RESEARCH REPORT SERIES
(Statistics #2007-21)

Examples of Easy-to-implement, Widely Used Methods of Masking for which Analytic Properties are not Justified

William E. Winkler

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

Report Issued: December 29, 2007

Examples of Easy-to-implement, Widely Used Methods of Masking for which Analytic Properties are not Justified

William E. Winkler[1], william.e.winkler@census.gov 071207

Abstract

This paper provides examples that illustrate the severe analytic distortions of many widely used masking methods that have been in use for a number of years. The masking methods are intended to reduce or eliminate re-identification risk in public-use files. Although the masking methods yield files that do not allow reproduction of the analytic properties of original, confidential files, in a number of situations they sometimes allow small amounts of re-identification using elementary methods and widely available software.

Keywords: Data Quality, Re-identification

1. Introduction

Microdata are much more useful for ad hoc analyses than tables in publications. Because of the need for public-use microdata, statistical agencies have often adopted methods to mask data to prevent or reduce the risk of re-identification in public-use files. Agencies have adopted a number of methods because of their ease of implementation without regard to whether the methods have been clearly justified in terms of preserving analytic properties in very particular situations with an individual data set or in general. Part of the difficulty arose because there were no clear cut methods and software for creating models or justifying non-trivial analytic properties when many of the methods were introduced in the 1960s and 1970s.

In this paper we demonstrate the sometimes severe degradation of a number of these widely used methods. These easy-to-implement methods include single-variable swapping, naively applied truncating, straightforward random sampling, rounding, rank swapping, single-variable microaggregation, and multi-variable microaggregation. After an extensive literature search and many informal communications, we have been unable to find *any* papers (with one very recent exception: Burkhauser, Feng, and Jenkins 2007) in which the analytic properties of possible or actual public-use files have been clearly justified (Winkler 2004, 2005). By *clearly justified*, we mean that the correspondence between certain aggregates in the masked, public-use data and aggregates in the original, confidential or the ability to support analyses such as regression or loglinear modeling is clearly shown. If there are limitations in the masked data, there are often no clear explanations of what few analyses are or are not possible.

A lone exception to the lack of analytic validity for the easy-to-implement methods is recent work introducing a new truncation methodology that allows a narrowly focused but important application. Burkhauser et al. (2007) provide convincing evidence that, if truncation is applied in a particular time-consistent manner, then it is possible to compute P90/P10 indexes and Gini indexes with data that are truncated with the new methodology

that are consistent with the comparable indexes that can be computed from the original, non-truncated data. These types of indexes are used in measuring income-inequality trends. As we will point out later, the new truncation methodology will not help in typical, single database applications where truncation is known to (sometimes severely) distort applications (e.g., Lane 2007, Figure 1).

The much more recent literature on modeling and synthetic-data generation (Fienberg 1997; Kennickell 1997; Abowd and Woodcock 2002, 2004; Reiter 2002, 2005; Raghunathan, Reiter, and Rubin 2003, Dandekar, Cohen and Kirkendal 2002) has much more clearly demonstrated some of the limitations of synthetic data to support more than a few analytic properties. With synthetic data, we must justify very specifically what analytic properties are preserved. Reiter has succinctly observed that specific properties that are not included in the model will not be included in the synthetic data. He has further observed that some properties may not be included in the model due to the simplifications needed for the modeling and the lack of data needed for accurate modeling. Other methods such as additive noise (Kim 1986; Fuller 1993; Yancey et al. 2002), Data Shuffling (Muralidhar and Sarathy 2006a,b), and Post Randomization (Gouweleeuw et al. 1993) have also justified a few analytic properties in masked files and the limitations of the masked files for general analyses.

The methods of creating models are not often adopted because of the perceived or actual difficulty in creating suitable models. The inability to model in a manner that preserves analytic characteristics and significantly reduces re-identification risk may change. Winkler (2007a, 2008) introduced general discrete-data modeling/edit/imputation methods for cleaning up data to improve quality. The methods can also be used to create synthetic data (Winkler 2007b) that preserves analytic properties while significantly reducing re-identification risk. They are no more difficult than standard methods of loglinear modeling (Bishop, Fienberg, and Holland 1975) and imputing for missing data under standard assumptions (Little and Rubin 2002). The methods are consistent with hot-deck assumptions that are widely used in surveys and can be extended to non-ignorable nonresponse in situations where models of the nonresponse mechanism are available. The new methods (Winkler 2007a, 2008) are much easier to implement than hot-deck, allow preservation of joint probabilities in a principled manner and preservation of a number of analytic characteristics (including variances). Typically hot-deck is naively applied without testing how collapsing can severely distort joint probabilities.

In this paper, we focus on methods for which there are no clearly demonstrated justifications that the methods preserve analytic properties (Winkler 2004, 2005) for a large number of real-world survey data situations. The methods include single-variable swapping, naïve truncating, sampling, rounding, rank swapping, single-variable microaggregation, and multi-variable microaggregation. The difficulties with single-variable swapping, truncating, and rounding have been noted by economists. Various advocates of methods such as microaggregation (Domingo-Ferrer and Mateo-Sanz 2002) have noted potential analytic problems but have not tried to establish a small number of situations in which the methods might be applied to preserve approximately one or two analytic properties.

