The Journal of Privacy and Confidentiality (2009) **1**, Number 1, pp. 111–124

Global Measures of Data Utility for Microdata Masked for Disclosure Limitation

Mi-Ja Woo*, Jerome P. Reiter†, Anna Oganian‡, and Alan F. Karr§

Abstract. When releasing microdata to the public, data disseminators typically alter the original data to protect the confidentiality of database subjects' identities and sensitive attributes. However, such alteration negatively impacts the utility (quality) of the released data. In this paper, we present quantitative measures of data utility for masked microdata, with the aim of improving disseminators' evaluations of competing masking strategies. The measures, which are global in that they reflect similarities between the entire distributions of the original and released data, utilize empirical distribution estimation, cluster analysis, and propensity scores. We evaluate the measures using both simulated and genuine data. The results suggest that measures based on propensity score methods are the most promising for general use.

Key words and phrases: Data confidentiality; data utility; propensity score; cluster; nonparametric distribution function.

1 Introduction

Statistical agencies and other data producers disseminate many forms of microdata, i.e., data on individual subjects (people, households, establishments, ...), to the public. These disseminators strive for releases that protect the confidentiality of subjects' identities and values of sensitive attributes. Many agencies meet this objective by altering—we use the term *masking*—the original data before release, for example, by aggregating categorical values, swapping data values for selected records, or adding noise to numerical values (22). These methods limit disclosure risk by reducing the information available to intruders attempting to identify individuals in the released data.

Disseminators also strive, however, to release data that yield high quality results to legitimate users of the data, for instance, analysts estimating statistical models. This objective competes with confidentiality protection, because reducing data quality in order to thwart identification also negatively impacts inferences. Disseminators must therefore balance confidentiality protection with data utility. This is effectively done first by quantifying disclosure risk and data utility, then selecting strategies that have high utility for acceptable risks (7), or by restricting attention to a risk-utility frontier (9).

In the broadest sense, the utility of a particular data release is the benefit to society of the released information. Benefits this general are nearly impossible to quantify and measure, because they depend on more than simply the released data. A narrower, more feasible approach is to characterize the quality of what can be learned from the masked data relative to what

*National Institute of Statistical Sciences, Research Triangle Park, NC, mailto:
†Duke University, Durham, NC
‡National Institute of Statistical Sciences, Research Triangle Park, NC
§National Institute of Statistical Sciences, Research Triangle Park, NC

can be learned from the original data. Such comparisons can be tailored to specific analyses or can be broadened to global differences in distributions. An example of analysis-specific measures is when the data disseminator designates a single regression model and computes the overlap in the confidence intervals for the regression coefficients estimated with the original and the masked data (9). One class of examples of global measures are functions of the differences between point estimates of the first and second moments (and possibly other summaries) based on the original and masked data (6; 12; 23). Another is statistical distances between the distributions of the original and masked data (4; 8).

Analysis-specific and global measures have different merits. The former can be closely tied to both analysts' inferences and the nature of the masking strategy. For example, it is straightforward to evaluate the effect of top-coding (e.g., releasing incomes above $100,000 only in a category of "$100,000 or more") on an estimated mean or percentile of a distribution, or the attenuation (toward zero) in the coefficients of a particular linear regression when adding random noise to its predictors. The price, of course, of such high attention to one analysis is inattention or even harm to other analyses.

Global measures, on the other hand, are broad yet blunt. They reflect large-scale features of the entire distribution of the released data, and may be disconnected from impacts on particular analyses. They can incorporate the nature of the masking, for example, utility as functions of the number of swaps when using data swapping or the additional variance when adding random noise to numerical data.

Many existing global measures have features that limit their value as general purpose measures. For instance, comparing first and second moments of continuous distributions does not reveal information about tails, nor does it handle nominal data. Some statistical distances between distributions are difficult to compute for high-dimensional, mixed data, especially when the population distribution is not known.

In this paper, we propose several new global measures of data utility. These are most appropriate for individual-level data, although they can be modified to assess utility at aggregated levels. The idea underpinning the measures is to characterize the extent to which it is possible to discriminate between the original and released data using common statistical techniques. Released data that are difficult to distinguish from the original data have relatively high utility. By contrast, when it is easy to discriminate, the released data have relatively low utility. To perform the discrimination, we utilize propensity scores, cluster analysis, and empirical distributions. We evaluate the measures using both synthesized and genuine data. We find that a measure based on propensity scores is the most promising for general use, because it reflects features of the entire distribution, works for mixed data, is computationally feasible to implement, and distinguishes different masking strategies.

