NISS

The Effect of Statistical Disclosure Limitation on Parameter Estimation for a Finite Population

Hang J. Kim and Alan F. Karr

Technical Report 183
October 2013

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709
www.niss.org

The Effect of Statistical Disclosure Limitation on Parameter Estimation for a Finite Population

Hang J. Kim

Duke University and National Institute of Statistical Sciences, Durham, NC 27708

(*hangkim@niss.org*)

Alan F. Karr

National Institute of Statistical Sciences, Research Triangle Park, NC 27709 (*karr@niss.org*)

Word counts of text: 4,897

ABSTRACT

In this paper we study the impact of statistical disclosure limitation in the setting of parameter estimation for a finite population. Using a simulation experiment with microdata from the 2010 American Community Survey, we demonstrate a framework for applying risk-utility paradigms to microdata for a finite population, which incorporates a utility measure based on estimators with survey weights and risk measures based on record linkage techniques with composite variables. The simulation study shows a special caution on variance estimation for finite populations with the released data that are masked by statistical disclosure limitation. We also compare various disclosure limitation methods including a modified version of microaggregation that accommodates survey weights. The results confirm previous findings that a two-stage procedure, microaggregation with adding noise, is effective in terms of data utility and disclosure risk.

KEY WORDS: American Community Survey; Data utility; Disclosure risk; Microaggregation; Risk-utility paradigm; Replicate weights

1. INTRODUCTION

When disseminating public survey or census data, many national statistical agencies and survey organizations publish the data in the form of tables, as well as records of individual respondents, also known as *microdata*. Due to analysis flexibility and details of the individual level information, a microdata set facilitates research by *legitimate users*, persons who use data without attempting to violate privacy or confidentiality. However, microdata can also be used by *intruders*, persons who try to reveal subjects' identities or values of sensitive variables (Cox et al. 2011). Therefore, statistical agencies must balance disseminating high-quality data with protecting data confidentiality and data subjects' privacy.

Statistical agencies often mask actual values of microdata before release by means of various statistical disclosure limitation (SDL) methods, such as adding noise, data swapping and microaggregation. Typically, there can be multiple candidates for the final data release, which arise from different choices of SDL techniques and their parameters, or from different randomizations. As a framework to choose the best release, or at least good releases, from the candidates, the so-called risk-utility (R-U) framework has been studied extensively (Duncan and Stokes 2004; Gomatam et al. 2005; Cox et al. 2011). The R-U framework helps the agencies find a feasible release between the two extremes of zero risk when disseminating no data and maximum utility when releasing the original dataset.

Despite the abundance of SDL literature, there has been limited attention to the role of *survey weights* (or simply *weights*) in the context of data utility of masked data from SDL implementation (Cox et al. 2011). In real world, datasets generated by simple random sampling (SRS) are rare; for complex designs, respondents in the survey have different weights. The base weights are typically set initially as inverses of the probability

of selection, but "final weights" may also reflect such phenomena as nonresponse, attrition in panel surveys, and poststratification. Weights are arguably not necessary for some analyses of survey data (Fienberg 2009), but they are essential for estimating a parameter of a finite population such as national total income. Most previous researchers evaluated data utility after SDL methods without consideration of survey weights (Fienberg 2009), for example, using simple (unweighted) mean estimator or K-L divergence as utility measures.

In this paper we study impact of SDL methods on parameter estimation for a finite population by a simulation experiment using microdata from the 2010 American Community Survey (ACS). Comparing the variance estimators of original data and masked data, we argue for special caution of using standard error estimator produced from the masked data using replicate weights. We also demonstrate a framework for applying risk-utility paradigms to survey data of a finite population, which incorporates a utility measure based on estimators with weights and risk measures based on record linkage techniques with composite variables. To accommodate survey weights, we propose a modified version of microaggregation.

The remainder of the paper is organized as follows. In Section 2, we review some SDL methods for microdata, the risk-utility frontier, and impact of releasing survey weights on data disclosure. In Section 3, we present a risk-utility framework applicable to survey samples of finite populations with the illustrative example using the 2010 ACS data. Section 4 presents the results of the simulation studies. Lastly, Section 5 concludes the paper with a brief discussion.

2. BACKGROUND

2.1 Notations

We evaluate disclosure risk based on *record linkage* techniques. Similar to the notation of Reiter (2005), let y_{il} denote the value of variable l for record i, for $l = 1, \ldots, p$ and $i = 1, \ldots, n$. Denote the datafile of the originally reported values by D which consists of n rows of $\boldsymbol{y}_i = \{y_{i1}, \cdots, y_{ip}\}$. We partition the vector of record i as $(\boldsymbol{y}_i^A, \boldsymbol{y}_i^U)$ where $\boldsymbol{y}_i^A$ is available to an intruder from his external datafile D_{ext} as well as from the released datafile D_{rel} by the agency and $\boldsymbol{y}_i^U$ is unavailable to the intruder except in D_{rel}.

Denote the unique unit identifier of record i in the original datafile D by y_{i0}, which is never released by the agency, and the unit identifier of record j in D_{ext} by y_{j0}. To disclose sensitive information of a target record, the intruder first attempts to identify the target record j in D_{ext} from D_{rel}, i.e. to find i such that $y_{i0} = y_{j0}$ (called *re-identification*). To prevent this, the agency changes the values of $\boldsymbol{y}_i^A$ into $\tilde{\boldsymbol{y}}_i^A$ before releasing with some protection methods. Now, D_{rel} denotes the released datafile from agency which consists of n rows of vectors $\tilde{\boldsymbol{y}}_i = (\tilde{\boldsymbol{y}}_i^A, \boldsymbol{y}_i^U)$.

2.2 SDL Methods for Microdata

In this section, we introduce some perturbative methods for microdata. This section is not a comprehensive review of SDL methods, but rather background on a set of methods studied previously in R-U settings (Karr et al. 2006; Oganian and Karr 2006; Woo et al. 2009; Cox et al. 2011). We refer readers to Willenborg and De Waal (2001) for other SDL methods for microdata and further discussion.