We note that, if some of the easy-to-apply methods are applied with substantial care in a few specific situations, then the methods may yield partially valid analyses. We have

not been able to find clear justifications of situations in which an individual method, while not being general applicable, might be applied to data that is in a particular form that is amenable to the method. We caution that the methods cannot be routinely applied and expected to yield even partially valid analytic properties (as the examples will illustrate).

The outline of this paper is as follows. Following this introduction, we provide the examples with explanatory remarks. The artificial data represent general continuous data or general multivariate normal data. Fuller (1993), using methods that he originally introduced for errors-in-variables analyses, showed that that it is often sufficient to consider multivariate normal data with covariance matrix I or Σ. The advantage of such artificial data is that it is often easier to preserve some analytic properties than with actual real data. If the methods distort the masked artificial microdata, then they are even more likely to distort real microdata. In the third section, we introduce some elementary methods for re-identification that can be used for bounding worst-case situations. The advantages of the methods are that they are also easy-to-implement and are based primarily on the analytic properties on the masked files. We again use the artificial multivariate normal data to demonstrate re-identification as in Fuller (1993) or Winkler (1998). Because of the smoothness of multivariate normal data, it contains relatively fewer outliers than comparable real-world continuous data. If it is possible to re-identify with the artificial multivariate normal data, then it will be even easier to re-identify with real-world data. The fourth section provides discussion of more-difficult-to-implement alternative methods for which analytic properties of masked microdata have been demonstrated. In the final section we provide concluding remarks.

2. Examples with Explanation

In this section we provide examples of the deleterious effects of some widely used masking methods. We only deal with continuous data. Winkler (2007b) deals with a few situations of discrete data. As Fuller (1993) has shown, in many situations (such as additive noise or measurement-error problems), general continuous data can often be considered multivariate normal with mean zero and covariance matrix equal to the identity matrix **I**. This is because general multivariate data **Y0** can often be transformed to approximate normality **Y1** and the mean of the data can be subtracted. The data **Y1** can further be transformed to **Y2** = $\Sigma^{-1/2}\, Cov(Y1)$ where $\Sigma^{1/2}$ is the square root of *Cov(Y1)* that is obtained via the Cholesky or Singular Value Decompositions.

For ease of understanding, we generate multivariate normal data corresponding to **Y1** rather than **Y2**. Whereas **Y1** may allow easier preservation of analytic properties than with **Y0**, it is often more difficult to re-identify in **Y1** than in **Y0**. With real data **Y0**, some outliers will often stick out more from the overall point cloud than with approximately multivariate normal data **Y1**. The point of the artificial data is that it may be easier to preserve analytic properties in the masked version of the artificial data and re-identification risk may be lower than with real data.

Artificial Data

For most of the following examples, we use a 4-field data set that is generated using SAS. The data set consists of 1000 records of which only the first 100 are used in some

analyses. The first two fields, X1 and X2, are strongly pairwise correlated and the last two, X3 and X4, are somewhat pairwise correlated. The values of the fields are in the following ranges: $1 \le x1 \le 100$; $17 \le x2 \le 208$; $1 \le x3 \le 100$; $61 \le x4 \le 511$. The correlations are based on normal, homoscedastic error.

Table 1. Correlations of Original Data

	x1	x2	x3	x4
x1	1.000	0.895	0.186	-0.102
x2	0.895	1.000	0.218	-0.065
x3	0.186	0.218	1.000	0.329
x4	-0.102	-0.065	0.329	1.000

Random Blanking

The first method involves random blanking of values of a single field in any record. We randomly blank 5% and 25% of x1s, x2s and x3s with no simultaneous blanking. This method is sometimes referred to as *local suppression*. SAS computes correlations based on the available pairs.

The correlations in Table 2a,b are based on pairs of values of fields that are present in records. While blanking is very easy to implement, it causes moderate-to-severe distortions in most of the correlations. Some correlations are somewhat the same because of the simplistic manner in which blanking was done. Any other analytic properties of the files such as coefficients in regression analyses would be even more severely distorted. With actual survey, the blanking would need to be done in a manner that best reduces re-identification risk. The less simplistic blanking would likely increase the bias of simple statistics such as correlations and regression coefficients.