The paper is organized as follows. In S2, we present four new measures of data utility. In S3, we use empirical studies to illustrate some features of the measures. In S4, we conclude with discussion about how to use global measures of data usefulness. We do not discuss model-specific usefulness measures or disclosure risks. For discussions of model-specific measures, see (9). For a review of measures of disclosure risks, see (17).

2 Global Data Utility Measures

The four global utility measures presented here capture differences in the distributions of the original and masked data. The first measure (S2.1) adapts propensity scores as a tool for evaluating differences in distributions of two sets of data. The second measure (S2.2) uses cluster analysis to determine whether records in the original and masked data have similar values. The third and fourth measures (S2.3) use Kolmogorov-Smirnov-type statistics to evaluate differences between the empirical distribution functions of the original and masked data.

None of the measures is specifically tied to the nature of the masking. This allows us to compute utility values on the same scale for any masking strategy, which facilitates comparisons of the data quality achieved by competing strategies applied on the same data set.

2.1 Propensity Score Measure

In the observational study literature, the propensity score is the probability of being assigned to treatment, given covariate values $\mathbf{x}$. Treatment assignment and covariates are conditionally independent given the propensity score (19). Thus, when two large groups have the same distributions of propensity scores, the groups should have similar distributions of covariates.

This theory suggests an approach for measuring data utility. First, we merge (by "stacking") the original and masked data sets, adding a variable T equal to one for all records from the masked data set and equal to zero for all records from the original data set. If variables have been dropped as part of the masking, they are also dropped in computation of propensity scores. Second, for each record in the original and masked data, we compute the probability of being in the *masked* data set—the propensity score. Third, we compare the distributions of the propensity scores in the original and masked data. When those distributions are similar, the distributions of the original and masked data are similar, so data utility should be relatively high.

Propensity scores can be estimated via a logistic regression of the "masked/original" variable T on functions of all variables $\mathbf{x}$ in the data set. The propensity scores are the predicted probabilities in this logistic regression (2). For example, suppose we fit the logistic regression in (1) in S3. Using the notation of this equation, we estimate the propensity scores p_i by substituting the estimates of the regression coefficients (the estimated βs, which are obtained by maximum likelihood estimation) into (1) and solving for p_i.

The similarity of the propensity scores for the masked and original observations can be assessed in numerous ways, for example, by comparisons of their percentiles in each group. A simple summary is to compute

$$U_p = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - c]^2\,, \tag{1}$$

where N is the total number of records in the merged data set, $\hat{p}_i$ is the estimated propensity score for unit i, and c equals the proportion of units with masked data in the merged data set. In many cases, the original and masked data sets would have the same size N_0, in which case, $N = 2N_0$ and $c = 1/2$. When the original and masked data have the same distribution, the propensity scores for all units should approximately equal c, so that U_p is near zero. At the other extreme, if $\hat{p}_i$ is nearly one for units i from the masked data and nearly 0 for units from the original data, then the two data sets are completely distinguishable and $U_p \sim 1/4$.

This measure is sensitive to the specification of the logistic regression used to estimate the propensity scores. For example, using an intercept only in the regression results in $\hat{p}_i = c$ for all i, regardless of the values in the masked data. The advice from the literature on propensity score estimation is useful in the data utility context as well: include all variables, with interactions and polynomial terms, considered important to be similar in the original and masked data.

2.2 Cluster Analysis Measure

Cluster analysis, a form of unsupervised machine learning, places records into groups whose members have similar values of selected variables. For a random partition of a data set into two groups of sizes N_a and N_b, we would expect that, on average, $N_a/(N_a + N_b)$ percent of the observations in each cluster belong to group a.

This observation motivates our second measure of data utility. Let subscripts O and M denote the original data and masked data, respectively. First, we again merge the original and masked data sets. Second, we perform a cluster analysis on the merged data with a fixed number of groups, G. Third, we calculate the following measure:

$$U_c = \frac{1}{G} \sum_{j=1}^{G} w_j \left[\frac{n_{jO}}{n_j} - c \right]^2, \tag{2}$$

where n_j is the number of observations in the j-th cluster, n_{jO} (n_{jM}) is the number of observations from the original (masked) data in the j-th cluster, w_j is the weight assigned to the j-th cluster, and $c = N_O/(N_O + N_M)$. The weights w_j can equal the approximate standard errors of the percentages in the clusters, or they can reflect the importance of particular clusters. Large values of U_c indicate disparities in the cluster memberships, which in turn suggest differences in the distributions of the original and masked data.