Adding Noise. A number of authors propose to add random noise to numerical data so that exact values of sensitive variables or subjects' identities cannot be identified by

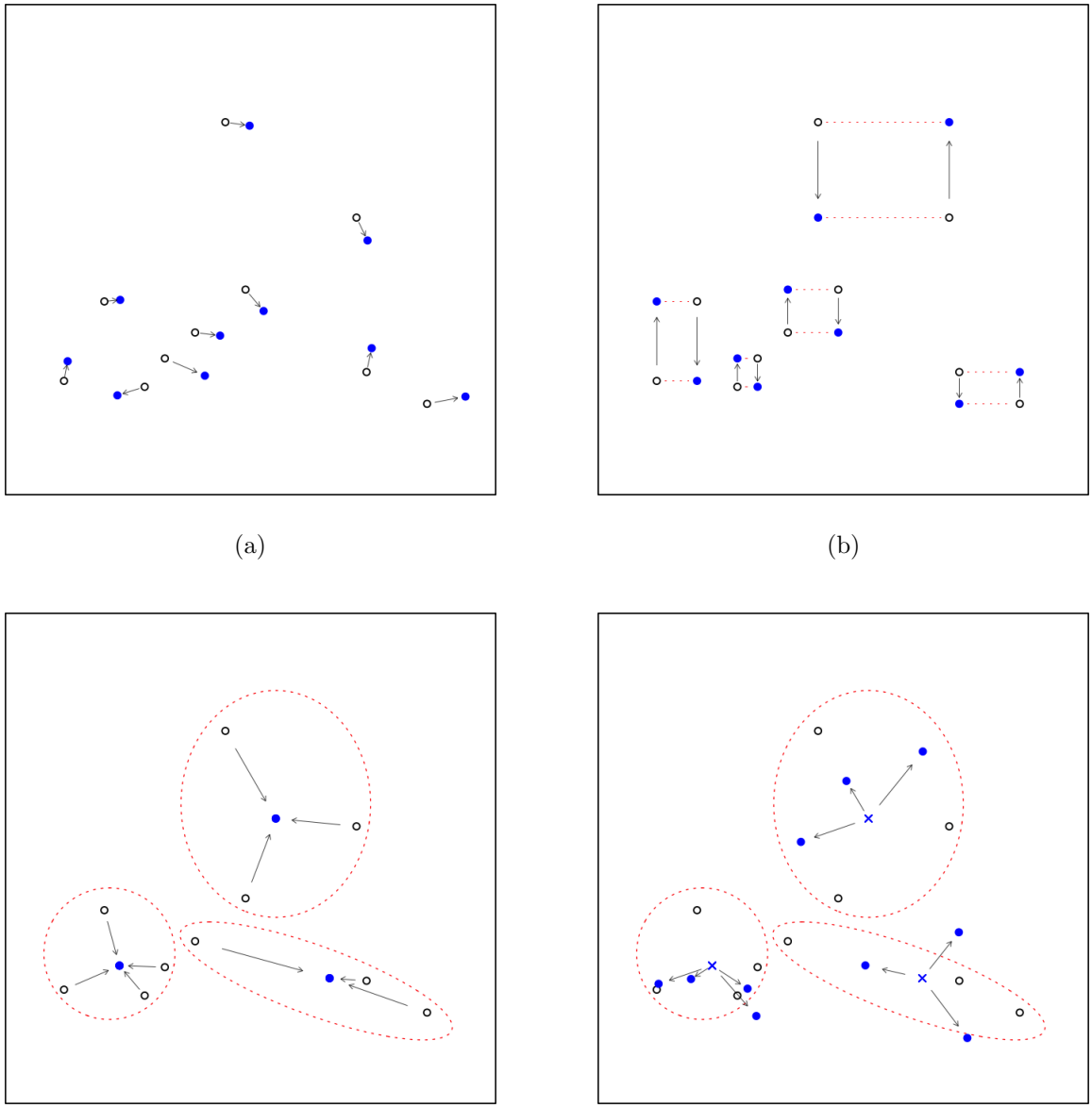

Figure 1: Illustration of four SDL methods: (a) adding noise, (b) rank swapping, (c) microaggregation, and (d) microaggregation with adding noise. Empty circles indicate the original records and solid circles indicate the masked records after SDL. For rank swapping in panel (b), only the variable of Y-axis is swapped. Panel (d) shows the two-step procedure where '$\times$' represent the masked values by microaggregation and the solid circles represent the values after adding random noise to '$\times$'.

intruders (Kim 1986; Sullivan and Fuller 1990; Tendick 1991). The masked record with *adding noise* is given by $\tilde{\boldsymbol{y}}_i^A = \boldsymbol{y}_i^A + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n$, where $\boldsymbol{\varepsilon}_i$ is the added noise. Typically, $\boldsymbol{\varepsilon}_i$ is assumed to follow $N(0, c^{\text{noise}}\ \Sigma^A)$, where Σ^A is the covariance matrix of $\{\boldsymbol{y}_1^A, \ldots, \boldsymbol{y}_n^A\}$ and c^{noise} is a positive constant. For example, Figure 1 (a) shows $\tilde{\boldsymbol{y}}_i^A$ as empty circles and $\boldsymbol{y}_i^A$ as solid circles. As c^{noise} increases, the amount of disclosure protection increases while data utility decreases.

Rank Swapping. Data swapping is to switch one or more attributes between randomly selected pairs of records. As a special form of data swapping, Moore (1996) suggests *rank swapping* to retain the dependence structure among variables. Instead of arbitrary swapping, rank swapping allows switching values within a range defined by a pre-determined parameter c^{rank}, with $0 < c^{\text{rank}} < 100$. For each variable l in $\boldsymbol{y}^A$, the method is implemented as follows:

(i) Sort $\{y_{1l}, \ldots, y_{nl}\}$ by its size and denote the sorted values by $\{y_{(1)l}, \ldots, y_{(n)l}\}$.

(ii) Randomly pick up $y_{(i)l}$ and $y_{(i')l}$. If the percentage difference of the indices is less than c^{rank}, i.e., $|i - i'| < n\ c^{\text{rank}}/100$, then swap the values and flag them.

(iii) Repeat (ii) with unflagged values until all records (or $n - 1$ records when n is odd) are swapped.

Figure 1 (b) shows $\tilde{\boldsymbol{y}}_i = (y_i^U, \tilde{y}_i^A)$ as empty circles and $\boldsymbol{y}_i = (y_i^U, y_i^A)$ as solid circles where X-axis and Y-axis represent the sets of U and A, respectively, i.e., only the variable of Y-axis is used for swapping. Rank swapping reduces disclosure risk because an intruder cannot be certain that any record is real; yet, it often distorts data when c^{rank} is large, and so reduces data utility.

Microaggregation. In *microaggregation*, the data are partitioned into groups, and the original response $\boldsymbol{y}_i^A$ is replaced by the average of subjects in its group $\mathcal{G}_i$. Figure 1

(c) shows an example of microaggregation. First, the original values represented by empty circles are grouped by a clustering method. The dotted curve indicates each cluster. Then, we calculate the average of the values in each cluster, which is represented by a solid circle. Lastly, $\boldsymbol{y}_i^A$ is replaced by the group average, i.e., $\tilde{\boldsymbol{y}}_{\text{mic},i}^A = \sum_{i' \in \mathcal{G}_i} \boldsymbol{y}_{i'}^A / |\mathcal{G}_i|$, where $c^{\text{mic}} = |\mathcal{G}_i|$ is the cardinality of $\mathcal{G}_i$. We refer readers to Fayyoumi and Oommen (2010) for variants of microaggregation with different clustering methods.

