USE OF STRATUM MIXING TO REDUCE PRIMARY-UNIT-LEVEL IDENTIFICATION RISK IN PUBLIC-USE SURVEY DATASETS

John L. Eltinge[1]

The preparation and release of public-use datasets involve a balance between information needs of researchers and the general public, and legal and ethical requirements to protect the confidentiality of responses. This generally leads to restrictions on release of both individual data records and associated design information. For example, extensive previous literature has considered several methods of masking individual responses; and De Waal and Willenborg (1997) discussed identification risks associated with release of sampling weights.

This paper considers identification risks associated with inclusion of stratum and primary-sample-unit (PSU) labels in public-use microdata releases. In general, standard approximate design-based analyses of stratified cluster sample survey data require the use of stratum and PSU labels, or of closely related replicate weights. However, because PSUs are often identical to counties or contiguous groups of counties, it is sometimes possible to match the estimated demographic characteristics of a sample PSU with publicly known county-level demographic profiles. Thus, release of nominally uninformative stratum labels h = 1, 2, …, L and PSU labels i = 1, 2, … can lead to explicit identification of some sample PSUs, especially if a PSU has a demographic profile that is distinct from other PSUs in the population. This is problematic because identification of a specific PSU can substantially increase the risk of identification of some individual sample elements within that PSU. In addition, allowing identification of primary units is a violation of specific disclosure-limitation policies at some agencies.

To address the problem of PSU-level identification risk, this paper considers a relatively simple method for masking stratum and PSU labels. The principal idea is to combine markedly different strata and primary units into groups that are not readily matched with specific counties in the original population. This grouping does not affect customary sampling weights or calculation of point estimators, but does coarsen the information available for calculation of variance estimators. In particular, the grouping allows the computation of variance estimators that are approximately design unbiased, but that are less stable than customary design-based variance estimators computed from complete stratum and primary-unit information.

Key Words: Confidentiality; Design-based inference; Deterministic identification risk; Microdata disclosure risk; Microdata masking; PSU-level profile vector; Stochastic identification risk; Stratified multistage sample survey.

1. INTRODUCTION

1.1 Identification Risk Incurred Through the Release of Survey Microdata

Government statistical agencies often devote considerable effort to production of public-use survey microdata files. In many cases, release of these data can be an important component of scientific and policy research by governmental, academic, business and nonprofit organizations. However, government

[1] John L. Eltinge, Bureau of Labor Statistics and Texas A&M University. Office of Survey Methods Research, Bureau of Labor Statistics, PSB 4915, 2 Massachusetts Avenue NE, Washington, DC 20212; Eltinge_J@bls.gov. This work was supported in part by the National Center for Health Statistics. The views expressed here are those of the author and do not necessarily reflect the policy of the Bureau of Labor Statistics or the National Center for Health Statistics. The author thanks John Horm, Van Parsons and Al Zarate for many helpful comments on disclosure limitation issues.

statistical agencies have also recognized that they must balance legitimate research interests against legitimate privacy interests of survey respondents. For example, the United States has stringent legislation restricting the disclosure of Federal survey microdata, and survey data collection efforts often provide potential respondents with explicit assurances regarding data confidentiality.

In response to these issues, the statistical literature has developed a considerable body of work associated with the quantification and reduction of risks that arise in microdata release. For some general background, see, e.g., Paass (1988), Duncan and Lambert (1989), Skinner (1992), Fuller (1993), Lambert (1993), Fienberg (1994), Cox and Zayatz (1995), the December, 1998 special issue of *Journal of Official Statistics* on data confidentiality, and references cited therein.

Following Chen and Keller-McNulty (1998) and others, the present paper will consider *identification disclosure*, which occurs when a given sample unit in a microdata release is subsequently linked with an informative label e.g., a person's name or street address. In the literature to date, work with identification risk has tended to focus on identification of individual sample respondents, e.g., individual persons or establishments. A partial exception is DeWaal and Willenborg (1997), who considered cases in which the release of sampling weights could lead to the identification of specific population cells (e.g., poststrata) to which specific sample units belonged.

