RESEARCH REPORT SERIES
(*Statistics #2002-01)*

Disclosure Risk Assessment in Perturbative Microdata Protection

William E. Yancey, William E. Winkler, Robert H. Creecy

Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: January 31, 2002

Disclosure risk assessment in perturbative microdata protection*

William E. Yancey, William E. Winkler, and Robert H. Creecy

U.S. Bureau of the Census
{william.e.yancey,william.e.winkler,robert.h.creecy}@census.gov

Abstract. This paper describes methods for data perturbation that include rank swapping and additive noise. It also describes enhanced methods of re-identification using probabilistic record linkage. The empirical comparisons use variants of the framework for measuring information loss and re-identification risk that were introduced by Domingo-Ferrer and Mateo-Sanz.
Keywords: additive noise, mixtures, rank swapping, EM Algorithm, record linkage

1 Introduction

National Statistical Institutes (NSIs) have the need to provide public-use microdata that can be used for analyses that approximately reproduce analyses that could be performed on the non-public, original microdata. If microdata are analytically valid or have utility (see [18]), then re-identification of confidential information such as the names associated with some of the records may become easier.

This paper describes methods for masking microdata so that it is better protected against re-identification. The masking methods are rank swapping ([13], also [5]), additive noise ([8]), mixtures of additive noise ([14]). The re-identification methods are based on record linkage ([6], also [16]) with variants that are specially developed for re-identification experiments ([10], [18]). The overall framework is based on variants of methods that score both information loss and re-identification risk that were introduced by [5].

The outline of this paper is as follows. Section 2 covers the data files that were used in the experiments. In section 3, we describe the masking methods, the re-identification methods, and the variants of scoring metrics based on different types of information loss and re-identification risk. Section 4 provides results. In Section 5, we give discussion. The final section 6 consists of concluding remarks.

* This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

2 Data Files

Two data files were used.

2.1 Domingo-Ferrer and Mateo-Sanz

We used the same subset of American Housing Survey 1993 public-used data that was used by [5]. The Data Extraction System (http://www.census.gov/DES) was used to select 13 variables and 1080 records. No records having missing values or zeros were used.

2.2 Kim-Winkler

The original unmasked file of 59,315 records is obtained by matching IRS income data to a file of the 1991 March CPS data. The fields from the matched file originating in the IRS file are as follows:

1. Total income
2. Adjusted gross income
3. Wage and salary income
4. Taxable interest income
5. Dividend income
6. Rental income
7. Nontaxable interest income
8. Social security income
9. Return type
10. Number of child exemptions
11. Number of total exemptions
12. Aged exemption flag
13. Schedule D flag
14. Schedule E flag
15. Schedule C flag
16. Schedule F flag

The file also has match code and a variety of identifiers and data from the public-use CPS file. Because CPS quantitative data are already masked, we do not need to mask them. We do need to assure that the IRS quantitative data are sufficiently well masked so that they cannot easily be used in re-identifications either by themselves or when used with identifiers such as age, race, and sex that are not masked in the CPS file. Because the CPS file consists of a 1/1600 sample of the population, it is straightforward to minimize the chance of re-identification except in situations where a record may be a type of outlier in the population. For re-identification, we primarily need be concerned with higher income individuals or those with distinct characteristics that might be easily identified even when sampling rates are low.

3 Methods

The basic masking methods considered are (1) rank swapping, (2) univariate additive noise, and (3) mixtures of additive noise. Record linkage is the method of re-identification. The methods of information loss are those described by Domingo-Ferrer *et al.* [5]. and some variants.

3.1 Rank Swapping

Rank swapping was developed by Moore [13] and recently applied by Domingo-Ferrer *et al.* [5]. The data X is represented by (X_{ij}), $1 \leq i \leq n, 1 \leq j \leq k$, where i ranges through the number of records and j ranges through the number of variables. For each variable $j, 1 \leq j \leq k$, (X_{ij}) is sorted. For each j, (X_{ij}) can be swapped with (X_{il}) where $|j - l| < pn$ and p is a pre-specified proportion. The programming needed for implementing rank swapping is quite straightforward.

3.2 Additive Noise

Kim [8] introduced independent additive noise ε with covariance proportional to the original data X so that $Y = X + \varepsilon$ is the resultant masked data. The term ε has expected value 0. He showed that the covariance of Y is a multiple of the covariance of X and gave a transformation to another variable Z that is masked and has the same covariance as X. He also showed how regression coefficients could be computed and how estimates could be obtained on subdomains. His work has been extended by Fuller [7]. In this paper, we will consider the basic additive noise $Y = X + \varepsilon$ as was also considered by Fuller. Masking via additive noise has the key advantage that it can preserve means and covariances. Additive noise has the disadvantage that files may not be as confidential as with some of the other masking procedures. Kim [9] has shown that means and covariances from the original data can be reconstructed on all subdomains using the observed means and covariances from the masked data and a few additional parameters that the data provider must produce. Fuller [7] has shown that higher order moments such as the regression coefficients of interaction terms can be recovered provided that additional covariance information is available. In most situations, specialized software is needed for recovering estimates from the masked file that are very close to the estimates from the original, unmasked file.