Microaggregation With Adding Noise. Microaggregation does not change sample means, but does decrease variances. As a solution to recover the variability taken out by microaggregation, Oganian and Karr (2006) suggest a two-step procedure to combine microaggregation and adding noise. Let $\tilde{\boldsymbol{y}}_{\text{mic},i}^A$ be masked records by microaggregation, plotted as '$\times$' in Figure 1 (d), and let $\tilde{\boldsymbol{y}}_i^A$ be the masked records by the two-step procedure, *microaggregation with adding noise*, plotted as solid circles in the figure. Then, the combined approach can be expressed as

$$\tilde{\boldsymbol{y}}_i^A = \tilde{\boldsymbol{y}}_{\text{mic},i}^A + \boldsymbol{\delta}_i$$

where $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \Sigma^{MA})$ and Σ^{MA} is the size of noise to recover variability lost during microaggregation. Oganian and Karr (2006) suggest to use $\Sigma^{MA} = \Sigma^A - \Sigma^M$ where Σ^M denotes the covariance matrix of $\{\tilde{\boldsymbol{y}}_{\text{mic},1}^A, \cdots, \tilde{\boldsymbol{y}}_{\text{mic},n}^A\}$.

2.3 Risk-utility Framework

The risk-utility framework helps a statistical agency to reason about identifying good, or even optimal, released datasets among candidate releases created from different choices of SDL methods or different parameters within a single SDL method, or different randomizations. The framework is imperfect in the sense that, while it does tell an agency how to think, it is often not concrete enough to tell the agency how to act (Cox

et al. 2011). We employ it nevertheless.

There are three conceptual components to the R-U framework: a quantified measure of disclosure risk, a quantified measure of data utility, and a value function that relates the two. In Figure 2, the dots correspond to different candidate releases, each with a risk measure and a utility measure. As the figure suggests, increasing utility is generally associated with increasing risk. The contours, or indifference curves, of the value function exhibit typical structure—value increases as risk decreases or utility increases, the other held fixed. The convexity of these curves shows decreasing marginal returns. For instance, when utility is high, a small increase in risk is acceptable only if there is a large accompanying increase in utility.

Perhaps the most important feature in Figure 2 is the *risk-utility frontier*, which consists of those candidate releases for which there is no other candidate that has both higher utility and lower risk. This frontier is analogous to production possibility curves or efficient frontiers in microeconomics. Of the eighteen candidates in the figure, the three candidates on the frontier are the only ones that need to be considered as releases. For every other candidate, there is a release on the frontier that dominates it in the sense of having both higher utility and lower risk. A specific value function makes the agency find the final release from three candidates; even if the agency is unable to specify a value function, its decision has been simplified dramatically, as the agency only needs to consider the three candidates on the frontier.

2.4 Impact of Survey Weights on Data Confidentiality

This paper deals with the impact of SDL on data utility for analyses involving survey weights. It is also possible that the weights themselves threaten disclosure risk. To illustrate, in the geographical aggregation analyses for use of chemicals (herbicides and pesticides) on agricultural crops reported in Karr et al. (2001), the weights were the most

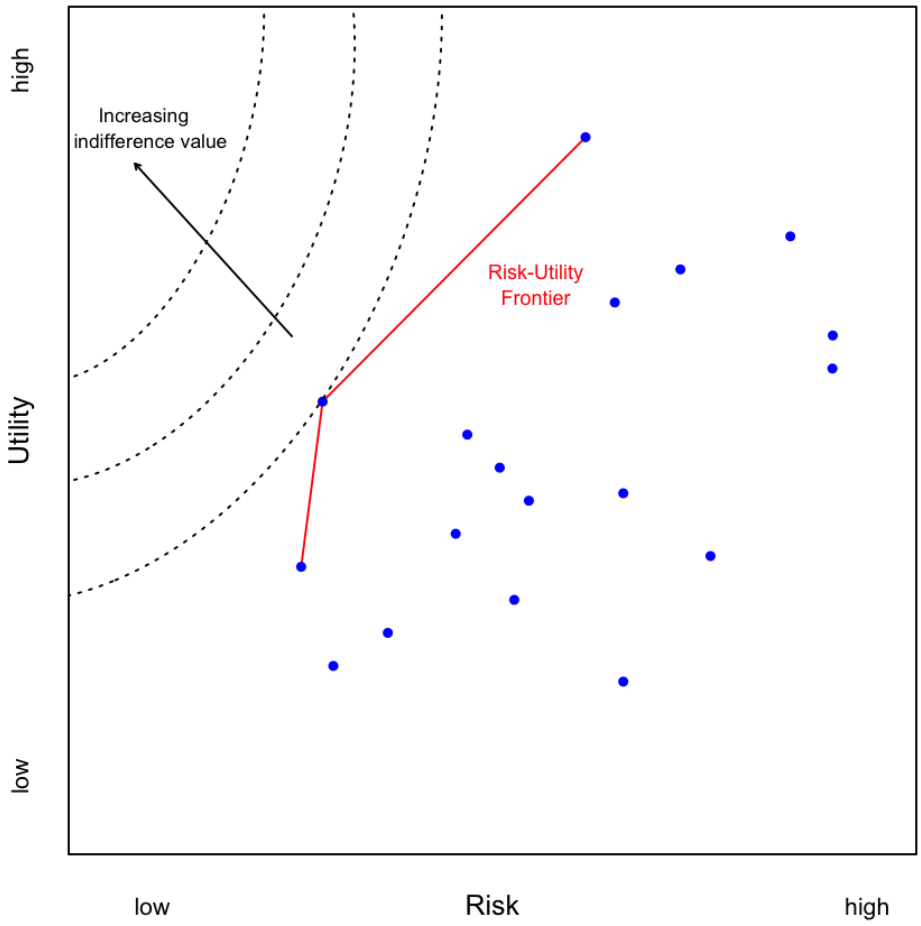

Figure 2: Description of the R-U framework. The solid circles represent different candidate releases. The solid line linking three circles indicates the risk-utility frontier. The dotted curves are the indifference curves of the value function.

sensitive values (more than then the chemical use levels), and were not released, although the released summaries were weighted.