The present work will consider the risk of identification of primary sample units (PSUs). This PSU-level focus is of practical interest because identification of a specific PSU can substantially increase the risk of subsequent identification of survey respondents contained within that PSU. Consequently, some agencies have microdata release rules stated explicitly in terms of primary sample units and other aggregate geographical units. For example, the National Center for Health Statistics (1997, p. 19) rules include the following.

> "B. Geographic places that have fewer than 100,000 people are not to be identified on the tape.
> C. Characteristics of an area are not to appear on the tape if they would uniquely identify an area of
> less than 100,000 people.
> D. Information on the drawing of the sample which might assist in identifying a data subject must not be released outside the Center. Thus, the identities of primary sampling units are not to be made available outside the Center.''

This paper will focus on the risk that a given primary sample unit in a microdata file can be identified by matching the sample demographic profile of that PSU with known demographic characteristics of related counties or other groups of counties. These risks are of practical concern because some public data users may have strong substantive reasons to attempt to identify specific primary units, despite agency policies to the contrary. For example, users may want to link a specific PSU with related socioeconomic, health or environmental variables available from other data sources.

1.2 Release of Stratum and Primary Sample Unit Labels

A simple but problematic option for microdata release is to permit public access to the full data file, including selection probabilities, stratum and primary unit labels, and response variables y. In this setting, the only information omitted from this release would be the name and address of the respondent and other explicit geographical identifiers. This release method can be feasible for relatively large geographical areas, e.g., the entire United States. However, this approach can be problematic for smaller geographical areas that contain only a moderate number of PSUs. The fundamental problem is that for a smaller area, there are relatively few candidate PSUs that could have been selected. Thus, if each candidate PSU is known to be a single county, and if PSU-level characteristics are known for each county, then one can identify county labels for each sample PSU (h,i) as follows.

1. For each PSU (h,i), define the profile vector

$$m_{hi} = (m_{hi1}, \ldots, m_{hiK}) \tag{1.1}$$

where each of the K elements in (1.1) represent, e.g., population proportions for specified demographic cells, urban or rural classificiations, managed health care classifications, or other PSU-level means or proportions that are publicly known.

2. Use released survey weights w and observed sample values y to compute weighted estimates

$$\hat{m}_{hi} = (\hat{m}_{hi1},...,\hat{m}_{hiK}) \tag{1.2}$$

of the profile vector for PSU (h,i).

3. Of practical interest are cases in which K, the dimension of the profile vector, is relatively large and the distances among the known true profile vectors are large relative to the errors

$$\hat{m}_{hi} - m_{hi} \ . \tag{1.3}$$

For these cases, relatively simple methods (e.g., minimum distance matching) will allow the observed estimate (1.2) for a sampled primary unit (h,i) to be matched with its associated true known vector with high probability. This in turn means that all sample units contained in PSU (h,i) are then known to belong to the identified county with the profile vector (1.1)

In considering the disclosure risk presented by steps (1) through (3), there are three points of special interest. First, disclosure risk is high for PSUs (h,i) that have profile vectors (1.1) that are very distinct from the profile vectors of other PSUs included in the data release. Conversely, disclosure risk is lower for groups of PSUs with relatively similar profile vectors (1.1). Second, due to the reasoning outlined in point (3) above, the magnitude of PSU-level identification risk depends on the following.

a. The specific profile variables that are publicly known *and* are also included in the public data release.

b. The distribution of the true profile vectors (1.1) within the population covered by the data release.

c. The magnitude of the differences (1.3).

Third, one can consider reduction of PSU-level identification risk through a combination of the following.

i. Mixture of several true PSUs into artificial pseudo-PSUs that reduce the profile-vector separation described above, without excessive degradation of the quality of legitimate full-population analyses.

ii. Suppression of variables y for which the PSU level means (1.1) are publicly known and are very distinct across PSUs.

1.3 Reduction of Identification Risk Through Stratum Mixing

The remainder of this paper examines some specific ways in which to implement steps (i) and (ii). Section 2 introduces a specific microdata-disclosure application involving subnational-level data releases for the National Health Interview Survey (NHIS), and reviews standard design-based variance estimation methods applicable to non-masked data produced by the NHIS. Section 3 reviews two competing criteria that are important in evaluation of any identification-reduction method: identification risk and inferential efficiency. Section 4 introduces a method, known as stratum mixing, for reduction of PSU-level identification risk. Section 5 closes with some additional comments on the main ideas in this paper.