Table 2a. Correlations of Blanked Data (available number of pairs)
5% Blank Rate among x1, x2, and x3

	x1	x2	x3	x4
x1	1.000	0.884	0.219	-0.103
	96	92	90	96
x2	0.884	1.000	0.122	-0.091
	92	96	90	96
x3	0.219	0.175	1.000	0.302
	90	90	94	94
x4	-0.103	-0.091	0.302	1.000
	96	96	94	100

Table 2b. Correlations of Blanked Data (available number of pairs)
25% Blank Rate among x1, x2, and x3

	x1	x2	x3	x4
x1	1.000	0.875	0.271	-0.125
	80	54	55	80
x2	0.875	1.000	0.122	0.045
	54	74	49	74
x3	0.271	0.122	1.000	0.331
	55	49	75	75
x4	-0.125	0.045	0.331	1.000
	80	74	75	100

Truncation

With naïve truncation, all values of a variable *X* above a certain bound *B* are replaced with the upper bound *B*. More recent truncation methods often chose a value *B1* that separates the upper 0.5% or upper 3% tail of the distribution (that varies better with inflation from year to year) and uses the average *AV1* of the *X* values above the truncation point as the replacement value. The more recent method preserves totals and will have slightly lesser deleterious effect on simple statistics such as correlations than the simplest truncation method for which we provide empirical examples.

For this example, we use CPS data from approximately 1992 and a second data source that has been matched to the CPS. In the second source, we have modeled the upper tail of the distribution of values and replaced the true values with synthetic values that are drawn from the modeled distribution. The synthetic data in the tail of the second distribution will not alter the results of the effects of truncation. We use a subset of size ~26,000 with cpswage and wage above 5.

Table 3. Data Characteristics for 26418 Records

Variable	Mean	Std Dev	Minimum	Maximum
wage	36007	46100	11	3932720
cpswage	35144	29098	11	434999

Table 4. Distortions in Correlation Due to Truncation

Income Truncation Value	Correlations between Wage and CPSWage
None	0.623
400,000	0.768
200,000	0.823
100,000	0.862
50,000	0.883

If both values are above the truncation point T, then the original value is replaced by T. If T=400,000, then two wages above 400,000 would each be replaced by 400,000.

Table 5. Regressions with Different Truncation Values

Income Truncation Value	R^2	Wage and CPSWage	
		Intercept	Slope
None	0.388	1346	0.986
400,000	0.590	4420	0.885
200,000	0.678	4740	0.863
100,000	0.742	4307	0.863
50,000	0.780	3085	0.886

We see that truncation has very deleterious effects on the analytic properties. With this particular (mostly real) data, the characteristics of the data above the truncation value are substantially different than those above the truncation value. We note that this type of analytic degradation would likely be observed with any artificial data. It has also been noted by Lane (2007, Figure 1) with real data.

Although there are sophisticated methods for filling in the truncated values from a hypothesized distribution (Little and Rubin 2002, Feng et al. 2006), such filling-in methods are not necessarily easy to implement. If the agency producing the truncated data also gives better information about the (smoothed) distribution for the truncated values, then many users would be able to improve on the basic filling-in methods because they have better information. It is not clear how easily the filling in could be performed in multivariate situations where the filling-in involves pairs of variables that are not necessarily correlated as there are in the above example.

The reason that individuals use truncation is to reduce re-identification risk. If someone has a relatively high income (upper 0.5% tail of a distribution) and is in a database that contains other (quasi-)identifying information such as age (or age range), sex, race, and a geocode such as State or large metropolitan area, then that individual and other characteristics in the database associated with the individual may easily be re-identified. If a large number of incomes need to be truncated in a large number of subdomains, then analytic properties will necessarily be much more seriously affected than in the above empirical example.

The new truncation methodology of Burkhauser et al. (2006) allows the comparison of P90/P10 and Gini indexes over a number of time periods. Their truncation method is based on modeling the distributions over all the time-periods that are used in an analysis and choosing a common truncation point based on a fixed proportion of the records in each time period. Their method differs from the conventional truncation in which a truncation value is chosen and used for several years and then a new (typically) higher truncation value is chosen is used for several more years. Each of the truncated values is replaced by the mean of the values being truncated rather than the value at the truncation

point. For any given year (i.e. database), their method will not solve the problems affecting correlations and regressions as illustrated in the above example.

Sampling

Sampling can reduce re-identification risk when the records that are sampled lie in the interior of point clouds of continuous variables. If a record is an outlier in a population, then it will also be an outlier in the sample and much more likely to be re-identified. If a record is an outlier in an uncontrolled random sample, then it is quite likely to be an outlier in the population when the number of variables is greater than 10. By an *uncontrolled random sample*, we mean a sample that has no explicit controls that would, say, prevent the sampling of population outliers.

In this subsection, we describe a number of degradations that sampling induces. The degradations are well-known to many sampling experts but may not be widely known to others. Sampling experts can sometimes better assure analytic properties in files with methods that control various aggregates so that the sampled files preserve the aggregates (after estimation procedures are followed). Alternatively, the sampling may sometimes be repeated until some properties are assured in a particular sample. In the discussion of section 4, we specifically consider re-identification when certain of the easy-to-use masking methods are combined with sampling.

Because most public-use files have 20 or more variables, it is very difficult to get a representative sample (particularly at the 1% or 10% rates that are often used). By *representative* we mean that the sample will support many of the analytic properties of the entire population. If we have a skewed variable x and take a 1% sample, then almost no samples will contain the 10 largest values of x. If we take a 10% sample, then almost half of the samples will not contain the 10 largest values of x. This means that the effect of sampling with some continuous variables can be like the effect of truncation in one of the previous examples. The value 10 is chosen for convenience and may vary according to the data and the perceived analytic characteristics that need to be approximately preserved.