3.3 Mixtures of Additive Noise

Roque [14] introduced a method of masking of the form $Y = X + \varepsilon$ where ε is a random vector with zero mean, covariance proportional to that of X, and whose probability distribution is a mixture of k normal distributions. The number k must exceed the dimension (number of variables) in the data X. The total covariance of ε is such that the $\mathrm{Cov}(Y) = (1 + d)\,\mathrm{Cov}(X)$ where $d, 0 < d < 1$, is pre-specified. The mean parameters of the component distributions are

solved by a nonlinear optimization method. With the empirical data used by Roque, the bias in the individual component means have the effect of making re-identification more difficult in contrast to re-identification when the simple normal noise method of Kim is used.

In this paper, we provide a simpler computational approach using factorization of $\mathrm{Cov}(X)$. The advantage is that no nonlinear optimization solver needs to be applied. In fact the basic computational methods are a straightforward variant of the methods used in [10]. The appendix gives more details.

Note that since additive noise preserves means μ and produces a scalar inflation of the covariance matrix, we can rescale the masked data records y from the masked data set Y by

$$y' = \frac{1}{\sqrt{1+d}} y - \left(1 - \frac{1}{\sqrt{1+d}}\right) \mu \tag{1}$$

so that the scaled data set Y' has expected value mean μ and $\mathrm{Cov}\,(Y') = \mathrm{Cov}\,(X)$.

3.4 Re-identification

A record linkage process attempts to classify pairs in a product space $A \times B$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Fellegi and Sunter [6] considered ratios R of probabilities of the form

$$R = \frac{\Pr(\gamma \in \Gamma | M)}{\Pr(\gamma \in \Gamma | U)} \tag{2}$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur or deal with different types of comparisons of quantitative data. The fields compared (surname, first name, age) are called matching variables. The numerator in (2) agrees with the probability given by equation (2.11) in [7] .

The decision rule is given by:

1. If $R > T_\mu$, then designate pair as a link.
2. If $T_\lambda \leq R \leq T_\mu$, then designate pair as a possible link and hold for clerical review.
3. If $R < T_\lambda$, then designate pair as a nonlink.

Fellegi and Sunter [6] showed that this decision rule is optimal in the sense that for any pair of fixed bounds on R, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds, T_μ and T_λ, are determined by the error bounds. We call the ratio R or any monotonically increasing transformation of it (typically a logarithm) a matching weight or

total agreement weight. Likely re-identifications, called matches, are given higher weights, and other pairs, called nonmatches, are given lower weights.

In practice, the numerator and denominator in (1) are not always easily estimated. The deviations of the estimated probabilities from the true probabilities can make applications of the decision rule suboptimal. Fellegi and Sunter [6] were the first to observe that

$$\Pr(\gamma \in \Gamma) = \Pr(\gamma \in \Gamma | M) \Pr(M) + \Pr(\gamma \in \Gamma | U) \Pr(U) \tag{3}$$

could be used in determining the numerator and denominator in (2) when the agreement pattern γ consists of simple agreements and disagreements of three variables and a conditional independence assumption is made. The left hand side is observed and the solution involves seven equations with seven unknowns. In general, we use the Expectation-Maximization (EM) algorithm [1] to estimate the probabilities on the right hand side of (3). To best separate the pairs into matches and nonmatches, our version of the EM algorithm for latent classes [16] determines the best set of matching parameters under certain model assumptions which are valid with the generated data and not seriously violated with the real data. In computing partial agreement probabilities for quantitative data, we make simple univariate adjustments to the matching weights such as are done in commercial record linkage software. When two quantitative items a and b do not agree exactly, we use a linear downward adjustment from the agreement matching weight to the disagreement weight according to a tolerance. For this analysis, we experimented with two methods of weight adjustment. For the raw data value a and the masked data value b, the d method adjustment is

$$w_{adj} = \max(\left\{ w_{adj} - \frac{(w_{agr} - w_{dis})|a - b|}{t \max(|a|, 0.1)} \right\}, w_{dis}),$$

and the l method adjustment is

$$w_{adj} = \max(\left\{ w_{adj} - \frac{(w_{agr} - w_{dis})|\log a - \log b|}{t \max(|\log a|, 0.1)} \right\}, w_{dis}),$$

where $w_{adj}, w_{agr}, w_{dis}$ are the adjusted weight, full agreement weight, and full disagreement weights, respectively and t is the proportional tolerance for the deviation ($0 \leq t \leq 1$). The full agreement weights w_{agr} and disagreement weights w_{dis} are the natural logarithms of (2) that are obtained via the EM algorithm. The approximation will not generally yield accurate match probabilities but works well in the matching decision rules as we show later in this paper. Because we do not accurately account for the probability distribution with the generated multivariate normal data, our probabilities will not necessarily perform as well as the true probabilities used by Fuller when we consider single pairs. We note that the d method is basically a natural linear interpolation formula based on the distance between the raw and masked values. The l method was originally devised by Kim and Winkler [11] for multiplicative noise masking, but it has proven generally more effective than the d method for the additive noise case. This is probably because the most identifiable records are the ones containing

large outlier values, and the l method tends to downweight less than the d method when both the input values are large.