2. DESIGN-BASED ANALYSES OF SUBNATIONAL DATA RELEASES FROM THE NATIONAL HEALTH INTERVIEW SURVEY

This work was motivated by the anticipated release of National Health Interview Survey (NHIS) microdata files for specific subnational level areas. This release is of practical interest because for some public health issues, subnational level analyses can potentially provide a valuable complement to customary national level analyses.

Data for the NHIS are collected through a stratified multistage sample design, with primary sample units generally identical to counties or groups of contiguous counties. In general, the National Center for Health Statistics has recommended that NHIS data be analyzed through customary design-based methods. For some general background on design-based analysis of data from the 1985-1994 NHIS design, see Casady, Parsons and Snowden (1986) and Parsons and Casady (1986). Closely related ideas apply to the analysis of data from the 1995-2004 NHIS design; see Judkins, Marker and Waskberg (1996).

In particular, consider a stratified multistage design with L strata and with n_h primary sample units selected from stratum h with replacement, h = 1, …, L. We emphasize here that the term "primary unit" refers to the first stage at which non-certainty selection has taken place. In addition, define the population total $Y = \sum_{h=1}^{L} Y_h$ where $Y_h = \sum_{i=1}^{N_h} Y_{hi}$; $Y_{hi} = \sum_{j \in U_{hi}} Y_{hij}$; N_h is the number of primary units in stratum h ; U_{hi} is the set of elements in primary unit (h,i); and Y_{hij} is a survey item associated with element j in primary unit (h,i). Also, define the point estimator, $\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$, where $\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} / n_h$ and $\hat{Y}_{hi}$ is the customary design unbiased point estimator of Y_h based only on data from primary unit (h,i).

Due to independence of sampling across strata,

$$V(\hat{Y}) = \sum_{h=1}^{L} V(\hat{Y}\ ).$$

In addition, with-replacement sampling of primary units within strata implies that for a given stratum h, the point estimators $\hat{Y}_{hi}$ are independent and identically distributed with mean Y_h ; and that an unbiased estimator of $V(\hat{Y}_h)$ is $\hat{V}(\hat{Y}_h) = \sum_{i=1}^{n_h} (\hat{Y}_{hi} - \hat{Y}_h)^2 / n_h(n_h - 1)$. Consequently, an unbiased estimator of $V(\hat{Y})$ is $\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \hat{V}(\hat{Y}_h)$. For simplicity, this development has restricted attention to simple weighted estimators of the population total Y. Related variance estimation methods apply to smooth nonlinear functions of population totals, e.g., means, ratios or solutions to standard estimating equations; see, e.g., Cochran (1977), Krewski and Rao (1981) and Shao (1996).

In addition, we emphasize that consideration here is restricted to subnational areas from which we have selected a relatively large number of primary sample units. For areas covered by a moderate or small (e.g., less than 20) number of selected PSUs, the limitation on the number of degrees of freedom for the associated design-based variance estimator can degrade the performance of the resulting inference methods (e.g., inflate confidence interval widths or reduce test power). For such cases, it may be advisable to use alternatives to standard design-based variance estimators. Examples include generalized variance functions (Wolter, 1985, Chapter 5) or average design effect approximations. Also, we exclude from consideration here analyses of small subnational areas, for which the realized sample size in a given year may be small enough that the resulting sampling variability of standard *point* estimators is unacceptably large. For such cases, it may be advisable to focus analytic effort on larger geographical areas, or on multiple-year analyses in the original area.

3. EVALUATION OF IDENTIFICATION RISK REDUCTION METHODS

In general, masking of public-release microdata involves a balance between the following two potentially conflicting goals.

1. Reduce (to the extent possible) the information provided at an undesirably fine level of aggregation. For example, one would want to ensure that for a given released record j, the probability that this record can be matched with a specific identified person is not much larger than the probability of a correct match associated with purely random guessing. In practical settings, direct evaluation of such risks is problematic and depends on the specific matching strategy employed by the data user; see, e.g., Fienberg et al. (1997) and references cited therein.