Weights may also contain information about design variables that are not released (Willenborg and De Waal 1996; De Waal and Willenborg 1997). This problem is demonstrated in Cox et al. (2011) where the original weights are tied closely to design strata. However, a modification of microaggregation called *indexed microaggregation*, in which the clustering is based on weights, but the response variables are altered, reduces the risk substantially. In particular, the relationship between weight and stratum is attenuated dramatically, but with almost no impact on estimates of means and covariances. Poststratification is also an issue, because use of known population controls can allow re-identification of data subjects (Willenborg and De Waal 1996, 2001).

On the other hand, Fienberg (2009) argues that only a few studies show that releasing weights increases disclosure risk by non-negligible amount, and suggests instead that model-based rather than design-based analyses would allow weights to be suppressed and obviate risks associated with them.

3. RISK-UTILITY FRAMEWORK WITH THE AMERICAN COMMUNITY SURVEY DATA

In this section, we propose a R-U framework applicable to the microdata from the 2010 American Community Survey data, which incorporates risk measures based on record linkage using composite variables and a utility measure reflecting survey weights.

3.1 The American Community Survey

The American Community Survey (ACS) is an ongoing data collection by the U.S. Census Bureau, about characteristics of the nation's population and housing. Data are collected from approximately three million housing units per year, mainly by mail, but also with follow-ups by telephone or personal interviews (U.S. Census Bureau 2009). The data files and the questionnaire are available at the ACS website, *http://www.census.gov/acs/www/*.

The Census Bureau releases one-year, three-year and five-year estimates for states, cities, counties and large population groups in the U.S. It also provides Public Use Microdata Sample (PUMS) files, which contain 5% samples of records for individual housing units. Inconsistent or missing responses are imputed by the subjects' responses to other items or hot-deck imputation.

The PUMS files contain weights, so that users can estimate characteristics of interest. The Census Bureau provides estimates for selected characteristics in the ACS homepage to assist data users in checking that the weights are correctly used in parameter estimation. Following this approach, in our simulation study we estimate the population mean per element of l-th variable $\bar{Y}_l = \sum_{i=1}^{N} y_{il}/N$ by

$$\bar{y}_{\cdot l} = \frac{\sum_{i=1}^{n} w_i y_{il}}{\sum_{i=1}^{n} w_i} \tag{1}$$

where w_i denotes the weight associated with the i-th record in the sample. Note that the total population estimator $\hat{N} = \sum_{i=1}^{n} w_i$ in the ACS is controlled so not subject to sampling error.

For standard error estimation, there are two approaches introduced in analysis of ACS PUMS data, the *replication method* and the *design factor method* (U.S. Census

Bureau 2009, chap. 12). For the replication method, the Census Bureau provides $R = 80$ sets of *replicates weights*, constructed using *successive difference replication* (SDR, Fay and Train 1995) where the variance estimator is obtained based on the squared difference between two neighboring units. Let r denote the index of a replicate, and $w_{i,r}$ denote the weight of record i in replicate r. For each replicate r, we calculate the replicate estimate $\bar{y}_{\cdot l,r} = \sum_{i=1}^{n} w_{i,r} y_{il} / \hat{N}$. Then, the standard error of $\bar{y}_{\cdot l}$ is estimated by

$$SE_{\text{rep}}(\bar{y}_{\cdot l}) = \sqrt{\frac{4}{80} \sum_{r=1}^{80} (\bar{y}_{\cdot l} - \bar{y}_{\cdot l,r})^2}. \tag{2}$$

The number of replicate weights $R = 80$ is equivalent to the order of a Hadamard matrix used to produce replicate factors, and Equation (2) is based on the replicate variance estimate in Fay and Train (1995). Note that the sampling fraction $f = n/N$ is dropped from the original expression of Fay and Train (1995). To produce the 'final' replicate weights $w_{i,r}$, all the weighing processes given to the survey weight w_i, such as the population control or raking, are also applied to the replicate base weight generated from the SDR implementation.

The second approach in ACS PUMS variation estimation, called as the design factor method, is to use a generalized variance formula which contains the design factor DF, given by

$$SE_{\text{des}}(\bar{y}_{\cdot l}) = DF \times \sqrt{99 \, \frac{s_l^2}{\sum_{i=1}^{n} w_i}}. \tag{3}$$

where

$$s_l^2 = \frac{\sum_{i=1}^{n} w_i y_{il}^2 - \left(\sum_{i=1}^{n} w_i y_{il}\right)^2 / \sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i - 1}.$$

The DF is calculated as the ratio of the variance estimator using the repletion method to the hypothetical variance calculated under the assumption of SRS, to adjust the increase in variance due to the actual sample design (U.S. Census Bureau 2009, chap. 12). The

sampling rate for the PUMS file is approximately $f = 1$ percent, and the value 99 in the equation represents the approximated value of the 1-year PUMS finite population correction factor calculated by $(100 - f)/f$.

3.2 Exploratory Analyses With the 2010 ACS PUMS Data

This section presents an exploratory analysis with the 2010 ACS PUMS data for the state of North Carolina. The data consist of 95,531 persons from 41,674 housing units, which represent the NC population with a size of 9,561,558. For our study, we choose five income variables, wage or salary income (`WAGP`), self-employment income (`SEMP`), interest, dividend, and net rental income (`INTP`), social security income (`SSP`), and a composite variable given by a person's total income subtracted by the sum of four income variables above (`OTHER`). The variables are all continuous, but some records with point masses at zero. `SEMP` and `INTP` are top and bottom-coded, and `WAGP` and `SSP` are top-coded only. We adjust the dollar variables by applying the inflation adjustment factor whose value is 1.007624 for all sample cases for 2010 ACS (U.S. Census Bureau 2011).

In the ACS PUMS file, all income fields of persons under 15 years old are marked as NA. Also, all persons under 16 years old have zero incomes for `WAGP` and `SEMP`. For simplicity, we analyze a subset of the original data, which consists of $n = 76,450$ persons who are 16 years old or older from 41,649 housing units.

Table 1 shows the result of exploratory analyses which contains information about top/bottom codings, the number of zeros, the number of negative values, the unweighted average, the weighted average, the standard error from the replication method, and the standard error from the design factor method.

Here are some observations from the exploratory analyses. First, we observe that many income variables have zero values. Second, the weighted averages of the income variables are usually smaller than the unweighted averages. The unweighted averages

consider the contribution of all respondents equally, so some large values of people with small weights can be exaggerated, while the weighted averages adjust the relatively small impact of the large values on estimation. The histogram in Figure 3 shows how much the weights vary. Lastly, the table provides standard errors with the two approaches introduced in Section 3.1. It is shown that $SE_{\text{des}}(\bar{y}_{\cdot l})$ are generally larger than $SE_{\text{rep}}(\bar{y}_{\cdot l})$. Note that, although both variance estimators are not unbiased, the Census Bureau publishes $SE_{\text{rep}}(\bar{y}_{\cdot l})$ as the direct variance estimators, arguing that the variance estimator is "accurate enough for analysis of the ACS data" (U.S. Census Bureau 2009, chap. 12).