If we have discrete variables, then it seems likely that most 1% (and some 10%) samples will not contain all of the patterns that are present in the entire population and the analytic results will be compromised. By a *pattern*, we mean a particular set of values that occur for the set of variables. If a file is sufficiently large to allow 1% sampling, then the majority of patterns that occur 40 of fewer in the original population will not occur in any fixed sample. If these patterns or a large subset of them are needed to preserve certain analytic properties, then the analytic properties will almost certainly not be present in some of the samples.

If the producer of the public-use file has the situation of 20 or more variables, then the producer will have a very difficult time of justifying analytic properties with many of the 1% (and possibly many of the 10%) random samples. A potential solution is to keep drawing random samples, until a particular sample can be justified in terms of analytic properties. After the sample is justified, then the 20-variable records can be examined carefully to determine those that are potential outliers in different combinations of variables and may need additional masking to reduce re-identification risk. The additional masking, however, may severely distort analytic properties.

Rounding

Another easy-to-implement method is rounding for which we also use the SAS rounding procedure. The naïve intuition is that, while rounding makes it more difficult to re-identify, rounding does not seriously compromise analytic properties. We will comment on re-identification again when we consider more sophisticated generalizations of masking that include rounding as a special case.

Table 6a. Correlations after Rounding. Base 10.

	x1	x2	x3	x4
x1	1.000	0.887	0.170	-0.115
x2	0.881	1.000	0.205	-0.062
x3	0.170	0.205	1.000	0.315
x4	-0.115	-0.062	0.315	1.000

Table 6b. Correlations after Rounding. Base 50.

	x1	x2	x3	x4
x1	1.000	0.739	0.096	-0.173
x2	0.739	1.000	0.183	-0.107
x3	0.096	0.183	1.000	0.289
x4	-0.173	-0.107	0.289	1.000

Table 6c. Correlations after Rounding. Base 100.

	x1	x2	x3	x4
x1	1.000	0.688	0.141	0.025
x2	0.688	1.000	0.092	0.057
x3	0.141	0.092	1.000	0.320
x4	0.025	0.057	0.320	1.000

With Round 100, most variables only take 2, 3, or 5 different values. In the first situation (Table 6a), correlations are slightly distorted; in all remaining situations of rounding, correlations are severely distorted. For instance, in Table 6b, the correlation between x1 and x2 is 0.739 and between x1 and x3 is 0.096; in Table 1, the correlations are 0.895 and 0.186, respectively. Although slight rounding can preserve simple analytic properties in some situations, we also need to be concerned with re-identification. Although rounding to base 100 is very extreme with the empirical data of Table 1, it may be necessary to reduce re-identification risk. Very elementary re-identification methods are covered in the next section and in the discussion. Although we do not cover specific re-identification risk for rounding, we do cover re-identification risk for rank swapping and

single-variable microaggregation (both covered below) that are each known to generalize rounding.

Perturbation Methods

With *perturbation* methods we change values of certain variables in a more sophisticated manner to make re-identification more difficult. The intent is to possibly preserve one or two analytic properties. In the following we consider the easiest methods. The ability of additive noise, general perturbation (Muralidhar and Sarathy 2002, 2006a,b), blanking and imputation, and general models to create synthetic data have been justified elsewhere.

1. Swapping
2. Rank Swapping
3. Microaggregation

4. Additive Noise
5. General Perturbation
6. Blanking and Imputation (Partial Model)
7. Create General Model and Draw Synthetic Data

Swapping

With this procedure, called *single-variable swapping*, we swap values of individual variables independent of other variables. The ideas of single-variable swapping were originally developed by Dalenius and Reiss (1982). Fienberg and MacIntyre (2005) provide a recent overview.

In the following, the first three of the variables are swapped.

Table 7a. Correlations after swapping (10% rate)

	x1	x2	x3	x4
x1	1.000	0.010	0.100	0.104
x2	0.010	1.000	-0.039	0.039
x3	0.100	-0.039	1.000	0.290
x4	0.104	0.039	0.290	1.000

Table 7b. Correlations after swapping (100% rate)

	x1	x2	x3	x4
x1	1.000	-0.876	0.936	-0.942
x2	-0.876	1.000	-0.980	0.935
x3	0.936	-0.980	1.000	-0.956
x4	-0.942	0.935	-0.956	1.000

The 10% swapping rate means that 10% of the values of each variable are swapped arbitrarily with other values of the records. We observe that correlations are *almost instantly destroyed* with swapping even at the 10% swapping rate. For instance, in Table 7a, the correlation between x1 and x2 is 0.010 and the correlation between x1 and x3 is 0.100; in Table 1, the correlations are 0.881 and 0.170, respectively. Re-identification rate after 100% swapping is effectively 0. At 10% swapping, the re-identification rate will typically be substantially greater than 0. If values of two variables (among x1, x2, x3, and x4) are sufficient to re-identify (by comparing the original file directly with the swapped version of the file), then there is greater than 99% probability that two or less values in any record would be swapped. Among 1000 records, we would expect to re-identify at least 990 records.