Many algorithms for clustering require pre-specifying the value of G. For measuring data utility, there is no obvious criterion for selecting G. We desire G to be large to detect local deviations in the distributions of the original and masked data, but require at least two records per cluster. One approach is to try several values of G on the original data set to examine sensitivity to the choice of G, and select the masking strategy that appears most often as the best choice, after considering disclosure risk as well.

In addition to issues of selecting G, this measure has other weaknesses. When two masking strategies yield the same value of U_c, it is not necessarily the case that the masked data sets they produce is equally useful. For example, the masked data points and original data points may be widely separated within clusters for one strategy and narrowly separated within clusters for another strategy, yet the percentages of cluster memberships could be the same. Additionally, the U_c does not account for the similarity of records that are close to each other but classified in different clusters.

2.3 Empirical CDF Measures

These measures assess the differences between the empirical distribution functions obtained from the original and masked data. Let S_X and S_Y be the empirical distributions obtained from the original data, X, and the masked data, Y, respectively. When X has dimension

$N_x \times d$, we have

$$S_X(x_1, .., x_d) = \frac{1}{N_x} \sum_{i=1}^{N_x} I(x_{i1} \leq x_1, \ldots, x_{id} \leq x_d) \tag{3}$$

where x_{ij} equals the value of the j-th variable for the i-th observation, and $I(\cdot)$ equals one when the condition inside the parentheses is true and equals zero otherwise. The $S_Y(y_1, \ldots, y_d)$ is defined similarly.

Let $Z = (O, M)$ be the merged data having dimension $N \times d(N_O + N_M) \times d$. We consider two measures of data utility constructed from the empirical distributions:

$$U_m = \max_{1 \leq i \leq N} |S_X(\mathbf{z}_i) - S_Y(\mathbf{z}_i)| \tag{4}$$

$$U_s = \frac{1}{N} \sum_{i=1}^{N} [S_X(\mathbf{z}_i) - S_Y(\mathbf{z}_i)]^2 . \tag{5}$$

The former is the maximum absolute difference, and the latter the average squared differences, between the empirical CDFs. To use this measure with nominal data, one has to transform the labels into a series of indicator variables. A drawback to these statistics is that they can have low power to detect differences in distributions (1).

3 Empirical Studies

In this section, we examine the performance of the global measures of S2 using empirical studies. The first study illustrates some features of the measures with synthesized data. It illuminates important issues for implementation of the measures. The second and third studies apply the measures on data from the U.S. Current Population Survey and the U.S. Public Elementary/Secondary School Universe Survey. They provide some empirical evidence with which to compare the measures.

3.1 Synthesized Data

Following (9) and (13), we use empirical studies having known distributions to illustrate features of the global measures, including the sensitivity of the measures to the statistical characteristics of the data. Specifically, we create eight versions of two-dimensional data, in which the distribution is symmetric or non-symmetric, the two variables are highly correlated or uncorrelated, and the sign of the correlation is positive or negative. Symmetric data are simulated from a bivariate t-distribution with two degrees of freedom. The non-symmetric data include one variable simulated from an F-distribution with ten degrees of freedom in the numerator and denominator, and another variable simulated from linear regressions on the first variable. The sample size in each scenario is $n = 10,000$.

To mask the original data, we use some of the methods investigated in (12). These include: (i) incorrectly simulated data, generated by a bivariate *normal* distribution whose parameters are the sample mean and covariance of the original data; (ii) microaggregation with three observations per group, where groups are determined using z-score projections; (iii) microaggregation with three observations per group followed by adding random noise, where the noise is generated from the bivariate normal distribution with mean equal to zero and variance equal

to the differences between the sample covariance matrix of the original and microaggregated data (see (13)); and, (iv) rank swapping independently for each variable, where each record's swapped value is randomly selected from the 1500 (15% of the data) values closest to the original value.