Table 1: Exploratory analysis results of a subset of 2010 ACS PUMS file of North Carolina, restricted to subjects who are 16 years old or older.

	WAGP	SEMP	INTP	SSP	OTHER
top/bottom coding	top	top/bottom	top/bottom	top	–
no. (%) of zeros	31,535 (41.2%)	72,141 (94.4%)	65,337 (85.5%)	57,925 (75.8%)	58,658 (76.7%)
no. (%) of neg.	0 (0.0%)	213 (0.3%)	215 (0.3%)	0 (0.0%)	0 (0.0%)
$\sum_i y_{il}/n$	22,383	1,369	1,605	2,932	3,377
$\bar{y}_{\cdot l} = \frac{\sum_i w_i y_{il}}{\sum_i w_i}$	21,814	1,262	1,304	2,455	2,930
$SE_{\text{rep}}(\bar{y}_{\cdot l})$	0.01	0.01	0.03	0.01	0.02
$SE_{\text{des}}(\bar{y}_{\cdot l})$	0.01	0.05	0.05	0.01	0.02

3.3 Risk Measure

What is an appropriate measure of disclosure risk varies with data and context. Characterizing intruder knowledge and behavior has long been recognized—and remains—a very challenging problem (Cox et al. 2011). Fellegi and Sunter (1969) study a

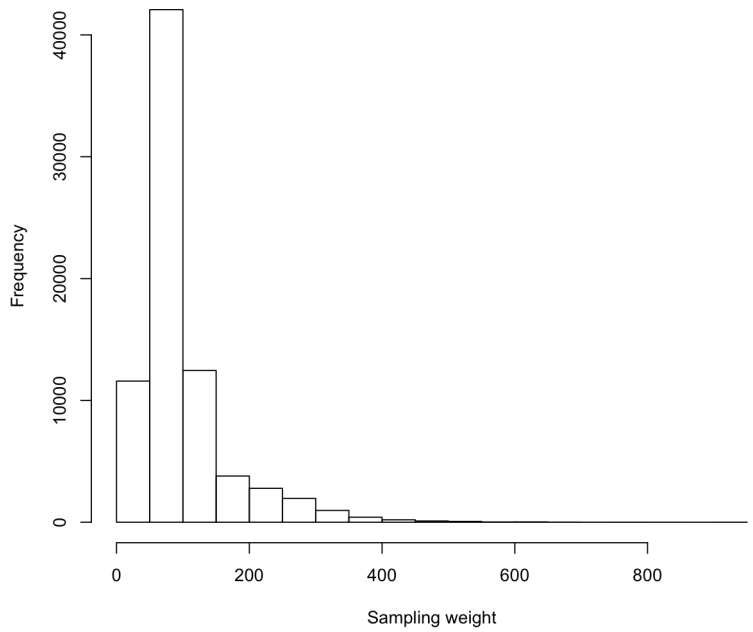

Figure 3: Distribution of weights of subjects in 2010 ACS PUMS file of North Carolina, who are 16 years old or older.

theoretical framework for record linkage based on likelihood ratios—specifically, the ratio of the conditional probabilities of *matches* and *non-matches* between a released record and a target record is used. The Fellegi-Sunter framework has been studied by many researchers, and was introduced to the area of SDL by Winkler (1998) and Lambert (1993).

While appealing, the Fellegi-Sunter framework can be difficult to implement. It requires (1) specification of a match score; (2) models for the distribution of the match score conditional on both matches and non-matches; and (3) specification of thresholds which classify the observed likelihood ratio as "match," "non-match," or requiring manual review. As an alternative when there is no sound way to resolve any of these questions, we adopt a *distance-based record linkage* techniques introduced in Domingo-Ferrer et al. (2001), based on the Euclidean (or more complicated, such as Hamming) distances between released records and target records.

In a departure from the existing literature, we assume that linkage is done using composite (sometimes also referred to as derived) variables rather than actual variables in D_{rel}. In our simulation study, we assume that the external datafile of intruder D_{ext} has true values of job income (`JOB`) which is the sum of `WAGP` and `SEMP` in the original agency's datafile D and true values of miscellaneous income (`MISC`) which is the sum of `INTP`, `SSP` and `OTHER` in D. Our choice of these particular composite variables is motivated by the fact that such information is often held by banks and credit rating agencies.

In our simulation study, we use a distance risk measure (see Domingo-Ferrer et al. 2001) with the composite variables produced by the following procedure:

1. Calculate the distance between the target record $\boldsymbol{y}_j^A$ in D_{ext} and the released value $\tilde{\boldsymbol{y}}_i^A$ in D_{rel} by

$$d_{j,i} = \sqrt{\sum_{l \in \{\texttt{JOB},\texttt{MISC}\}} (y_{jl}^A - \tilde{y}_{il}^A)^2}, \qquad \forall\ i, j = 1, \ldots, n.$$

2. Find the *nearest unit* i_j and the *second nearest unit* i'_j for each target unit j.

3. If y_{i0} of the nearest unit correctly matches with y_{j0} of the target unit, we say that the target unit j is "linked." If $y_{i_j 0}$ or $y_{i'_j 0}$ matches with y_{j0}, we say the target unit j is "linked to the second nearest".

   Finally, there are two risk measures, PL is the percentage of target records correctly linked to the nearest unit, and PL2 is percentage of target records correctly linked to the second nearest units, i.e.,

$$\text{PL} = 100\ \frac{\sum_{j \in D_{\text{ext}}} I[\ y_{j0} = y_{i_j 0}\ ]}{\sum_{j \in D_{\text{ext}}} 1} \tag{4}$$

and

$$\text{PL2} = 100 \, \frac{\sum_{j \in D_{\text{ext}}} I[\ y_{j0} = y_{i_j 0} \text{ or } y_{j0} = y_{i'_j 0}\ ]}{\sum_{j \in D_{\text{ext}}} 1} \tag{5}$$

where $I[\cdot]$ is the indicator function which has the value one if the condition holds or zero otherwise.

3.4 Utility Measure

Selection of appropriate utility measures is challenging, in part, because utility is specific to data, analysis, and even analyst. Utility measures range from narrow and specific measures to broad, but blunt, measures, with little middle ground (Karr et al. 2006).

