when strata were grouped by method (4). Observed undercoverage increased as variability of the target variable between PSU decreased. This is in accord with the negative shift observed in both Figures 1 and 2.

Estimated confidence intervals depend on the assumed degrees of freedom (DF). That is why the correct definition of DF is very important, particularly for small domains. To define DF, Rust (1986) and Lu et al. (2006) used the Satterthwaite approximation $DF(v) = 2\left[E(v)\right]^2 / Var(v)$, where v is the variance estimator. The simple definition of DF_d used above follows from this formula if the variance contributions from all PSU and strata are equal. We calculated $DF(v)$ for ungrouped and grouped variance estimators for the whole sample and for domains defined by Census regions. Pairing of strata caused a 50% reduction of DF_d for variance estimators in all domains for all methods of grouping strata. $DF(v)$ were reduced to $\sim 70\%$ for all domains when methods (1,2 and 4) were used for pairing strata. These methods require ordering strata by size (see Section 3.2) before grouping. In method (3) strata were paired at random and $DF(v)$ was reduced to $\sim 60\%$. This corresponds to the wider spread of deviation between variance estimates observed for method (3) in Figure 1. We can conclude that using the proper method of grouping strata allows us to retain more DF.

4. Application to survey data

Proposed methods for grouping strata were tested with application to actual survey data. Standard errors of the means and totals were estimated using the original design and grouped strata for 47 binomial survey variables of different nature and different variability between PSU. Distribution of the coefficient of variation of the PSU means for these variables defined as $\mathrm{CV}(y_{psu}) = \mathrm{StdErr}(\bar{y}_{psu})/\mathrm{Mean}(\bar{y}_{psu})$ is presented in Table 2. This distribution overlaps with estimated variability of three synthetic variables generated in the simulation study described above.

Table 2: Quantiles of the distribution of coefficients of variation $\mathrm{CV}(\bar{y}_{psu})$ of PSU means for selected 47 survey variables.

	Min	10%	25%	50%	Mean	75%	90%	Max
$\mathrm{CV}(y_{psu})$	0.11	0.26	0.36	0.49	0.74	0.82	1.72	2.78

For each of these variables we calculated relative deviations between estimates of the standard errors of the means $R^{(g)}_{mean}$ and totals $R^{(g)}_{tot}$ using grouped and ungrouped strata (3.1). Distribution of these values over selected survey variables is presented in Figure 3.

Distributions of $R^{(g)}_{mean}$ and $R^{(g)}_{tot}$ for the real data could not be compared directly to similar distributions calculated for the synthetic variables presented in Figure 1. However, obvious similarities can be easily noticed. First, the proposed method (1) for grouping strata produces minimal deviations from the standard errors of the totals estimated using ungrouped strata. Second, preferential grouping of large and small PSU realized in method (4) produces a consistent negative shift of the standard error estimates of the totals. Third, this shift was not observed for the standard error estimates of the means. This comparison provides confidence in the applicability of the developed methodology to real survey data.

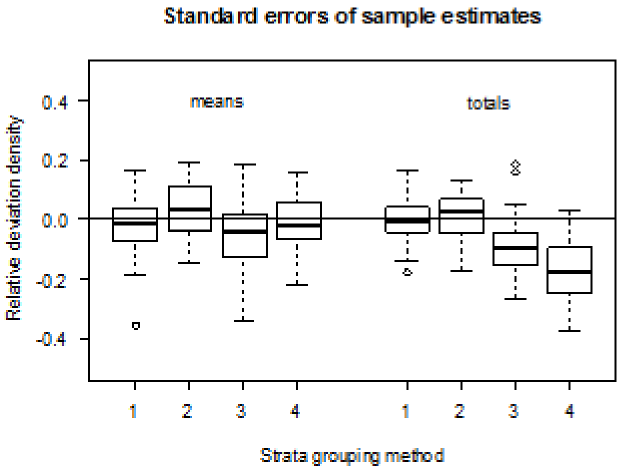

Figure 3: Distribution of the relative deviations between grouped and ungrouped estimates of the standard errors of the means $R^{(g)}_{mean}$ and totals $R^{(g)}_{tot}$ (3.1) over 47 survey variables.

5. Conclusions

In this paper we propose an optimal algorithm for pairing strata and PSU for confidentiality protection minimizing expected squared difference between grouped and ungrouped variance estimates of the totals (1.8). Expectations are taken over the distribution of the target variable $\bar{y}_{hi}$ conditional on the realized sample selection and are dependent on the sampled PSU sizes N_{hi}.

In agreement with the theory, the simulations indicate that the advantages of the proposed algorithm for calculating standard errors of the *totals* are more pronounced for small and intermediate variabilities of the target variable in comparison with variability of PSU sizes. At the same time relative deviations between grouped and ungrouped estimates of the standard errors of the *means* are uniform for all grouping methods across all variabilities of the target variable. We explain this by the non-trivial relation between the variances of the means and totals, see Wolter (1985), p.236. We expect that variances of other estimates, such as regression coefficients and ratios, will also be robust to possible deviations from the proposed method for grouping strata. Our understanding of the reasons for deviations between grouped and ungrouped estimates of the standard errors was further validated by application to real survey data.

The main question of this research was: how reliable are variances estimated using grouped strata and PSU? Results of simulations presented in Figure 2 and Table 1 demonstrate in most cases good and robust finite population properties of the standard errors and coverage probabilities by the corresponding confidence intervals estimated using grouped survey design, after taking into consideration degrees of freedom reduction. Exceptions occur when larger PSU are preferentially combined to smaller PSU from different strata, violating the rules (4-7) of the algorithm described in Section 2. This results in significant bias of the estimated standard errors and deviation from the nominal coverage probabilities. Combining PSU at random, which is the case for grouping methods (2,3), may be acceptable when the number of groups is large enough. But sometimes estimates may be required for domains with a limited number of grouped strata, for example in Census regions. In this case following the rules (4-7) of the proposed algorithm could be important for unbiased estimation of standard errors.

Acknowledgment
Susan Schappert stimulated my interest in this problem and greatly improved the language of the paper. Discussions with Alan Dorfman helped me to clarify understanding of the methodology.

References

Lee, K. (1972). The Use of Partially Balanced Designs for Half-Sample Method of Variance Estimation. *Journal of the American Statistical Association*, 67:324–334.

Lee, K. (1973). Using Partially Balanced Designs for the Half-Sample Method of Variance Estimation. *Journal of the American Statistical Association*, 68:612–614.

Lu, W. W., Brick, M. J., and Sitter, R. R. (2006). Algorithms for Constructing Combined Strata Variance Estimators. *Journal of the American Statistical Association*, 101:1680–1692.

Mayda, J., Mohl, C., , and Tambay, J.-L. (1996). Variance estimation and confidentiality: They are related! In *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 135–141.

McCarthy, P. J. (1966). *Replication: An Approach to the Analysis of Data From Complex Surveys*. U.S. Government Printing Office, Wshington, DC.

Nixon, M., Brick, M., Kalton, G., and Lee, H. (1998). Alternative Variance Estimation Methods for the nhis. *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 326–331.

Rao, J.N.K.. and Shao, J. (1996). On balanced half-sampled variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91:343–348.

Rust, K. (1984). *Techniques for Estimating Variances for Sample Surveys*. PhD thesis, University of Michigan.

Rust, K. (1986). Efficient Replicated Variance Estimation. *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 81–87.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York, NY.