For the propensity score method, we consider two logistic regressions using the merged data, Z, and an indicator variable, T, for data set membership:

$$\text{Model I} : \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 z_{i,1} + \beta_2 z_{i,2} + \beta_3 z_{i,1} z_{i,2} + \beta_4 z_{i,1}^2 + \beta_5 z_{i,2}^2 \quad (6)$$

$$\text{Model II} : \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 z_{i,1} + \beta_2 z_{i,2} + \beta_3 z_{i,1} z_{i,2} + \beta_4 z_{i,1}^2 + \beta_5 z_{i,2}^2 + \beta_6 z_{i,1}^2 z_{i,2}^2 + \beta_7 z_{i,1}^3 + \beta_8 z_{i,2}^3, \quad (7)$$

where $p_i = P(T_i = 1|Z_i)$. This enables us to investigate the sensitivity of the propensity score measure to the model specification.

For U_c, we classify records into clusters with the average linkage method (11). This is the default method in the software package SAS. We allow G to vary according to $G \in \{100, 250, 500, 750, 1000\}$, corresponding to 1% to 10% of the records. We set the cluster weights equal to the number of observations in each cluster.

The results across the eight data sets are reasonably summarized by comparing only the symmetric and non-symmetric versions of the data with high positive correlations. The utility measures for these two scenarios are displayed in Tables 1 and 2. To aid in comparisons, U_p and U_s are multiplied by $N = 20000$ in the tables. Larger values indicate lower data utility.

The propensity score measures differ for the two logistic regression models. Model I, seemingly erroneously, identifies the incorrectly simulated data as the most useful in both the symmetric and non-symmetric cases, whereas Model II identifies the incorrectly simulated data as the least useful. Since Model I includes only terms up to the second moment, which are nearly perfectly matched in the original and masked data when simulating from the bivariate normal, it attaches high utility to the incorrectly simulated data. However, these data poorly match the original data on higher-order moments. This is revealed in Model II, which includes higher-order terms.

The clustering measure can be sensitive to the group size. In the non-symmetric data, using $G < 1000$ identifies the microaggregated data as more useful than the incorrectly simulated data, whereas using $G = 1000$ does the opposite, although the difference in the U_c values for the two procedures when $G = 1000$ is small compared to the differences in the U_c values across procedures. This result can be explained as follows. When using microaggregation, the points from the masked data belonging to the same microaggregation group typically are placed in the same cluster. These clusters may not contain many original data records when the cluster sizes are small, as is likely the case for large G, and the variation of the original data values within some microaggregated groups is large, as is the case in the non-symmetric data.

The two CDF measures order the masking methods identically. The U_s measure appears to differentiate the methods more clearly than does the U_m measure. This is due partly to the different scale of the measures, and also because U_s, which considers differences across all data points, summarizes more aspects of the distributions than does U_m, which looks only at the maximum.

Comparing across measures, we see that they can give different orderings of procedures.

		Masking method			
		Incor. Simul.	Micro	Micro + Noise	Swap
Propensity	U_p (Model I)	**0.000**	16.048	0.002	9.997
	U_p (Model II)	596.970	32.890	**1.060**	18.292
Cluster	$U_c(G = 100)$	11.145	0.692	1.202	**0.521**
	$U_c(G = 250)$	5.357	0.811	2.996	**0.414**
	$U_c(G = 500)$	3.052	0.719	1.755	**0.309**
	$U_c(G = 500)$	2.126	0.619	1.313	**0.274**
	$U_c(G = 1000)$	1.652	0.556	1.055	**0.253**
CDF	U_m	0.234	0.051	0.156	**0.009**
	U_s	282.491	12.194	89.209	**0.127**

Table 1: Values of data utility measures for symmetric data. Values are comparable within but not among rows. Within each row, the masking method producing the highest utility is indicated in **boldface**.

The propensity score method with Model II selects microaggregation plus noise as most useful, whereas the clustering and CDF measures select rank swapping as most useful. Rank swapping precisely preserves (unweighted) univariate distributions at the expense of attenuating correlations. This preservation is highly valued by the clustering and CDF measures. The propensity score measure with Model II, however, strives to ensure that the second and higher moments are preserved, which is not done by swapping.

As discussed in (13), microaggregation plus noise can be argued to be the "best" method: the microaggregation preserves higher-order characteristics of the data, and the addition of noise restores the variability removed by microaggregation. Only the propensity score method with Model II is consistent with this reasoning.