In this paper, we focus on quality of the mean estimator of each variable and, for illustration, propose the use of the total absolute deviation (TAD) as a utility measure,

$$\text{TAD} = \sum_{l \in L} |\bar{y}_{\cdot l}^{\text{orig}} - \bar{y}_{\cdot l}^{\text{rel}}| \tag{6}$$

where $\bar{y}_{\cdot l}^{\text{orig}}$ denotes the estimate of population mean of variable l of the form (1) with D_{orig}, $\bar{y}_{\cdot l}^{\text{rel}}$ denotes the method-specific estimate with D_{rel} which is processed by a SDL method, and $L = \{\texttt{WAGP}, \texttt{SEMP}, \texttt{INTP}, \texttt{SSP}, \texttt{OTHER}\}$.

4. SIMULATION STUDY: ACS DATA

This section shows the simulation study and its results to illustrate SDL frameworks for finite population data using a subset of the 2010 ACS PUMS data introduced in Section 3.

4.1 Statistical Disclosure Limitation Methods

We apply four SDL methods introduced in Section 2.2 to the 2010 ACS PUMS data: adding noise (`Noise`), rank swapping (`Rank`), microaggregation (`Mic`), and microaggregation with adding noise (`MicN`). In the simulation study, we set different degrees of masking with `Noise` and `Rank` by setting various values of c^{noise} and c^{rank}. For `Mic` and `MicN`, we employ principal components projection as a multivariate ranking method (Anwar 1993; Defays and Nanopoulos 1993) to make groups, each with the size of $c^{\text{mic}} = 3$. For this, we do principal component analysis with D_{orig}, and then cluster records into groups by similarity of the first principal components. The size of the last cluster can be more than three, but less than six.

"Traditional" microaggregation replaces the original attribute values by the simple (unweighted) average of records in each micro-aggregated group. Specifically, suppose that subject i is a member of group $\mathcal{G}_i$ that is determined by a projection method. Then, the masked value for subject i under traditional microaggregation is given by

$$\tilde{y}_{old,i} = \frac{\sum_{i' \in \mathcal{G}_i} y_{i'}}{|\mathcal{G}_i|}.$$

To accommodate weights, we propose to use weighted averages instead, given by

$$\tilde{y}_{new,i} = \frac{\sum_{i' \in \mathcal{G}_i} w_{i'} y_{i'}}{\sum_{i' \in \mathcal{G}_i} w_{i'}}.$$

Table 2 is a simple example comparing the two approaches. Assume that we cluster the original values y_i by their sizes into the groups $\mathcal{G}_i$ of $|\mathcal{G}_i| = 3$. Then, we have three groups, as shown in the table. The illustrative example shows that the masked records by the traditional microaggregation $\tilde{y}_{old,i}$ (in the fourth column) preserve the unweighted mean estimate of y_i; yet, its weighted mean estimate differs from that of y_i. By contrast,

the released data with the modified version of microaggregation $\tilde{y}_{new,i}$ (in the last column) does not change the weighted mean estimate of $\{y_i\}$. The indexed microaggregation procedure discussed in Cox et al. (2011) is similar in concept to our proposed method.

Table 2: Comparison of a "traditional" microaggregation method and the proposed microaggregation method. w_i denotes the weight, y_i the original value, $\tilde{y}_{old,i}$ the micro aggregated value using the standard (unweighted) method, and $\tilde{y}_{new,i}$ the micro aggregated value using the proposed (weighted) method.

Group $\mathcal{G}_i$	w_i	y_i	$\tilde{y}_{old,i}$	$\tilde{y}_{new,i}$
1	1	1	2	2.5
1	3	2	2	2.5
1	6	3	2	2.5
2	1	4	5	5.5
2	2	5	5	5.5
2	5	6	5	5.5
3	1	7	8.5	9
3	2	8	8.5	9
3	3	9	8.5	9
3	4	10	8.5	9
$\sum y_i/n$		**5.5**	**5.5**	6.0
$\sum w_i y_i / \sum w_i$		**15.9**	14.5	**15.9**

4.2 Simulation Study for Variance Estimation

The first simulation study is to compare the mean estimates $\bar{y}_{\cdot l}$ and the two types of standard errors $SE_{\text{rep}}(\bar{y}_{\cdot l})$ and $SE_{\text{des}}(\bar{y}_{\cdot l})$ from masked datafiles D_{rel} with some SDL methods. We choose three SDL methods that have similar amount of disclosure protection: adding noise with $c^{\text{noise}} = 0.49$ (`Noise49`), rank swapping with $c^{\text{rank}} = 5$ (`Rank5`) and microaggregation with $c^{\text{mic}} = 3$ (`MicN`), whose values of PL are .40, .35 and .39 for `Noise49`, `Rank5` and `MicN`, respectively. We sample $M = 20$ replicates masked datasets $D_{\text{rel},m}$ from each method, where $m = 1, \ldots, M$.

Table 3 shows the averaged values of mean estimates and two standard errors over

Table 3: Point estimators and variance estimators of D_{orig} and D_{rel}. The selected SDL have similar amount of disclosure protection; PLs of (`Noise49`,`Rank5`,`MicN`) are (.040,.035,.039).

	$\hat{\theta}$				
	WAGP	SEMP	INTP	SSP	OTHER
$\mathbf{D}_{\text{orig}}$	**21,814**	**1,262**	**1,304**	**2,455**	**2,930**
Noise49	21,780	1,263	1,312	2,453	2,932
Rank5	21,884	1,290	1,325	2,462	2,955
MicN	21,825	1,259	1,305	2,455	2,932
	$CV_{\text{rep}} = SE_{\text{rep}}(\hat{\theta})/\hat{\theta}$				
	WAGP	SEMP	INTP	SSP	OTHER
$\mathbf{D}_{\text{orig}}$	**.013**	**.004**	**.028**	**.005**	**.018**
Noise49	.059	.301	.373	.077	.142
Rank5	.070	.456	.338	.057	.171
MicN	.025	.228	.358	.048	.096
	$CV_{\text{des}} = SE_{\text{des}}(\hat{\theta})/\hat{\theta}$				
	WAGP	SEMP	INTP	SSP	OTHER
$\mathbf{D}_{\text{orig}}$	**.010**	**.049**	**.049**	**.013**	**.019**
Noise49	.012	.062	.061	.016	.024
Rank5	.010	.050	.049	.013	.019
MicN	.010	.051	.051	.013	.019

twenty replicate masked datasets $D_{\text{rel},m}$ for each SDL. From the table, we see that `MicN` and `Noise49` produce the mean estimates close to those of D_{orig} while `Rank5` results in biased estimates. Specifically, `Rank5` always overestimates the parameters. This happens because the high values associated with small weights are swapped with small values associated with high weight in the rank swapping method. It implies that rank swapping is more ineffective in parameter estimation of population parameters than it is in previous SDL literature, which does not consider sampling weights.