To force 1-1 matching as an efficient global approach to matching the entire original data sets with the entire masked data sets, we apply an assignment algorithm due to [16]. Specifically, we use pairs $(i, j) \in I_0$ where I_0 is given by

$$I_0 = \min \left\{ \sum_{(i,j) \in I} w_{ij} \middle| I \subset J \right\},$$

where w_{ij} is the comparison weight for record pair (i, j), and J is the set of index sets I in which at most one column and at most one row are present. That is, if $(i, j) \in I$ and $(k, l) \in I$, then either $i \neq k$ or $j \neq l$. The algorithm of Winkler is similar to the classic algorithm of Burkard and Derigs (see *e.g.*, [16]) in that it uses Dijkstra's shortest augmenting path for many computations and has equivalent computational speed. It differs because it contains compression/decompression routines that can reduce storage requirements for the array of weights w_{ij} by a factor of 500 in some matching situations. When a few matching pairs in a set can be reasonably identified, many other pairs can be easily identified via the assignment algorithm. The assignment algorithm has the effect of drastically improving matching efficacy, particularly in re-identification experiments of the type given in this paper. For instance, if a moderate number of pairs associated with true re-identifications have probability greater than 0.5 when looked at in isolation, the assignment algorithm effectively sets their match probabilities to 1.0 because there are no other suitable records with which the truly matching record should be combined.

The proportion re-identified is the re-identification risk (PLD),computed using an updated version of the probabilistic re-identification software that has been used in [10], [14], and [5]. Domingo-Ferrer et al [5] also used distance-based record linkage (DLD) for situations in which Euclidean distance is used. DLD can be considered a variant of nearest-neighbor matching. Because this DLD is highly correlated with record linkage [5], we only use PLD. Domingo-Ferrer *et al.* also used interval disclosure (ID) that we do not believe is appropriate because it is far too weak a disclosure-risk. See [5] for a definition of ID.

For the two data sets examined below, we counted the number of re-identified matches used to compute the re-identification risk somewhat differently. For the Domingo-Ferrer data, since the data set is small, we counted all of the correct matches in the set of linked pairs reported by the record linkage software. For the larger Kim-Winkler data set, a more realistic count of correctly re-identified matches is given by counting matches with agreement weights above a cutoff value T_μ, where the rest of the correct matches in the file are scattered sparsely among a large number of incorrect match pairs. In practice, these sporadic matches would not be detectable and their inclusion would produce an intolerably high false match rate.

3.5 Information-Loss and Scoring Metric

In [5] a number of formulas are suggested for measuring the "information loss" or amount which the masked data set has been statistically altered from the original raw data set. The idea is to compute some kind of penalty score to indicate how much the masked data set statistically differs from the original. The problem becomes one of identifying what one considers to be significant statistical properties and then to define a way to compute their difference for two data sets.

For original data set X and masked data set Z, both $n \times m$ arrays, one might want to consider a measure of the change in data, *i.e.* a measurement for $Z - X$. The original suggestion was for something like

$$\frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{|x_{ij} - z_{ij}|}{|x_{ij}|}$$

but this is undefined whenever we have an original data element $x_{ij} = 0$. One can replace the denominator by a specified constant when $x_{ij} = 0$, but then the value of this score can vary greatly with the choice of constant, especially when the data set has a lot of zero elements, as in the Kim-Winkler data. Furthermore, the size of this data perturbation score tends to be several orders of magnitude larger than the other information loss scores described below, so that it totally dominates all of the other scores when they are combined. Initially we tried to improve this situation by modifying the above formula to

$$IL1 = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{|x_{ij} - z_{ij}|}{0.5\left(|x_{ij}| + |z_{ij}|\right)}$$

which helps with the score magnitude problem, since it is now bounded by $IL1 \leq 2$, but it only reduces but does not eliminate the possibility of dividing by zero. However, beyond these problems, by scaling the statistic by the individual data points, the resulting statistic is not very stable. Leaving aside zero data values, when the data values are very small, then small adjustments in the masked data produce large effects on the summary statistic. If we view our data set as independent samples from a common distribution, it is more stable to measure variations in the sample values by scaling them all by a value common to the variable. Thus if X and Z are independent random variables both with mean μ and variance σ^2, then the random variable $Y = X - Z$ has mean 0 and variance $2\sigma^2$. Hence a common scale for Z would be its standard deviation $\sqrt{2}\sigma$. In our case, we can estimate that standard deviation with the sample standard deviation S. This motivates the proposed modification for the data perturbation information loss statistic given by

$$IL1s = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{|x_{ij} - y_{ij}|}{\sqrt{2} S_j}.$$

This uses a common scale for all values of the same variable in the data set, the denominator is not zero unless the values of the variable are constant throughout the data set, and while the statistic does not have an *a priori* upper bound, the values in our empirical studies tend to be closer in magnitude to the other information loss statistics.

In summary we would suggest one of two approaches for a data perturbation score in the context of information loss measures for masked data sets. One approach would be to leave it out entirely. If the ideal of data masking is to try to preserve statistical properties of a data set while making individual records difficult to identify, when we are trying to assess the degree to which the statistical properties are preserved, perhaps we should not include a measure of how much individual records have been perturbed. The other approach, if one does want to include such a measure, then it would be better to use a more uniform and intrinsic scaling method, such as in $IL1s$.