3.2 CURRENT POPULATION SURVEY DATA

In this section, we use a subset of data from the March 2000 Current Population Survey (CPS) to illustrate the utility measures. The data comprise 51,016 heads of households and five variables, including age, race (4 categories), marital status (seven categories), household property taxes, and household income. An expanded version of this data set was used in (17; 18).

To alter the data, we apply different combinations of the following masking techniques (3; 5; 10):

- Round ages to the nearest multiple of five;
- Swap randomly 10% or 30% of races;
- Swap randomly 10% or 30% of marital statuses;
- Add random noise to positive property tax values drawn from $N(0, 290^2)$, where 290^2 is 1% of the variance of the positive property tax values. When masked property tax values are negative, re-draw until obtaining positive values. Zero values are not masked.
- Microaggregate income with 20 records per group.

		Masking method			
		Incor. Simul.	Micro	Micro+Noise	Swap
Propensity	U_p (Model I)	**0.000**	314.734	0.005	35.294
	U_p (Model II)	511.651	398.395	**0.385**	54.939
Cluster	$U_c(G = 100)$	5.844	4.00	0.568	**0.607**
	$U_c(G = 250)$	2.554	2.268	0.395	**0.331**
	$U_c(G = 500)$	1.422	1.342	0.326	**0.258**
	$U_c(G = 750)$	1.013	0.995	0.305	**0.236**
	$U_c(G = 1000)$	0.812	0.847	0.289	**0.231**
CDF	U_m	0.136	0.064	0.027	**0.010**
	U_s	85.433	20.002	1.986	**0.083**

Table 2: Values of data utility measures for non-symmetric data. Values are comparable within but not among rows. Within each row, the masking method producing the highest utility is indicated in **boldface**.

We first consider releasing data where age is rounded but the other four variables are not masked. We then add either (i) a random swap of 10% of races, (ii) a random swap of 10% of marital statuses, or (iii) two random swaps of 10% of races and 10% of marital statuses. Building on this last data set, we make the fifth and sixth masked data sets by adding noise to property taxes and microaggregating incomes. We repeat the process used to make the fourth through sixth data sets using 30% swap rates instead of 10%. We refer to these datasets with the combinations of masked variables' names and the masking parameters. For example, we write *Age+Race30+Mar30* when the masked data are generated by rounding age, swapping 30% of races, and swapping 30% of marital statuses (property taxes and incomes are not masked). We use these strategies solely for illustration; we do not believe these optimize the tradeoffs between disclosure risk and data usefulness. For example, synthetic data approaches (14; 15; 16) may provide better tradeoffs, even though we do not analyze such approaches here.

To compute the propensity score utility, we use a generalized additive model (20; 21) that includes main effects and first-order interactions among all five variables. The continuous variables are modeled using smoothing splines, and the coefficients are estimated using the "gam" routine in the software package *R*. The results are similar when estimating propensity scores with a logistic regression that includes all terms in a third-order polynomial in all five variables. To compute the cluster utility, we considered $G = 1,250$ and $G = 2,500$, corresponding to roughly 2.5% and 5% of observations. For all measures, we transform nominal variables into sets of indicator variables.

Unlike the study in S3.1, here we can (partially) order the masking strategies *a priori* with respect to the "degree of distortion." Using the notation in Table 3, this partial ordering is shown in Figure 1. Here, an arrow from one method to another means that the latter distorts the data more, and the relationships are transitive. Thus, *Age* distorts the data the least. *Age+Race10* and *Age+Mar10* distort the data more than *Age* because two attributes are altered rather than one. There is no reason *a priori* to know whether either of *Age+Race10* or *Age+Mar10* distorts the data more than the other, although it is clear from Table 3 that distortion is higher for *Age+Mar10*. Similarly, *Age+Race10+Mar10* causes more distortion than either *Age+Race10* or *Age+Mar10*. Analogous interpretations hold for the other arrows

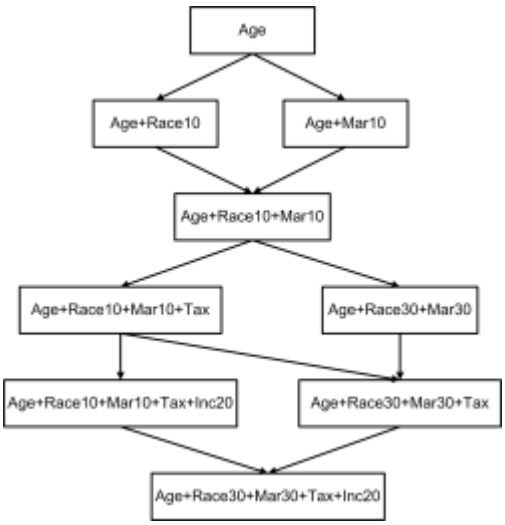

Figure 1: Partial ordering of the masking methods applied to the CPS data, in terms of data distortion. An arrow from Method A to Method B means that the latter distorts the data more.

in Figure 1.