A more interesting observation from the table is that CV_{rep} estimated from D_{rel} are far from those estimated from D_{orig}. This is because the construction of the replicate weights is tied to the sampling design. As we change the original values with SDL methods, the replicate weights are no longer associated with the same values as the sampling weights are. As we see from mean estimate $\hat{\theta}$ in the first table, mismatch between the masked values and sampling weights impacts on point estimation but the biases from masking are not huge. However, the change of values combined with the replicate weights impacts much on the variance estimation, which results in the poor result of CV_{rep} through all three SDL methods.

Now, we raise a question regarding the statement that $SE_{\text{rep}}(\bar{y}_{\cdot l})$ generally "produce more accurate estimates of a standard error" than $SE_{\text{des}}(\bar{y}_{\cdot l})$ (U.S. Census Bureau 2011). This statement seems valid when the original values are associated with the replicate weights. However, if the reported values are changed by SDL, the impact is unpredictable and may be significant, as the simulation study shows. In Table 3, CV_{des} of D_{rel} are very close to those of D_{orig}. Although it is originally assumed that CV_{rep} of D are good estimators of the true coefficients of variation, following SDL, CV_{des} of D_{rel} are relatively close to CV_{rep} of D compared to CV_{rep} of D_{rel}.

One possible explanation is that the ACS replication methodology is based on a specific sort order for the data, together with a selection of 780 row pairs from an 80×80

Hadamard matrix (U.S. Census Bureau 2009). When the data are subjected to SDL, this relationship is broken, which may result in the problems noted above.

4.3 Illustrative Example of R-U Framework for a Finite Population Estimation

The second simulation study illustrates the R-U framework applied to a finite population survey. We apply four SDL methods with different degrees of protection to the ACS data. For adding noise, we use the different values of $c^{\text{noise}} = (0.36, 0.49, 0.64)$. For rank swapping, we set $c^{\text{rank}} = (3, 5, 7)$. We sample $M = 20$ replicate masked datasets $D_{\text{rel},m}$ from `Noise`, `Rank` and `MicN`, while `Mic` has only one realization due to its non-random character.

Table 4: Utility and risk measures of D_{rel} from different SDL methods. Given the similar degree of disclosure protection, D_{rel} with `MicN` has the largest utility in terms of the total absolute deviation (TAD) of the form in (6). The standard errors of TAD range from 6 to 14. Those of PL and PL2 range from .001 to .003.

	`Noise36`	`Noise49`	`Noise64`	`Rank3`	`Rank5`	`Rank7`	`Mic`	`MicN`
Inverse-utility								
TAD	139	162	185	116	158	235	0	85
Risk								
PL	.05	.04	.03	.08	.04	.02	1.75	.04
PL2	.09	.07	.06	.15	.07	.05	3.42	.08

To compare the SDL methods, we calculate the risk measures and utility measure introduced in Sections 3.3 and 3.4. In terms of utility only, `Mic` has the smallest value, actually zero, of TAD as it exactly preserves the weighted mean of the original data as shown in Table 2. However, the method does not work well in terms of disclosure protection because it has the large values of PL and PL2. Adding noise with $c^{\text{noise}} = 0.64$ (`Noise64`) and rank swapping with $c^{\text{swap}} = 7$ (`Rank7`) produces released data with the

maximal level of data protection, but relatively low utility compared to the other methods. `MicN` seems promising since it shows the minimum level of TAD compared to the other methods with similar levels of PL and PL2, i.e., `Noise49` and `Rank5`.

Figure 4 shows the inverse-utility measure (TAD) and a risk measure (PL) for different SDL realizations. The plot suggests that `MicN` are generally superior to other approaches. Especially, most realizations of `MicN` are located bottom-left corner, which represent the largest utility with a moderate level of disclosure risk. This result is consistent with Oganian and Karr (2006), in which microaggregation with adding noise provided high-quality data while keeping disclosure risk low.

Agency must choose one of six realizations on the frontier – `Mic` lies on the frontier but is not shown here due to the huge amount of its disclosure risk. Choice of the final release depends on the agency's indifference curve. For an agency caring only about risk, the top-left value on the frontier representing a release from `Rank7` will be selected. At the other extreme, if only utility is under consideration, then the release from `Mic` (not shown in the figure) would be chosen. For R-U value tradeoff exhibiting decreasing marginal returns, i.e., convex curves of constant value as in Figure 1, the agency will choose among the other four releases on the frontier from `Rank7`, `Noise65`, `Noise49` and `MicN`.

We note briefly some additional analyses whose results are not shown in this paper. First, we implemented `Mic` and `Micp` with the traditional (unweighted) microaggregation instead of the weighted microaggregation. Similar to the result in Table 2, the data releases based on the traditional microaggregation method have large bias, and therefore low utility. Second, z-scores projection approach was used as an alternative grouping method for `Mic` and `Micp` instead of principal components projection. The results from the z-score approaches are similar to those from principal components projection approach shown in this paper. Finally, to assess generalizability of these results, we