An alternative to single-variable swapping is to swap several variables simultaneously across records. We will call this method *group swapping.* As an instance, if a record $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ik}, y_{i1}, y_{i2}, \ldots, y_{il})$ has its *y* –values swapped as group with the *y*-values in $\mathbf{X}_j=(x_{j1}, x_{j2}, \ldots, x_{jk}, y_{j1}, y_{j2}, \ldots, y_{jl})$. If the swapping is on the entire file or only within a set of subdomains that partition the entire file, then the means and correlations of the *y*-variables will be preserved. If the group swapping of the *y*-variables is within subdomains, then the means and correlations of *y*-variables will also be preserved on the subdomains. The correlations across *x* and *y* variables will show the severe deterioration as in the single-variable swapping situations. We still need to be careful because two or more of the *y*-variables may be sufficient to re-identify in some of the records just as we can re-identify with two variables in the example of the previous paragraph.

In some situations, individuals may perform very limited group swapping to reduce analytic deterioration. Limited swapping might best be applied to only a set (or subset) of the most easily re-identified pairs of records across the original file **X** and the masked file **Y**. The swapping was limited to a set of subdomains that partitioned the files X and Y. Even a very limited group swapping rate of less than 0.005 can significantly affect analytic properties (Kim and Winkler 1995) in subdomains that were significantly different from the subdomains in which swapping was controlled. By *controlled to subdomains*, we mean that the group swapping of a set of values in a record must be done with another record in the same subdomain.

Rank Swapping

In ranking swapping, values of fields are swapped across records within a specified proportional range according to an independent sort of the values (Moore 1997). The procedure is:

Begin with the set of records $\mathbf{X} = (x_1, \ldots., x_n)$. Sort each continuous variable x_i.
Randomly swap values of x_i within a *p*-percent range according to the rank ordering.

The difference between ordinary swapping in the previous section and rank swapping is that rank swapping puts restrictions on the range of values (in terms of the ordered distribution) in which swapping can occur. Rank swapping at the rate p=100% is the same as single-variable swapping. Swapping at a 10% rate means that a 10% of values of

a variable are swapped with 10% of the values of the variables in the entire range of the distribution of values.

If x_1 has 1000 values, then after sort of x_1 get values that we rename ($y_{1,1}$, …, $y_{1,1000}$). If p% is 5%, then we must randomly swap any k^{th} value $y_{i,k}$ with values between $y_{i,k-50}$ and $y_{i,k+50}$. The swapping is done without replacement. That is, each swap represents a pair of values that, after being swapped, are not eligible to be swapped again. The 5% range represents 50 records in either direction. The tails of the distributions (i.e., either lower or upper) must be dealt with via heuristics.

As the value of p goes to 0, the amount of distortion in the rank-swapped (masked) file **Y** decreases.

Table 8a. Correlations after Rank Swap 5%

	x1	x2	x3	x4
x1	1.000	0.883	0.199	-0.063
x2	0.883	1.000	0.206	-0.060
x3	0.199	0.206	1.000	0.366
x4	-0.063	-0.060	0.366	1.000

Table 8b. Correlations after Rank Swap 10%

	x1	x2	x3	x4
x1	1.000	0.854	0.171	-0.077
x2	0.854	1.000	0.169	-0.052
x3	0.171	0.169	1.000	0.364
x4	-0.077	-0.052	0.364	1.000

Table 8c. Correlations after Rank Swap 20%

	x1	x2	x3	x4
x1	1.000	0.733	0.121	-0.163
x2	0.733	1.000	0.152	-0.077
x3	0.121	0.152	1.000	0.306
x4	-0.163	-0.077	0.306	1.000

We observe that rank swapping at a 5% rate approximately preserves correlations. At 10% and 20% rank swapping rates, deterioration of correlations is substantial. For instance, in Table 8b, the correlation between x1 and x2 is 0.854 and between x1 and x3 is 0.171; in Table 1, the correlations are 0.881 and 0.186, respectively. In section 4, we will show that it is possible to get quite high re-identification rates with rank-swapped files.

Microaggregation
Microaggreation replaces values of variables with a single value such as an average or median that is within the subrange of the values of the variable. It has substantial similarity to rank swapping and is also quite easy to implement.

Single-ranking k-microaggregation. $\mathbf{X} = (x_1, \ldots, x_n)$. Sort each continuous variable x_i. Group values into successive groups of size *k* (or more but less than *2k-1*). Replace the values in each group by the group-mean or group-median. Repeat for each variable.

Because it is known that single ranking microaggregation can yield very high re-identification rates (Winkler 2002, Muralidhar 2003 private communication, Muralidhar and Sarathy 2006a), individuals have introduced methods for multivariable microaggregation. Domingo-Ferrer and Mateo-Sanz (2002) noted that the single-variable microaggregation can cause modest distortion in regression coefficients with small values of *k*. We will cover re-identification methods for microaggregated files in section 3.