The values of the data utility measures appear in Table 3. For simplicity, we multiply U_p and U_m by $N = 102,032$, and multiply each U_c by its G. All of the measures except clustering with $G = 1250$ order the strategies consistently with Figure 1. Clustering with $G = 1250$ assigns higher utility to *Age+Race30+Mar30+Tax+Inc20* than to *Age+Race30+Mar30+Tax*, which simply makes no sense. Unanimously, the other methods show that adding *Inc20* to either *Age+Race10+Mar10+Tax* or
Age+Race30+Mar30+Tax engenders negligible further distortion.

Pursuing this point, Table 2 summarizes the discrimination of the measures in terms of how they group the masking methods using small as compared to large differences in utility. For instance, every method except U_m assigns higher utility to *Age+Race10* than to *Age+Mar10*. One interpretation is that race has fewer categories than marital status, so that swapping it creates less distortion. At best, U_m fails to distinguish *Age+Race10* from *Age+Mar10*. Indeed, U_m seems to have effectively no capability to discriminate, and the capability of U_s, which does not distinguish *Age+Race30+Mar30* from *Age+Race30+Mar30+Tax* and *Age+Race30+Mar30+Tax+Inc20*, is less than that of U_p or either of the cluster measures. We emphasize, however, that further research is necessary in order to know definitively which differences of utility values are truly meaningful.

Uniquely among the utility measures, U_p assigns *much* higher utility to *Age+Race10* than to *Age+Mar10*, to the point that the utility of *Age+Race10+Mar10* is not much worse than that of *Age+Mar10*. Given the lack of discrimination for U_m and U_s, and the inconsistent results of U_c when $G = 1,250$, the U_p seems to be the preferred measure, as it was for the simulated data.

	Propensity	Cluster		CDF	
Masked data	U_p	$U_c(G = 1250)$	$U_c(G = 2500)$	U_m	U_s
Age	0.41	7.74	35.79	0.03	2.29
Age+Race10	2.78	33.26	70.16	0.39	7.96
Age+Mar10	46.41	102.78	168.07	0.35	12.55
Age+Race10+Mar10	50.12	127.70	210.82	0.39	18.22
Age+Race10+Mar10+Tax	208.79	388.14	545.06	0.39	20.48
Age+Race10+Mar10+Tax+Inc20	208.80	390.12	547.30	0.39	20.49
Age+Race30+Mar30	360.01	520.17	718.15	0.40	50.22
Age+Race30+Mar30+Tax	525.31	751.39	991.52	0.40	52.13
Age+Race30+Mar30+Tax+Inc20	525.44	741.25	1003.04	0.40	52.14

Table 3: Values of data utility measures with CPS data.

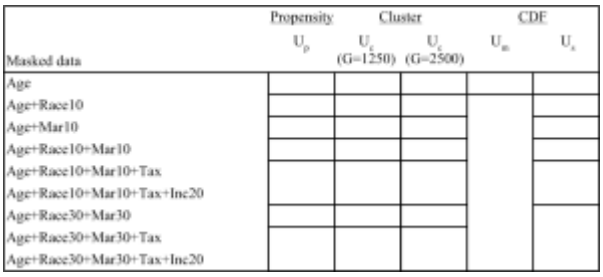

Figure 2: Similarity of masked CPS data sets as determined by utility measures. Masked versions with similar utility values are grouped in blocks.

3.3 SCHOOL DATA

Finally, we illustrate the utility measures using data from the 2003–2004 Public Elementary/Secondary School Universe Survey (PSUS), collected by the National Center for Education Statistics. The PSUS, collected annually, is the Department of Education's primary database on public elementary and secondary education. We use data on individuals from nine states, including locale (eight categories), total full time equivalent teachers, counts of free and reduced-price lunch eligible students, and counts of migrant students. For illustrations, we use only complete cases, which total 16,405 individuals [1].

To perturb this data, we mimic the masking strategies used previously for the CPS data.