The other information loss statistics that we compute are the same as some of those suggested by Domingo-Ferrer. To measure the variation in the sample means, we compute

$$IL2 = \frac{1}{m} \sum_{j=1}^{m} \frac{|\bar{x}_j - \bar{y}_j|}{|\bar{x}_j|}.$$

In theory, this score could also have the problem of zero or relatively small denominators, but since the sample means for our data sets are summary statistics for nonnegative whole real numbers, this did not seem to be a problem.

For variations in the sample covariance matrix, we compute

$$IL3 = \frac{2}{m\left(m+1\right)} \sum_{j=1}^{m} \sum_{k=1}^{j} \frac{\left|\mathrm{Cov}\left(X\right)_{jk} - \mathrm{Cov}\left(Y\right)_{jk}\right|}{\left|\mathrm{Cov}\left(X\right)_{jk}\right|}$$

for variations in the sample variances, we compute

$$IL4 = \frac{1}{m} \sum_{j=1}^{m} \frac{\left|\mathrm{Cov}\left(X\right)_{jj} - \mathrm{Cov}\left(Y\right)_{jj}\right|}{\left|\mathrm{Cov}\left(X\right)_{jj}\right|}$$

and for variations in the sample correlation matrix, we compute

$$IL5 = \frac{2}{m\left(m-1\right)} \sum_{j=1}^{m} \sum_{k=1}^{j-1} \left|\mathrm{Cor}\left(X\right)_{jk} - \mathrm{Cor}\left(Y\right)_{jk}\right|.$$

We wish to combine these information loss statistics into a summary information loss score. While it's not clear what sense it really makes to combine these numbers, and even if we do, its not clear what appropriate weighting we should give to them, in the absence of deeper insight, we just compute a straight average. However, we may choose which statistics we wish to include. As we have noted, the data perturbation measure $IL1$ is somewhat numerically problematic. Moreover, for the purposes of data masking, it is not clear if one cares

how much individual data records are perturbed as long as the overall statistical structure of the data set is preserved. Thus one information loss penalty score can be computed by leaving out $IL1$ to get

$$s0 = \frac{IL2 + IL3 + IL4 + IL5}{4}.$$

On the other hand, one can leave it in to get

$$s1 = \frac{IL1 + IL2 + IL3 + IL4 + IL5}{5}.$$

Another objection to combining all theses scores is that $IL3, IL4$, and $IL5$ are redundant. With the covariance, variance, and correlation, if we know two of these things, then we know the third. Furthermore, the covariance score $IL3$ is to a lesser degree subject to the same kind of scaling instability as found with $IL1$, namely that the smallest values make the largest contributions to the score. In particular, we observe that the score tends to be dominated by those components corresponding to the smallest correlations. Thus as an alternative summary information loss statistic, we suggest using the rescaled data perturbation score and leaving out the covariance score to get

$$s2 = \frac{IL1s + IL2 + IL4 + IL5}{4}.$$

For comparison purposes, for the empirical results, we combine each of these information loss scores with the re-identification score $reid$, as discussed in Section 3.4 to obtain an overall data masking score. Specifically, the resulting scores are given by

$$Ascore = 100\left(\frac{S0 + reid}{2}\right)$$
$$Dscore = 100\left(\frac{S1 + reid}{2}\right)$$
$$Sscore = 100\left(\frac{S2 + reid}{2}\right)$$

4 Results

4.1 Domingo Data Statistics

The Domingo data set consists of 1080 records from which we have masked 13 real variables. We can observe in Table 1 how the information loss scores increase with increasing noise level d. Rescaling the masked data has no mathematical effect on the mean and correlation. Since the rescaling somewhat contracts the data, there tends to be some decrease in the data perturbation scores. The effects of rescaling are most significant in the covariance and especially the variance scores.

Table 1. Domingo Data Information Loss Statistics

	IL1	IL1s	IL2	IL3	IL4	IL5	s0	s1	s2
rnkswp05	0.129	0.091	0.000	0.130	0.000	0.016	0.036	0.055	0.027
rnkswp10	0.219	0.155	0.000	0.195	0.000	0.036	0.058	0.090	0.048
rnkswp15	0.294	0.208	0.000	0.224	0.000	0.070	0.073	0.118	0.069
add01	0.194	0.137	0.001	0.036	0.014	0.002	0.013	0.038	0.039
add10	0.371	0.263	0.004	0.168	0.115	0.007	0.073	0.114	0.097
mixadd01	0.204	0.063	0.002	0.028	0.012	0.002	0.011	0.050	0.020
mixadd05	0.326	0.140	0.004	0.088	0.053	0.004	0.037	0.095	0.050
mixadd10	0.398	0.199	0.006	0.152	0.105	0.005	0.067	0.133	0.079
mixadd20	0.489	0.281	0.008	0.273	0.207	0.007	0.124	0.189	0.126
scalmixadd01	0.202	0.063	0.002	0.021	0.003	0.002	0.007	0.046	0.017
scalmixadd05	0.316	0.137	0.004	0.048	0.007	0.004	0.016	0.076	0.038
scalmixadd10	0.379	0.190	0.006	0.067	0.010	0.005	0.022	0.093	0.053
scalmixadd20	0.449	0.258	0.008	0.095	0.014	0.007	0.031	0.114	0.072

The matching software has two methods, the d method and the l method as discussed in Section 3.4 for measuring agreement between two real values. In either case, it interpolates between the agreement weight and the disagreement weight. In Table 2, we see that when the perturbations are small, the d method does a little better than the l method. However, when the perturbations get large, the l method is better able to see past moderate perturbations to large values.