Multivariable Microaggregation
In this procedure, all variables in $X=(x_{ij})$ are used in a clustering procedure to determine a set of cells in which each cell contains between *k* and *2k-1* records. The procedure for creating a *k*-microaggregated file is known to be NP Hard but heuristic procedures may be used (Domingo-Ferrer and Mateo-Sanz 2002). Typically each x-variable $\mathbf{X}_j$ is transformed into a common range and the metric used is the standard Euclidean metric on R^n.

Although the empirical data **X** are not described in detail, the data **X** are taken from a large file with many variables in which subsets had characteristics similar to the characteristics of the data in Table 1. Only four variables are considered. The file is a 100-record subset of a larger file. The file is somewhat similar to the artificial data used for most of the empirical examples. The data are divided into 10 clusters. To produce masked data **Y**, each record in a cluster is replaced by its centroid (average of the 10 records). This is 10-microaggregation.

The two correlation tables (9 and 10) demonstrate that *k*-microaggregation will not preserve correlations. For instance, in Table 9, the correlation between x3 and x4 is 0.95; in Table 10, the correlation between x3 and x4 is -0.305. It is likely that higher moments are also seriously compromised. The file of the empirical example is merely an extract of a file in which two variables are chosen that are (highly) correlated and a third variable that is relatively uncorrelated with the first two.

Table 9. Extracted File – Original Correlations

	x1	x2	x3	x4
x1	1.000	-0.274	0.316	0.273
x2	-0.274	1.000	0.044	0.039
x3	0.316	0.044	1.000	0.935
x4	0.273	0.039	0.935	1.000

Table 10. 10-Microaggregated File – Correlations

	x1	x2	x3	x4
x1	1.000	-0.300	0.347	-0.361
x2	-0.300	1.000	0.041	0.096
x3	0.347	0.041	1.000	-0.305
x4	-0.361	0.096	-0.305	1.000

Most public-use files have *twenty or more variables*. If we have a substantial number of variables, then we can divide the variables into a number of groups G1, …, Gs where the variables in each group are relatively (somewhat) correlated with each other and the correlation across groups is much weaker. The groups G1, …, Gs could be obtained via a general clustering algorithm. We could repeat the multivariable *k*-microaggregation procedure for each subgroup and produce a public-use file **Y**. Analytic properties are also unlikely to be effectively preserved with the cluster-group-microaggregate-within-group procedure. Nin et al. (2008) introduce methods for placing variables in groups that better preserve certain analytic properties after microaggregation. There are still re-identification issues related to the group-then-microaggregate procedures that we will deal with in the next section.

Summarizing comments

We have demonstrated that several of the widely used, easily implemented masking methods distort analytic properties to the point where the masked files are unusable for some (or most) analyses. We are unaware of *any* situations where the most basic of the easily applied methods have been applied to certain specific types of data and several of the analytic properties of the masked data have been justified in comparison to the original, confidential microdata. By *justifying analytic properties*, we mean that a masked file will allow approximate reproduction of one or two analytic properties (statistics) beyond the simple reproduction of means and covariances from the original, confidential file.

3. Elementary Methods for Evaluating Re-identification Risk

In this section we introduce methods of evaluating re-identification risk that are quite straightforward to implement. The methods provide an upper bound that is useful for the providers of the masked microdata. The methods are intended to deal with 'worst case' scenarios. An advantage of the methods is that they often delineate subsets of records that appear easy to re-identify because the records are outliers in some sense that is unexpected in the original analyses by the data providers.

The most straightforward method of investigating re-identification risk is for an agency to use various types of clustering software (or more powerful nearest neighbor or record linkage software) to compare original data **X** with the corresponding masked data **Y**. Although this will give re-identification rates that are too high, the agency can extrapolate the re-identification rates downward by assuming that only outliers in the original data **X**

may be re-identified. With certain types of masking, records in the interior of **X** may be difficult to distinguish from each other and corresponding records in the interior of **Y** may be difficult to distinguish from each other and from records in **X**.

This methodology of directly comparing **X** with **Y** is much more straightforward than alternatives based of cryptographic protocols (Chawla et al. 2005) or Dwork (2006), for various modeling methods (Elamir and Skinner 2006) and (Skinner and Shlomo 2007) that are based on distributional characteristics and loglinear models, for statistical models (Reiter 2005) that differ from the aforementioned models of Skinner and others, and for methods that involve the collection of (possible) intruder data **Z** from public sources and direct re-identification between **Y** and **Z**. We touch on these methods somewhat in the discussion.