- Swap randomly 10%, 20%, or 30% of state and locale indicators.
- Add random noise to full time equivalent teachers drawn from $N(0, 0.1\sigma^2)$, where σ^2 is the variance of the observed values of full time equivalent teachers. Negative values are not allowed.
- Micro-aggregate reduced lunch counts with 20 per group.
- Round migrant student counts to the nearest multiple of five.

We focus on evaluating the effects of selecting different swap rates for state and locale. That

[1] The details of this data set and the data files are displayed at http://nces.ed.gov/ccd/pubschuniv.asp.

Swap rates		Propensity	Cluster		CDF	
STATE	LOCALE	U_p	$U_c(G = 400)$	$U_c(G = 800)$	U_m	U_s
0%	0%	0.95	0.80	2.69	0.14349	0.02059
0%	10%	40.24	17.25	31.26	0.14379	0.02067
10%	0%	58.22	23.52	41.48	0.14373	0.02066
10%	10%	105.02	31.59	58.25	0.14398	0.02073
10%	20%	170.32	36.87	70.08	0.14459	0.02090
20%	10%	184.63	38.87	72.82	0.14404	0.02074
20%	20%	241.83	41.38	82.80	0.14434	0.02083
20%	30%	323.59	41.38	86.94	0.14440	0.02085
30%	20%	326.45	43.21	89.07	0.14471	0.02094
30%	30%	395.29	43.32	88.37	0.14465	0.02092

Table 4: Values of data utility measures with PSUS data.

is, we assume that the agency has implemented the noise addition, micro-aggregation, and rounding to protect the corresponding variables, and it seeks to evaluate the additional impact on utility of different swap rates. As before, these strategies are for illustrations and are not likely to be optimal for this dataset.

We compute the propensity score utility using a generalized additive model that includes main effects and interactions among all five variables. To compute the cluster utility, we select $G = 400$ and $G = 800$, again corresponding to 2.5% and 5% of observations. All nominal variables are split into a series of indicator variables.

The values of the data utilities are summarized in Table 4. The U_p measure orders the strategies appropriately, whereas the U_s and the U_m measures do not. In fact, neither the U_s nor U_m measures hardly discriminate among the masking strategies. The cluster measures lose discrimination capability for high swap rates. When going from 20% to 30% swaps, U_c increases only slightly when $G = 400$ and actually drops when $G = 800$. Once again, the results suggest that the propensity score utility outperforms the other measures.

For the propensity score measure, we also considered a logistic regression using a second-order polynomial in the five variables. We found that this model appropriately ordered the data sets and gave similar results as the generalized additive model.

4 Conclusions

When deciding on competing masking strategies, data disseminators need to assess the disclosure risk and data utility of each strategy. The global differences in the distributions of the original and masked data sets are an important aspect of the quality of the masked data.

In this paper, we have presented and evaluated four global measures of data utility. We believe that the empirical results with both synthesized and genuine data suggest that the propensity score method is particularly promising as a general use global utility measure. It behaved as expected for increasing intensity of data alteration, and it most clearly distinguished the qualities of the different masking strategies.

The key to implementing the propensity score approach is using a sufficiently detailed model. When there are many variables in the data set, this is a substantial modeling task, especially when the sample size is not large enough to handle a model with many terms. It may be possible to simplify the modeling task with minimal sacrifice in the quality of the measure. Variables that remain unmasked in both data sets can be included only through interactions with variables that have been masked, since the marginal distributions of unmasked variables are identical in the original and masked data sets. Variables deemed relatively unimportant for statistical analyses can be excluded from the propensity score models. Generalized additive models can simplify modeling relative to polynomials in logistic regressions.

In addition to using the measures to compare the merits of different masking strategies, data disseminators might want to interpret the values of the measures on an absolute scale. For example, they may want to determine if the utility of the proposed release is "good enough" for many users' purposes. Related to the issue of absolute scale is the interpretation of differences in the measures across masking methods. For example, the disseminator may want to know qualitatively how much utility is lost when a particular measure increases by some value Δ.