The re-identification software produces a list of linked pairs in decreasing matching weight. For this small data set, the re-identification rate is computed as the total number of correctly linked pairs out of the total number of records in the data file. This is a rather optimistic re-identification score since most of the true matches are mixed among many false matches, and an analyst would probably have difficulty picking many of them out. In any event, we can see that for this data set with so few records and so many matching variables, a 1% noise level does not provide adequate masking, but the re-identification rate drops off rapidly with increasing noise level.

For an overall data masking score, we combine the information loss score with the re-identification score. Since we computed three data loss scores, we compute three overall scores in Table 3.

4.2 Kim-Winkler Data Statistics

The Kim-Winkler data consists of 59,315 records each containing 11 real variables for income data. In Table 4 we show the information loss statistics for our additive mixed noise masking for the whole data set. We note that the $IL1$ date perturbation statistic tends to higher than that for the Domingo data, possibly due to the large number of zero entries in the Kim-Winkler data, whereas the $IL1s$ data perturbation metric is about the same as for the Domingo data.

Table 2. Domingo Data Reidentification Rates

	d metric	l metric
rnkswp05	0.8861	0.9620
rnkswp10	0.2694	0.7287
rnkswp15	0.0491	0.3444
add05	0.7972	0.7500
add10	0.2296	0.3167
mixadd01	0.7667	0.7176
mixadd05	0.1482	0.3556
mixadd10	0.0574	0.2194
mixadd20	0.0139	0.1009
scalmixadd01	0.7704	0.7370
scalmixadd05	0.1602	0.3537
scalmixadd10	0.0648	0.2417
scalmixadd20	0.0269	0.1241

Table 3. Domingo Data Scoring Metrics

	d Metric			l Metric		
	Ascore	Dscore	Sscore	Ascore	Dscore	Sscore
rnkswp05	46.11	47.06	46.66	49.90	50.85	49.45
rnkswp10	16.37	17.97	15.87	39.34	40.94	38.84
rnkswp15	6.11	8.36	5.91	20.87	23.12	30.67
add01	40.51	41.76	41.81	38.15	39.40	39.45
add10	15.13	17.18	16.33	19.49	21.54	20.69
mixadd01	38.88	40.81	39.31	36.42	38.36	36.86
mixadd05	9.27	12.16	9.93	19.64	22.53	20.30
mixadd10	6.22	9.53	6.80	14.32	17.63	14.90
mixadd20	6.88	10.13	6.98	11.23	14.48	11.33
scalmixadd01	38.95	40.90	39.46	37.21	39.16	37.72
scalmixadd05	8.94	11.94	10.06	18.47	21.47	19.59
scalmixadd10	4.43	8.00	5.97	13.19	16.75	14.73
scalmixadd20	2.89	7.06	4.94	7.75	11.92	9.80

Table 4. Kim-Winkler Data Information Loss Statistics, 11 Variables

	IL1	IL1s	IL2	IL3	IL4	IL5	s0	s1	s2
mixadd01	1.165	0.060	0.002	0.014	0.010	0.000	0.007	0.238	0.018
mixadd05	1.308	0.135	0.005	0.056	0.051	0.001	0.028	0.284	0.048
mixadd10	1.381	0.191	0.007	0.106	0.101	0.001	0.054	0.319	0.075
mixadd20	1.457	0.270	0.010	0.201	0.201	0.002	0.104	0.374	0.121
scalmixadd01	1.163	0.060	0.002	0.008	0.001	0.000	0.003	0.235	0.016
scalmixadd05	1.302	0.132	0.005	0.018	0.002	0.001	0.006	0.265	0.035
scalmixadd10	1.370	0.183	0.007	0.025	0.002	0.001	0.009	0.281	0.048
scalmixadd20	1.438	0.249	0.010	0.033	0.003	0.002	0.012	0.297	0.066

Table 5. Kim-Winkler Data Information Loss Statistics, 8 Variables

	IL1	IL1s	IL2	IL3	IL4	IL5	s0	s1	s2
rnkswp05	0.174	0.123	0.000	0.525	0.000	0.197	0.180	0.179	0.080
rnkswp10	0.280	0.198	0.000	0.609	0.000	0.211	0.205	0.220	0.102
rnkswp15	0.362	0.256	0.000	0.605	0.000	0.214	0.205	0.236	0.118
add01	1.271	0.897	0.006	0.018	0.009	0.019	0.013	0.331	0.235
add01_sw	1.286	0.909	0.006	0.018	0.009	0.009	0.013	0.335	0.232
mixadd01	1.304	0.061	0.003	0.012	0.010	0.000	0.006	0.266	0.019
mixadd05	1.443	0.137	0.006	0.052	0.051	0.001	0.027	0.311	0.049
mixadd10	1.512	0.194	0.008	0.101	0.101	0.001	0.053	0.345	0.076
mixadd20	1.582	0.274	0.012	0.199	0.201	0.002	0.103	0.397	0.122
scalmixadd01	1.302	0.061	0.003	0.005	0.001	0.000	0.002	0.262	0.016
scalmixadd05	1.437	0.134	0.006	0.011	0.002	0.001	0.005	0.291	0.037
scalmixadd10	1.503	0.185	0.008	0.015	0.002	0.001	0.007	0.306	0.049
scalmixadd20	1.567	0.252	0.012	0.020	0.003	0.002	0.009	0.321	0.067

In general the scaled data tends to get better results reducing the covariance measures $IL3, IL4$ than in the Domingo data case.