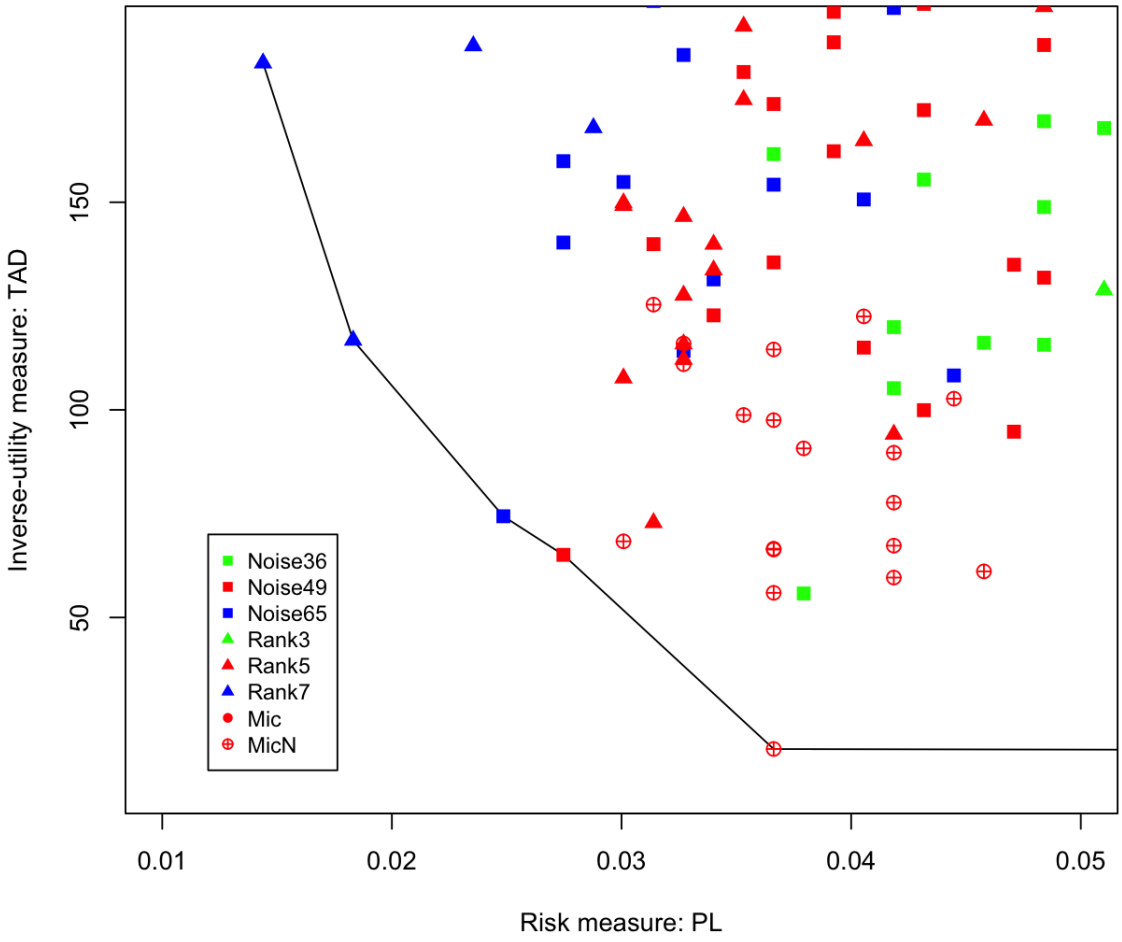

Figure 4: The risk-utility map with the replicate data sets from four SDL methods with different parameters. Note that Y-axis represents the inverse of utility that is opposite to that used in Figure 2. Therefore, the indifference value increases as a release approach to the origin.

repeated the analysis using 2010 ACS data from Pennsylvania, which yielded very similar results with those of North Carolina dataset.

5. CONCLUSIONS

In this paper we study the impact of statistical disclosure limitation on parameter estimation considering weights and articulate a framework for applying R-U paradigms for SDL in the setting of a finite population. The finite population setting is ubiquitous in analysis of survey data, but the utility appropriate to the weighted estimator has not previously been investigated.

In the course of developing the R-U framework with survey weights, we introduce some innovations: a risk measure based on record linkage using composite variables and a modified version of microaggregation that accommodates weights. We illustrate the framework by means of an experiment using publicly released microdata from the 2010 ACS. The results confirm previous findings that microaggregation with adding noise is an effective SDL strategy.

From the simulation experiment, we show that masked data from some SDL methods result in point estimators of population mean that are close to those of original data. However, variance estimator may be more impacted by SDL. Especially, the agency-recommended variance estimation using replicate weights can be seriously biased when it is performed using a masked dataset. It is possible that first performing SDL and then constructing replicate weights may attenuate the problem, and this is a subject for future research.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation [SES–1131897] to Duke University and the National Institute of Statistical Sciences (NISS). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Lawrence Cox and Jerome Reiter for insightful comments and discussions.

REFERENCES

Anwar, N. (1993), "Micro-Aggregation—The Small Aggregates Method," internal report, Luxembourg: Eurostat.

Cox, L. H., Karr, A. F., and Kinney, S. K. (2011), "Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act," *International Statistical Review*, 79, 160–183.

De Waal, A. G., and Willenborg, L. C. R. J. (1997), "Statistical Disclosure Control and Sampling Weights," *Journal of Official Statistics*, 13, 417–434.

Defays, D., and Nanopoulos, P. (1993), "Panels of Enterprises and Confidentiality: The Small Aggregates Method," in *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pp. 195–204.

Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. (2001), "Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk," in *Pre-proceedings of ENK-NTTS*, pp. 807–826.

Duncan, G. T., and Stokes, S. L. (2004), "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding," *Chance*, 17, 16–20.

Fay, R. E., and Train, G. F. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," in *American Statistical Association Proceedings of the Government Statistics Section*, pp. 154–159.

Fayyoumi, E., and Oommen, B. J. (2010), "A Survey on Statistical Disclosure Control and Micro-Aggregation Techniques for Secure Statistical Databases," *Software: Practice and Experience*, 40, 1161–1188.

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

Fienberg, S. E. (2009), "The Relevance or Irrelevance of Weights for Confidentiality and Statistical Analyses," *Journal of Privacy and Confidentiality*, 1, 183–195.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005), "Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk–Utility Framework for Remote Access Analysis Servers," *Statistical Science*, 20, 163–177.

Karr, A. F., Lee, J., Sanil, A. P., Hernandez, J., Karimi, S., and Litwin, K. (2001), "Disseminating Information but Protecting Confidentiality," *IEEE Computer*, 34, 36–37.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006), "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality," *The American Statistician*, 60, 224–232.

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," in *American Statistical Association Proceedings of the Survey Research Method Section*, pp. 303–308.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*, 9, 313–331.

Moore, R. A. (1996), "Controlled Data-Swapping Techniques for Masking Use Microdata Sets," Statistical Research Report 96/04, US Bureau of the Census, Statistical Research Division. Available at *http://www.census.gov/srd/www/byyear.html*.

Oganian, A., and Karr, A. F. (2006), "Combinations of SDC Methods for Microdata Protection," in *Privacy in Statistical Databases 2006, Lecture Notes in Computer Science*, eds. J. Domingo-Ferrer and L. Franconi, Berlin: Springer, pp. 102–113.

Reiter, J. P. (2005), "Estimating Risks of Identification Disclosure in Microdata," *Journal of the American Statistical Association*, 100, 1103–1112.

Sullivan, G., and Fuller, W. A. (1990), "The Use of Measurement Error to Avoid Disclosure," in *American Statistical Association Proceedings of the Survey Research Method Section*, pp. 802–807.

Tendick, P. (1991), "Optimal Noise Addition for Preserving Confidentiality in Multivariate Data," *Journal of Statistical Planning and Inference*, 27, 341–353.

U.S. Census Bureau (2009), "Design and Methodology: American Community Survey," U.S. Government Printing Office, Washington, DC. Available at *http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf*.

——— (2011), "PUMS Accuracy of the Data (2010)," electronic document. Available at *http://www.census.gov/acs/www/data_documentation/pums_documentation/*.

Willenborg, L. C. R. J., and De Waal, T. (1996), *Statistical Disclosure Control in Practice*, Springer–Verlag.

——— (2001), *Elements of Statistical Disclosure Control*, NY: Springer–Verlag.

Winkler, W. E. (1998), "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," *Research in Official Statistics*, 1, 87–014.

Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009), "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation," *Journal of Privacy and Confidentiality*, 1, 111–124.