As SAS procedures are widely available and understood, we describe primarily analytic means of re-identification using elementary procedures available in SAS. Each original data record $X_{i.}$ and each masked data record $Y_{j.}$ can be thought of as a point in R^n. An intruder might have additional data $Z_{k.}$ that corresponds to both $X_{i.}$and $Y_{j.}$ that contains identifying information such as name, address, and date-of-birth. By comparing Z_k with $Y_{j.}$, the intruder might determine that certain individuals were likely on the original file $X_{i.}$ In our situation, we compare $X_{i.}$ directly with $Y_{j.}$ Because our data $X_{i.}$ are possibly of higher quality than the data $Z_{k.}$ of the intruder, re-identification will be easier and we can better evaluate the risk of re-identification. The methods of re-identification are supplemental to the (similarly) analytic methods that we describe in the next section and are far simpler than the record linkage methods. In the discussion of section 3, we describe some of the weaknesses of the re-identification methods in that they may not be representative of some real world situations. We begin by describing nearest neighbor.

Nearest-Neighbor

Using matrix or array notation we have

X = (x_{ij}) original data, **Y**=(y_{ij}) masked data

As a simplistic procedure, we compare each record (row) of **X** with every row in **Y**. If **X** has *n* rows and **Y** has *m* rows, we do $n \times m$ comparisons. Using the standard Euclidean metric, for each row $(x_{i0,j})$ denote the three closest neighbors by $(y_{i01,j})$, $(y_{i02,j})$ and $(y_{i03,j})$. If $(x_{i0,j})$ corresponds to one of $(y_{i01,j})$, $(y_{i02,j})$ and $(y_{i03,j})$, assume that a re-identification has taken place. The worst situation is when $(x_{i0,j})$ corresponds to $(y_{i01,j})$ (i.e., closest y-record).

Some agencies might have a re-identification policy that states that a re-identification has occurred if one of the closest 3 *y*-records correspond to the *x*-record. Or they might state that might consider the closest 5 or 10 *y*-records. In the first case, we might crudely state that an *x*-record has a 1/3 chance for being re-identified. In the later cases, we might state that an *x*-record has a 1/5 or 1/10 chance of re-identification.

Although there are nearest-neighbor software packages available on STATLIB, we wish to use less powerful procedures in SAS that provide a crude approximation. We choose the SAS procedure according the analytic properties of the data. We will deal

with increasingly more sophisticated examples and methods for masking data and for re-identification later.

In the following, we perform lazy person's nearest-neighbor via SAS clustering. To do this we combine **X** and **Y** records in one file. As **Y** files, we use the different variants of the rounded files.

```
proc fastclus data=d1 maxc=50 maxiter=100 out=d2;
  var x1 x2 x3 x4;
```

The simplistic SAS clustering procedure yields the following re-identification rates with the rounded data given in Tables 6a,b,c.

Table 11. Re-identification Rates with Rounded Data

Round 10 – 100% against highest nearest-neighbor.
Round 50 - ~100% against highest nearest neighbor.
Round 100 - <5% against highest nearest neighbor.

The clustering is designed so that each original record in **X** is clustered (or matched) with one record in **Y** in a manner that minimizes the total sum of squares of the differences between the pairs of records in **X** and in **Y**. Although the clustering procedure is quite effective with this data, this clustering is known to be less sophisticated (i.e., powerful) than nearest-neighbor matching. If **X** were a 10% sample of a larger file **X1** and **Y** is produced by the rounding **X**, then we would still have substantial re-identification with the clustering procedure. If we increased the number of variables, decreased the sampling proportion, and used nearest- neighbor matching or record linkage, then the re-identification rates would still be substantial. We describe this further in the case of micro-aggregation described below for which re-identification is typically much more difficult than in the situation of rounding.

We summarize: While rounding moderately or significantly reduces analytic properties, re-identification risk can remain high. The re-identification risk is based on the fact the points in R^n can still be quite far apart. To preserve analytic properties, restrictions on the locations of the points in R^n may need to be substantial. The set of restrictions over the entire set of **X** and **Y** data along with minimally knowledgeable use of clustering (crude nearest neighbor) yields re-identification. If we were to increase the number of variables to six or eight, then the re-identification rate would increase very significantly from the re-identification rate with four variables.

Some individuals believe that it is not useful to compare a masked version with the original version of the file directly. The direct comparison can be useful because the data provider can determine unexpected re-identifications as in Kim and Winkler (1995). Kim and Winkler only considered those records as being re-identifications that were outliers in the various subdomains in which they were needed to maintain analytic properties. Non-

outliers were not considered re-identifications because the sampling fraction was low. In the discussion, we will deal with the sampling-fraction situation by referring to work by Winkler (1998) in which substantial re-identification (above 50% of true matches determined) occurs at 10% sampling rates with only six variables.

If we use the fast clustering procedure with the rank-swapped data in Tables 8a,b,c, re-identification at the three rates (5%, 10%, 20%) is effectively zero. Can we conclude that rank swapping effectively protects against re-identification? **No**. We may need to try different re-identification procedures. We try to *manually* re-identify using our knowledge of how a rank-swapping file is produced. The 5% rank swapped (i.e., masked) file **Y** is produced from file **X** by swapping values within a 5% range according to sort ordering.

In the following we describe how to do the re-identification manually. Software would do the re-identification more quickly, particularly on larger files.