For our re-identification, we only used eight of the income variables, so we computed the information loss scores based on just these eight variables. In Table 5, they show a generally slight increase over the eleven variable scores. For the re-identification scores, we computed the proportion of correctly linked pairs out of the total number of records, as in the case of the Domingo data. In this case, we only compute the results using the l interpolation metric in Table 6, since it is much more effective on this data. However, reporting the total number of correct matches in the full link file is probably even more misleadingly optimistic than in the Domingo data case. For the Domingo data, the true matches tend to be distributed throughout the link file. As the noise level of the masking increases, this distribution becomes more sparse and random. In the case of this data set, there are twenty or so records that are extreme outliers with one or more income categories much higher than the values for the mass of the records. Many of

Table 6. Kim-Winkler Data Reidentification Rates, l Metric

	Total File Matches	20% Zone
rnkswp05	0.8032	0.8032
rnkswp10	0.6072	0.6072
rnkswp15	0.4855	0.4855
add01	0.0590	0.0420
add01_sw	0.0010	0.0000
mixadd01	0.0841	0.0960
mixadd05	0.0346	0.0027
mixadd10	0.0240	0.0018
mixadd20	0.0149	0.0011
scalmixadd01	0.0844	0.0098
scalmixadd05	0.0355	0.0031
scalmixadd10	0.0249	0.0022
scalmixadd20	0.0174	0.0016

these records fail to be successfully masked from the re-identification through most noise levels, especially using the l metric. Thus there are always several clearly true matches at the top of the match-weight sorted link file. However, as the matching weights decrease, the proportion of true matches rapidly drops off as we include more and more false links in with a decreasing number of true matches. Thus is seems reasonable to cut off the count of true matches at some point, since beyond this point, any true matches will only appear sporadically among the preponderance of false matches and are unlikely to be discerned by the analyst. Here we choose a rather low cutoff point of 20%. This means that at this point, the number of linked pairs at this matching weight or higher contain 20% true matches and 80% false links. Below this point, the true matches become much rarer. In Table 7 are the overall data masking scores for the Kim-Winkler data. The data masking tends to be more effective here, especially at lower additive noise levels. Again inclusion of the $IL1$ data perturbation score tends to dominate and obscure the rest of the results.

Subpopulation Information Loss Statistics The additive noise procedures are supposed to preserve means and covariances on arbitrary subpopulations, at least when these statistics are properly corrected, according to the method of Kim [9]. In Tables 8 and 9 we compute the information loss scores for two subpopulations. We see that even for the corrected means and covariances, there are still generally somewhat higher scores than for the full data set. We especially note that the scaled data sets fail to recover the original data covariance and variance values as well.

In Table 9 we see slightly better information loss scores for a slightly larger subpopulation.

Table 7. Kim-Winker Data Scoring Metrics

	Full File Matches			20% Zone Matches		
	Ascore	Dscore	Sscore	Ascore	Dscore	Sscore
rnkswp05	98.35	98.21	88.32	98.35	98.21	88.32
rnkswp10	81.23	82.72	70.94	81.23	82.72	70.94
rnkswp15	69.03	72.17	60.81	69.03	72.17	60.81
add01	3.60	19.50	14.70	2.75	18.75	13.85
add01_sw	0.70	16.80	11.65	0.65	16.75	11.60
mixadd01	4.52	17.49	5.14	0.80	13.77	1.41
mixadd05	3.10	17.26	4.16	1.51	15.66	2.57
mixadd10	3.85	18.44	5.00	2.74	17.33	3.84
mixadd20	5.92	19.45	6.78	5.23	18.76	6.09
scalmixadd01	4.33	17.33	5.03	0.60	13.60	1.30
scalmixadd05	2.02	16.35	3.62	0.40	15.45	2.00
scalmixadd10	1.58	16.54	3.71	0.45	15.41	2.58
scalmixadd20	1.33	16.91	4.22	0.49	16.12	3.43

Table 8. S4 Return Type Information Loss, 8 Variables, 5885 Records

	IL1	IL1s	IL2	IL3	IL4	IL5	s0	s1	s2
rnkswp05	0.114	0.081	1.407	39.020	158.950	0.123	48.875	39.923	40.141
rnkswp10	0.199	0.141	0.397	1.682	3.980	0.179	1.560	1.287	1.174
rnkswp15	0.277	0.196	0.151	0.918	0.799	0.174	0.511	0.464	0.330
add01	1.364	0.964	0.027	0.343	0.240	0.017	0.156	0.398	0.312
add_sw01	1.364	0.964	0.027	0.343	0.240	0.017	0.156	0.398	0.312
mixadd01	1.399	0.246	0.057	0.101	0.008	0.006	0.043	0.314	0.079
mixadd05	1.559	0.550	0.128	0.232	0.024	0.014	0.099	0.391	0.179
mixadd10	1.631	0.777	0.181	0.342	0.040	0.021	0.146	0.443	0.255
mixadd20	1.697	1.099	0.256	0.539	0.069	0.033	0.224	0.519	0.364
scalmixadd01	1.397	0.245	0.057	0.130	0.008	0.006	0.050	0.320	0.079
scalmixadd05	1.552	0.536	0.128	0.298	0.024	0.014	0.116	0.403	0.175
scalmixadd10	1.621	0.741	0.181	0.440	0.040	0.021	0.170	0.460	0.246
scalmixadd20	1.681	1.004	0.256	0.693	0.069	0.033	0.263	0.547	0.341