2. Maintain (to the extent feasible) the information available for estimands defined at appropriately high levels of aggregation. In practical terms, this involves satisfactory performance of design-based point estimators and confidence intervals for selected parameters of the subnational population.

An ideal masking method would achieve a maximum information reduction according to criterion (1) while incurring little or no information loss according to criterion (2). The extent to which a given masking method satisfies these criteria will depend on the specific measures of information content used in (1) and (2), and on the properties of the sample design and population. Sections 3.1 and 3.2, respectively, will consider some examples of criteria (1) and (2) applied to PSU-level identification risk.

3.1 Reduction of Matching Probabilities

One could consider several different criteria to evaluate PSU-level identification risk. For a general discussion of related criteria developed previously for element-level identification risk, see the references listed in Section 1.1. For the present work, consider

$$P(E \in U \mid x, w, d)\,, \tag{3.1}$$

the probability that an element E is contained in a primary sample unit U, conditional on a released observation dataset x (e.g.., element-level questionnaire responses), a released set of survey weights *w*, and additional released design information *d* (e.g., stratum and primary-unit labels, or associated replicate weights).

We note three points regarding the conditional probabilities (3.1); additional related results are in Eltinge (1999). First, purely random "guessing" will result in the assignment of some elements to their correct PSUs. For example, given u primary sample units in a released subnational area, random (equal probability) assignment of PSU names to elements E would (in large samples) result in the assignment of a proportion 1/u of the elements in each PSU to the correct PSU. In this case, if expression (3.1) is approximately equal to 1/u for all elements E and all PSUs U, then the design information d has been masked as much as is reasonably possible.

Second, consider a current public-data release of observations x and weights w. Then in weighing the risks attendant in the release of design information d, principal interest focuses on the incremental increase in identification risk (3.1), beyond the risk incurred by the current weight-and-data-only release. Define the incremental identification risk incurred by the release of design information d as,

$$I_{dEU} = P(E \in U \mid x, w, d) - P(E \in U \mid x, w) \tag{3.2}$$

According to this criterion, the design information *d* would be ideally masked if $I_{dEU} = 0$ for all E and U.

Third, note that the specific value of expression (3.2) will vary across i and j. Thus, a given release d will induce a distribution of values (3.2), taken across the population of elements E, for a given unit U. This distribution would then lead to several possible summary evaluation criteria. For example,

the maximum of the absolute value of expression (3.2) across all elements would be a very conservative criterion. The mean, median or upper quantiles (e.g., the 90th or 75th percentiles) of the absolute value of expression (3.2) would be somewhat less conservative. In addition, note that masking methods that focus on deletion or modification of a small or moderate number of outlying data points would be consistent with an evaluation criterion that focused on the maximum or upper tail quantiles of expression (3.2).

3.2 Loss of Inferential Efficiency

Consider a specific parameter q associated with the full population or with a subnational area satisfying the disclosure limitation criteria considered in Section 3.1. Ideally, a masking procedure would produce the same point estimator, variance estimator and confidence interval as would be produced in a customary design-based analysis of the original (non-masked) dataset. For masking methods that retain the original observations y and weights w, but modify the design information d, masking will affect the analysis primarily through the variance estimator

$$\hat{V}_m(\hat{q}) \tag{3.3}$$

where expression (3.3) is computed from the modified design information provided in the masked dataset. Direct assessment of variance estimator bias follows from evaluation of the *misspecification effect*

$$meff(\hat{V}_m) = \hat{V} / \hat{V}_m$$

where $\hat{V}$ denotes the design-based variance estimator for $\hat{q}$ computed from the non-masked design information; cf. Skinner (1989). Similarly, assessment of variance estimator stability follows from the related degrees-of-freedom term f for the masked-data variance estimator (3.3).