Such absolute statements are difficult to make with global utility measures, because the metrics are not the typical kind of statistical inference that users make with data. One approach is to derive a null distribution for each measure, for example by repeatedly resampling the observed data with replacement, then computing the measures using the resampled and original data sets. The disseminator can determine the fraction of times the candidate masked data value exceeds the values from the null distribution, as is done in classical randomization tests. We examined this approach for the propensity score measures using the CPS data and found that, except for data with age recodes only or small swap rates, the masked data values were far in the tails of the null distributions. This suggests that the measures are effective, in that they pick up deviations between the altered and observed data distributions. It also suggests that the types of masking used here have big impacts on the multivariate distributions.

Another approach recognizes that global data utility is only one component of data utility. Data disseminators also can compare inferences for specific models judged to be representative of the types of analyses done with the released data. When both global and specific measures are computed, data disseminators need to combine the results to determine a multivariate data utility measure. Developing methods for integrating multiple measures of data utility, as well as multiple measures of disclosure risk, is an area of future research.

References

[1] Baringhaus and Franz (2004). "On a new multivariate two-sample test." *Journal of Multivariate Analysis*, 88: 190–206. 115

[2] Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data.*. Chapman and Hall, 2 edition. 113

[3] Dalenius, T. and Reiss, S. P. (1982). "Data-swapping: A Technique for Disclosure Control." *Journal of Statistical Planning and Inference*, 6: 73–85. 117

[4] Dobra, Feinberg, Karr, and Sanil (2002). "Software systems for tabular data releases." *Int. J. Uncertainty, Fussiness and knowledge Based Systems*, 10(5): 529–544. 112

[5] Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). "Practical Data-oriented Microaggre-

gation for Statistical Disclosure Control." *IEEE Transactions on Knowledge and Data Engineering*, 14: 189–201. 117

[6] 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 of disclosure control methods." In *ETK–NTTS '2001*, volume 9, 807–825. Luxembourg: Eurostat. 112

[7] Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001). "Disclosure risk vs. data utility: The R-U confidentiality map." Technical report. Www.niss.org/downloadabletechreports.html. 111

[8] Gomatam, S., Karr, A. F., and Sanil, A. P. (2006). "Data swapping as a decision problem." *Journal of Official Statistics*, 21: 635–656. 112

[9] Karr, A. F., N., K. C., Oganian, A., P., R. J., and Sanil, A. P. (2006). "A framework for evaluating the utility of data altered to protect confidentiality." *The American Statistician*, 60: 224–232. 111, 112, 115

[10] Kim, J. J. (1986). "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation." 303–308. ASA. 117

[11] Murtagh, F. (1983). "A survey of recent advances in hierarchical clustering algorithms." *Computer Journal*, 26: 354–359. 116

[12] Oganian, A. (2003). "Security and Information Loss in Statistical Database Protection." Ph.D. thesis, University Politècnica de Catalunya. Http://vneumann.etse.urv.es/publications/theses/tesi.oganian.pdf. 112, 115

[13] Oganian, A. and Karr, A. F. (2006). "Combinations of SDC methods for microdata protection." In *Lecture Notes in Computer Science 4302*, volume 102–113. Springer-Verlag. 115, 116, 117

[14] Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). "Multiple imputation for statistical disclosure limitation." *Journal of Official Statistics*, 19: 1–16. 118

[15] Reiter, J. P. (2003). "Inference for partially synthetic, public use microdata sets." *Survey Methodology*, 29: 181–189. 118

[16] — (2004). "Simultaneous use of multiple imputation for missing data and disclosure limitation." *Survey Methodology*, 30: 235–242. 118

[17] — (2005). "Estimating risks of identification disclosure for microdata." *Journal of the American Statistical Association*, 100: 1103–1113. 112, 117

[18] — (2005). "Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study." *Journal of Royal Statistical Society, Series A*, 168: 185–205. 117

[19] Rosenbaum, P. R. and Rubin, D. B. (1983). "The Central Role of the propensity score in observational studies for Causal Effects." *Biometrika*, 70: 41–55. 113

[20] Wahba, G. (1990). *Spline Models for Observational Data.*. Philadelphia: SIAM. 118

[21] Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). "Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy." *Annals of Statistics*, 23: 1865–1895. 118

[22] Willenborg, L. C. R. J. and de Waal, T. (2001). *Elements of Statistical Disclosure Control.*. New York: Springer-Verlag. 111

[23] Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002). "Disclosure risk assessment in perturbative microdata protection." In Domingo-Ferrer, J. (ed.), *Inference Control in Statistical Databases*, 135–152. Berlin: Springer-Verlag. 112

Acknowledgements

This research was supported by NSF grant EIA–0131884 to 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.