Table 9. Schedule C Subset Information Loss, 8 Variables, 7819 Records

	IL1	IL1s	IL2	IL3	IL4	IL5	s0	s1	s2
rnkswp05	0.207	0.146	0.082	0.711	0.523	0.250	0.391	0.355	0.250
rnkswp10	0.322	0.228	0.125	0.796	0.573	0.261	0.438	0.415	0.297
rnkswp15	0.410	0.290	0.126	0.765	0.473	0.266	0.408	0.408	0.299
add01	1.221	0.863	0.013	0.021	0.017	0.003	0.013	0.255	0.224
add01_sw	1.250	0.884	0.057	0.431	0.296	0.084	0.217	0.424	0.330
mixadd01	1.410	0.220	0.012	0.065	0.009	0.005	0.023	0.288	0.062
mixadd05	1.560	0.492	0.027	0.175	0.026	0.012	0.060	0.360	0.139
mixadd10	1.627	0.696	0.038	0.299	0.043	0.019	0.100	0.405	0.199
mixadd20	1.688	0.984	0.054	0.531	0.075	0.030	0.172	0.476	0.286
scalmixadd01	1.408	0.219	0.012	0.084	0.009	0.005	0.028	0.304	0.061
scalmixadd05	1.553	0.480	0.027	0.225	0.026	0.012	0.073	0.369	0.136
scalmixadd10	1.617	0.664	0.038	0.385	0.043	0.019	0.121	0.420	0.191
scalmixadd20	1.674	0.899	0.054	0.682	0.075	0.030	0.210	0.503	0.265

5 Discussion

For the research community, there are two general difficulties with comparing masking methods. The first is that the suitable test files are needed. The test files should have variables in which the distributions are representative of actual databases. Some of the test files should have quite skewed distributions. Others should have a large number of zeros for several of the variables The second is that the information-loss metrics should be reasonably robust across different types of databases. We observed that some of the metrics that we have used in this paper are sensitive to the skewness of distributions and the proportions of zeros associated with a variable.

Much of prior research (e.g., [8], [7], [10]) dealt with situations where only a few specific analyses were demonstrated to be approximately reproduced with a masked data file. In most of the situations, special software was needed to do many of the analyses, particularly on subdomains. If masked files are required to reproduce more than one or two sets of analyses from original, unmasked data, then we suspect the special methods and software will be typically needed.

6 Concluding Remarks

This paper provides a comparison of rank swapping with various methods of additive noise. In the comparison of [5], rank swapping provided the best trade-off between information-loss and disclosure risk with measures used in the earlier work. With the same data and the same metrics, rank swapping provides better results than the types of mixtures of additive noise that we provide in this paper. With other, much larger data [10] that represents actual public-use situations, scaled mixtures of additive noise perform best with the same scoring metrics.

For the scoring methods used here, using scaled masked data produces improved scores since the information loss scores are improved by better covariance agreement while the re-identification risk is only slightly worse. This suggests that additional scoring metrics and more applications to different data situations are needed. The scoring metrics, particularly the components of information loss, need to be better connected to additional analyses and specific characteristics of data.

7 References

References

[1] Dempster, A. P., Laird, N. M. and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, **B, 39** (1977) 1–38.

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

[3] De Waal, A.G. and Willenborg, L.C.R. J.: A View on Statistical Disclosure Control of Microdata, *Survey Methodology*, **22**, (1996) 95–103.

[4] De Waal, A.G. and Willenborg, L.C.R. J.: Optimal Local Suppression in Microdata, *Journal of Official Statistics*, **14**, (1998) 421–435.

[5] Domingo-Ferrer, J., Mateo-Sanz, J. and Torra, Vincenc: Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk, *Proceedings of ETK-NTTS* '2001, (2001) to appear.

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

[7] Fuller, W. A.: Masking Procedures for Microdata Disclosure Limitation, *Journal of Official Statistics*, **9**, (1993) 383–406.

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

[9] Kim, J. J.: Subdomain Estimation for the Masked Data, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, (1990) 456–461.

[10] Kim, J. J. and Winkler, W. E.: Masking Microdata Files, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, (1995) 114–119.

[11] Kim, J. J. and Winkler, W. E.: Multiplicative Noise for Masking Continuous Data, *American Statistical Association Proceedings of Secure Survey Research Methods*, (to appear).

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

[13] Moore, R.: Controlled Data Swapping Techniques for Masking Public Use Microdata, *U.S. Bureau of the Census, Statistical Research Division Report 96/04 (1996).*

[14] Roque, G. M. , *Masking Microdata Files with Mixtures of Multivariate Normal Distributions*, Unpublished Ph.D. dissertation, Department of Statistics, University of California–Riverside (2000).