In formal inferential work, variance estimation generally is viewed as an intermediate step toward construction of a confidence interval or test statistic. Consequently, it is arguably preferable to focus directly on the performance of the nominal (1 - a)100% confidence intervals,

$$(\hat{q}_{mL}, \hat{q}_{mU}) = \hat{q} \pm t_{f,1-a/2}(\hat{V}_m)^{1/2}$$

where $t_{f,1-a/2}$ is the upper $1-a/2$ quantile of a t distribution on f degrees of freedom. Of principal interest are the true coverage rate $1-h$, say, and the distribution of the interval widths $2t_{f,1-a/2}(\hat{V}_m)^{1/2}$.

4. STRATUM MIXING TO REDUCE PSU-LEVEL IDENTIFICATION RISK

4.1 Stratum Mixing

To reduce the risk of identification of selected PSUs in a released dataset, one may group the original true strata and true PSUs into coarser pseudo-strata and pseudo-PSUs. To develop this idea, consider again the complex sample design and point estimators reviewed in Section 2. For simplicity of notation, assume now that L is even and that each n_h is even. Then one relatively simple way to combine strata and PSUs is as follows. First, pair the L old strata into L/2 new pseudo-strata. For notational convenience, assume that this pairing matches strata consecutively in the index h, to produce the pairs, (1, 2), (3, 4), …, (L-1, L). Section 4.2 below considers some specific methods for determining these pairings.

Second, for each old stratum h, randomly partition the PSUs into two groups, denoted $G_{h(1)}$ and $G_{h(2)}$, each with $n_h / 2$ selected PSUs. Then for the new pseudo-stratum g containing the old strata 2g-1 and 2g, define the two new pseudo-primary units,

$$G^*_{g(1)} = G_{2g-1(1)} \cup G_{2g(1)} \text{ and } G^*_{g(2)} = G_{2g-1(2)} \cup G_{2g(2)}$$

where $\cup$ denotes the union of sets.

Third, define the associated point estimators

$$Y^*_{g(m)} = \sum_{i \in G_{2g-1(m)}} 2\hat{Y}_{2g-1,i} / n_{2g-1} + \sum_{i \in G_{2g(m)}} 2\hat{Y}_{2g,i} / n_{2g} \qquad (4.1)$$

where, as in Section 2, $\hat{Y}_{hi}$ is a design unbiased estimator of the population total Y_h, based only on data from primary unit i in stratum $h = 2g-1, 2g$. Note that $Y^*_{g(m)}$ is based only on data from the new pseudo-stratum $G^*_{g(m)}$. In addition, expression (4.1) is design unbiased for $Y_{2g-1} + Y_{2g}$; and

$$Y^* = \sum_{g=1}^{L/2} (Y^*_{g(1)} + Y^*_{g(2)}) / 2$$

is design unbiased for the population total Y, and is algebraically identical to $\hat{Y}$.

Fourth, due to the with-replacement sampling of PSUs in the original design, $Y^*_{g(1)}$ and $Y^*_{g(2)}$ are independent and identically distributed random variables. Thus, a design unbiased estimator of $V(Y^*)$ is

$$\hat{V}_m = \sum_{g=1}^{L/2} (Y^*_{g(1)} - Y^*_{g(2)})^2 / 4$$

The preceding discussion focused on simple pairwise combinations of PSUs. In principle, one could consider three-fold or higher-order levels of combination, especially for strata in which disclosure risks appear to be especially problematic.

4.2 Three Forms of Stratum Mixing

The preceding discussion took as given the ordering of strata $1,...,L$ that led to the pairings, $(1,2),(3,4),...,(L-1,L)$. Three specific methods for ordering and pairing strata are as follows.

Purely random mixing. In purely random mixing, the stratum ordering $1,...,L$ is determined through a random permutation (e.g.., Kennedy and Gentle, 1980, p. 241) of the original stratum labels.

Deterministic mixing. Here, the pairing of strata is based on criteria that are purely deterministic, e.g., the distance between strata measured on a metric determined by variables that are known before data collection takes place.

Data-driven mixing. In data-driven mixing, the pseudo-strata are formed by pairing original strata that have mean profile vectors that are far apart.