[15] Tendick, P. and N. Matloff, N.: A Modified Random Perturbation Method for Database Security, *ACM Transactions on Database Systems*, **19**, (1994) 47–63.

[16] Winkler, W. E.: Advanced Methods for Record Linkage, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, (1994) 467–472.

[17] Winkler, W. E.: Matching and Record Linkage, in B. G. Cox (ed.) *Business Survey Methods*, New York: J. Wiley, (1995) 355–384.

[18] Winkler, W. E.: Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, *Research in Official Statistics,* **1**, (1998) 87–104

8 Appendix: Additive Mixture Noise Methodology

As in the case of normal additive noise, to the given raw data set X, an $n \times m$ array, we wish to produce a masked data set Z by adding a masking noise array Y

$$Z = X + dY$$

where the records of Y are independent samples of a distribution with zero mean and $\mathrm{Cov}(Y) = \mathrm{Cov}(X) = \Sigma$. Typically for additive noise we chose a normal distribution $N(0, \Sigma)$; for mixture noise we may choose a normal mixture distribution $\sum_{k=1}^{K} \omega_k N(\theta_k, \Sigma_k)$. For a probability distribution, the weights are constrained so that $\sum_{k=1}^{K} \omega_k = 1, \omega_k > 0$. To obtain zero mean, we must have $\sum_{k=1}^{K} \omega_k \theta_k = 0$. When we choose $\Sigma_k = \sigma_k \Sigma$, for $K > m$, in general to obtain total covariance Σ, the component means θ_k must further satisfy a (underdetermined) system of quadratic equations that can be solved numerically, as addressed in [14]. However, it is computationally simpler to produce colored noise from white noise. That is, if $\Sigma^{\frac{1}{2}}$ is a square root of the positive definite symmetric matrix Σ,

$$\Sigma = \Sigma^{\frac{1}{2}} \left(\Sigma^{\frac{1}{2}}\right)^T$$

and w is an uncorrelated random vector with mean 0 and identity covariance I, then the random vector y,

$$y = \Sigma^{\frac{1}{2}} w$$

has mean 0 and covariance Σ. To produce w, we need m independent components w_j of mean 0 and variance 1. For standard normal additive noise, we can choose each w_j to be distributed as $w_j \sim N(0, 1)$; for mixture distribution noise, we may choose each w_j to be distributed as

$$w_j \sim \sum_{k=1}^{K} \omega_k N\left(\theta_k, \sigma^2\right)$$

for some choice weights and common variance $\sigma^2 < 1$. Such a mixture distribution has mean $\sum_{k=1}^{K} \omega_k \theta_k$ and variance $\sigma^2 + \sum_{k=1}^{K} \omega_k \theta_k^2$. To obtain component

means θ_k so that the mixture distribution has mean 0 and variance 1, we can start out with arbitrary numbers $\psi_1, \psi_2, \ldots, \psi_{K-1}$ and let

$$\psi_k = -\frac{1}{\omega_K} \sum_{k=1}^{K-1} \omega_k \psi_k$$

and compute

$$S = \sum_{k=1}^{K} \omega_k \psi_k^2$$

and let

$$\theta_k = \sqrt{\frac{1-\sigma^2}{S}} \psi_k$$

The simplest case occurs when we choose the weights ω_k to be equal and the smallest number of components $K = 2$. In this case we have $\theta_1 = \sqrt{1-\sigma^2}$ and $\theta_2 = -\sqrt{1-\sigma^2}$. The mixture distribution

$$\frac{1}{2} N\left(\sqrt{1-\sigma^2}, \sigma^2\right) + \frac{1}{2} N\left(-\sqrt{1-\sigma^2}, \sigma^2\right)$$

differs most from the standard normal distribution when we choose a value of σ^2 near 0, where we get a bimodal distribution with modes near ± 1. A weakness of using standard normal additive noise for masking is that most samples from the distribution tend to be near zero and hence produce small perturbation to the data. In this simplest mixture model, samples from the distribution tend to be near either 1 or -1, and thus should produce a more substantial data perturbation. For the data masking for these empirical studies, we used a value of $\sigma^2 = 0.025$.

Using this uncorrelated, zero mean mixture distribution, we generated a data set

$$W = \begin{pmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_n^T \end{pmatrix}$$

where each vector w_i is drawn from the above zero mean, identity covariance mixture distribution. Using a square root $\Sigma^{\frac{1}{2}}$ of the sample covariance matrix of X, we compute a colored noise data set

$$Y = \begin{pmatrix} \left(\Sigma^{\frac{1}{2}} w_1\right)^T \\ \left(\Sigma^{\frac{1}{2}} w_2\right)^T \\ \vdots \\ \left(\Sigma^{\frac{1}{2}} w_n\right)^T \end{pmatrix}$$

and for different noise proportion parameters d, we compute a masked data set

$$Z = X + dY\,.$$

Since the resulting data set Z theoretically has covariance $(1+d)\,\Sigma$, for each d we also compute a scaled masked data set Z_s related by

$$z_s = \frac{1}{\sqrt{1+d}} z + \left(1 - \frac{1}{\sqrt{1+d}}\right) \mu$$

where μ is the (sample) mean vector of the masked data.