4.3 Distinctions Between Stratum Mixing and Customary Stratum Collapse

Stratum mixing is similar to customary stratum collapse (see, e.g., Rust and Kalton, 1987; Wolter, 1985, Section 2.5; Hartley, Rao and Kiefer, 1969; Hansen, Hurwitz and Madow, 1953; and references cited therein), in the *limited* sense that both operations construct variance estimators by combining information across the strata used in the original sample design. However, the stratum mixing method introduced in Section 4.1 is conceptually and operationally distinct from customary stratum collapse in three ways. First, stratum collapse generally is carried out because the original sample design involved selection of a single primary sample unit per stratum, so that standard design-based variance estimators could not be computed from the resulting survey data. In stratum mixing, the original design involves two or more

primary units selected per stratum, and customary design-based variance estimators could have been computed, provided one had access to the relevant original stratum and PSU labels.

Second, under stratum collapse with a one-PSU-per-stratum design, all elements sampled from a given stratum are placed in the same pseudo-PSU. Under stratum mixing, the set of n_h primary units contained in the original stratum h is randomly partitioned into groups assigned to the first and second pseudo-PSUs, respectively, of the new pseudo-stratum g containing the original stratum h.

Third, under mild regularity conditions, a stratum-collapse-based variance estimator is conservative, i.e., has an expectation that is greater than or equal to the true variance. Stated briefly, the potential positive bias arises from the fact that collapse-based variance estimators will involve squared differences of random variables with different means. On the other hand, deterministic and purely random stratum mixing use squared differences of random variables with the same means. This equal-mean condition in turn leads to some approximate unbiasedness properties described in Section 4.4 below.

4.4 Reduction of Matching Probabilities and Reduction of Inferential Efficiency

Recall from Section 1.2 that release of design information d led to an increase in PSU level identification risk, through possible matching of known true-PSU profile vectors m_{hi} with estimated PSU-level profile vectors $\hat{m}_{hi}$. This risk can be reduced through the Section 4.1 method of stratum mixing. For example, consider the extreme case in which all L combined pseudo-PSUs (g,i) have the same expectation for the weighted sample profile vector $\hat{m}_{gi}$. Then according to the profile-matching criterion, the release of the pseudo-stratum and pseudo-PSU labels will result in zero incremental increase in the risk of disclosure. For a given real dataset, one would seek to approximate this idealized result as closely as possible. To achieve this, one method would be to pair strata in a way that would minimize the sum of squared distances among the resulting weighted sample vectors $\hat{m}_g$. Note that in an informal sense, this is the coverse of standard multivariate cluster-formation algorithms, which are intended to form clusters that are as far apart as possible. Here, we seek to form pseudo-strata by pairing original strata that were far apart; under conditions, this ensures that the resulting combined pseudo-strata are close together, relative to the distribution of the original strata.

The inferential performance of stratum mixing is driven in large part by the expectations and variances of the resulting variance estimators. Under mild regularity conditions, deterministic and purely random mixing lead to variance estimators that are approximately design unbiased, but are less stable than customary design-based variance estimators. Under additional regularity conditions (primarily involving normality of PSU-level estimators within the original strata), data-driven mixing has properties similar to those of deterministic mixing and purely random mixing.

The abovementioned loss of variance estimator stability is attributable to the reduction in the effective number of PSUs arising from the use of stratum mixing. The resulting loss in inferential efficiency (e.g.,. through the resulting inflation in standard design-based confidence interval widths or the reduction of the power of design-based tests) is the price incurred by restricting the amount of design information included in the public data release. In cases for which L, the remaining effective number of degrees of freedom, is moderate or large (e.g., greater than 20), the loss of inferential efficiency can be relatively modest. For cases in which L is smaller, but a relatively large proportion of population is in self-representing strata, the loss of degrees of freedom can be mitigated through a more extensive partition of the self-representing strata into finer groups; see, e.g., Parsons and Eltinge (1999).

5. DISCUSSION

In closing, we note three additional points. First, the discussion in Sections 1 through 4 focused on profile vectors involving means or proportions. In principle, these profile vectors could also include more complex parameters, e.g., population variances. If there is public knowledge of the distribution of a distinctly non-normal continuous variable (e.g., income), then one could also consider inclusion of

specified population quantiles in the profile vector. Matching work then could also involve comparison of distributions through design-based forms of Cramer-von Mises or Shapiro-Wilk-Francia statistics.

Second, empirical performance of the proposed methods will depend heavily on: (1) the specific variables included in the profile vector; and (2) the distribution of these variables within PSUs, across PSUs within strata, and across strata. In particular, note that if the original PSU labels are highly informative for the elements of the profile vector, then it would be especially important to account for the design in the data analysis. However, this same case would be one in which it would be especially important to mask the original PSU membership. Consequently, in considering variables for inclusion in a candidate profile vector, there is special interest in variables for which PSU labels are informative.

Finally, evaluations of stratum-mixing methods lead to discussion of deterministic and stochastic criteria for evaluation of identification risk. Deterministic criteria are of special interest for cases in which the available public data (e.g., the abovementioned county-level demographic profiles) are very rich, leading to unambiguous matches of sample PSUs with known counties. However, in some applications the PSU-level sample profile vectors have limited dimensions and display sampling variability that is not small relative to differences among the publicly known true profile vectors. This in turn implies that some PSUs can be correctly matched with specified counties with probabilities that are high but not equal to one. Consequently, it is important to consider stochastic evaluation criteria to identify these high-probability cases, and to quantify associated risks. These risks involve trade-offs among: (1) sampling variability in PSU-level sample profile vectors; (2) the distribution of public profile vectors; and (3) differences between the public profile vectors and the expectation of the sample profile vectors, due to, e.g., definitional issues, aggregation issues or time effects.

REFERENCES

Casady, R.J., Parsons, V.L. and Snowden, C.B. (1986). Simplified variance estimation for the National Health Interview Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 412-417.

Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* **14**, 79-95.

Cochran, W.G. (1977). *Sampling Techniques, Third Edition.* New York: Wiley.

Cox, L.H. and Zayatz, L.V. (1995). An agenda for research in statistical disclosure limitation. *Journal of Official Statistics* **11**, 205-220.

De Waal, A.G. and Willenborg, L.C.R.J. (1997). Statistical disclosure control and sampling weights. *Journal of Official Statistics* **13**, 417-434.

Duncan, G. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7**, 207-217.

Eltinge, J.L. (1999). Evaluation and reduction of cluster-level identification risk for public-use survey microdata files. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* to appear.

Fienberg, S.E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics* **10**, 115-132.

Fienberg, S.E., Makov, U.E. and Sanil, A.P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics* **13**, 75-89.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383-406.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory, Volumes I and II.* New York: Wiley.

Hartley, H.O., Rao, J.N.K. and Kiefer, G. (1969). Estimation with one unit per stratum. *Journal of the American Statistical Association* **64**, 841-851.

Judkins, D.R., Marker, D. and Waksberg, J. (1996). *National Health Interview Survey: Research for the 1995 Design.* Report from Westat, Inc. to U.S. National Center for Health Statistics under contract no. 200-89-7021. Rockville, Maryland: Westat, Inc.

Kennedy, W.J. and Gentle, J.E. (1980). *Statistical Computing*. New York: Marcel Dekker.

Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics* **9**, 1010-1019.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics* **9**, 313-331.

National Center for Health Statistics (1997). *NCHS Staff Manual on Confidentiality*. Hyattsville, Maryland: National Center for Health Statistics.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics* **6**, 487-500.

Parsons, V.L. and Casady, R.J. (1986). Variance estimation and the redesigned National Health Interview Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 406-411.

Parsons, V.L. and Eltinge, J.L. (1999). Stratum partition, collapse and mixing in construction of balanced repeated replication variance estimators. Paper presented at the 1999 Joint Statistical Meetings, Baltimore, Maryland.

Rust, K. and Kalton, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics* **3**, 69-81.

Shao, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics* **27**, 203-254.

Skinner, C.J. (1989). Introduction to Part A, pp. 23-58 in C.J. Skinner, D. Holt and T.M.F. Smith, eds., *Analysis of Complex Survey Data*. New York: Wiley.

Skinner, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica* **46**, 21-32.

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