Procedure for quick re-identification in rank-swapped or microaggregated files
Take files **Y** and **X**. Assume the unique identifier in **Y** is different than in **X** or put one in if it is missing. Create four copies **Xs1**, **Xs2**, **Xs3**, and **Xs4** of **X** where **Xsi** is created by sorting **X** according to variable **Xi**, respectively. Choose any record *y* in **Y**. For the first variable *y1o* in *y*, find the corresponding 5-10 records in **Xs1** that correspond to *y1o*. If there are more than 10 records, choose an alternative *y1o*. For *y1o*, write down the unique identification numbers A1o corresponding to the *x*-records in **Xs1**. Take a different *yi*-variable that is relatively uncorrelated with *y1*. Take *yio* from record y and compare it with **Xsi**. Write down the unique identification numbers Aio corresponding to *yio*. Typically, the intersection Aio∩A1o will contain the desired *single* unique identifier that relates y to a unique x-record.

With rank swapping and the above procedures with *only* two variables, we can sometimes achieve nearly 100% re-identification rates. We observe that we can use the analytic characteristics to manually re-identify. In section 4, we will note additional re-identification for files that have valid analytic properties. Nin et al. (2007a) provide a rank-swapping method that is enhanced with a random swap that somewhat reduces re-identification risk. Their empirical work (also work in Nin et al. 2007b) confirms the re-identification risk with rank swapping and with single-variable microaggregation.

If we have a number of variables that we divide into several groups G1, …, Gk in which there is relatively higher correlation between variables within groups and relatively lower correlation across groups, then it is likely that multi-variable microaggregation within groups will still yield moderate re-identification risk in many situations. The re-identification could be done with a procedure that mimics the above single-variable re-identification procedures. If the original file has 10,000 records, there are three groups G1, G2, and G3, and we 100-multi-variable microaggregate in each group, then any given record will be associated with 100 records A1o, A2o, and A3o in each of groups G1, G2, and G3, respectively. If the intersection of A1o, A2o, and A3o is a unique record (which it typically will be), then a re-identification occurs. Nin et al. (2008) provide methods for systematically improving the clustering of the variables that enter groups G1, …, Gk to enhance analytical properties but demonstrate that re-identification risk can still be moderate or greater.

Summarizing comments
We have provided an indication that some re-identification may be possible with these files that have been masked with these elementary methods. Our methods were primarily to use the analytic properties of the files to demonstrate situations where some records could be re-identified. Much more powerful re-identification methods that very specifically use analytic properties to construct new metrics in nearest-neighbor or record linkage software are likely to re-identify at higher rates. Better use of analytic knowledge of the characteristics of a population file will raise re-identification even further.

4. Discussion

The discussion consists of a critique of the re-identification risk and a listing of some masking methods for which analytic properties have been justified.

Putting a large number of analytic restraints on synthetic data may necessitate that the synthetic data agree exactly with the original, confidential microdata. As an example, a file of economic data for businesses (i.e., continuous data) contains twelve variables that are known to (approximately) be normal. If users want the public-use data have twelve variables to preserve means and covariances on a number of subdomains and the some of the subdomains contain 30 or fewer records, then the number of analytic constraints exceeds the number of degrees of freedom. The only data that can preserve the analytic properties on the small subdomains are the original microdata.

4.1. Re-identification Risk

Many individuals have assumed that sampling provides subsets of a population with *automatically justified* analytic properties and severely reduced re-identification risk. If the masked, public-use file **X1** has variables that are distorted in certain ways and the intruder file **Y** has distortions that naturally occur in the values of the differing variables, then re-identification risk is substantially reduced (or eliminated).

These beliefs may be somewhat valid in a narrow range of situations but are misplaced in general because of several factors. Most public-use files **X1** are created from confidential files **X** that are typically believed to be of high quality and contain 20 or more variables. Historical beliefs were that intruders would not have most of the variables in the public-use file and most re-identification would be via naïve, exact comparison methods. Sweeney dramatically demonstrated that many public use files could not be considered confidential because they contained ZIP code, date-of-birth, and sex that uniquely identified as much as 87% of the population using readily available voter-registration data. The point is that three variables (in files having sometimes more than 100 variables) allowed easy manual re-identification. Sweeney also demonstrated how to obtain publicly available data that could assist in re-identification of small numbers of records in some files using manual methods only.

Kim and Winkler (1995) demonstrated that a very small proportion of records could be re-identified because of analytic restrictions of the public-use files on some subdomains. Their conservative assumption was that outliers in the subdomains (even at small

sampling rates) would be outliers in the original population. They demonstrated that 6-8 quantitative variables on the subdomains (defined by several discrete variables) were often sufficient for re-identification. With multivariate normal data as in Fuller (1993), Winkler (1998) compared a 10% sample file that had been masked with additive noise to the original population file. In the six-variable situation, Winkler was able to identify more than 70% of the records with probability greater than 50%. If analytic properties need to be maintained on subdomains and there are 10 or more variables, then it seems that moderate re-identification can occur with sampling fractions approaching 1%.

We note that the re-identification would be substantially more straightforward with either rounding or with single-variable micro-aggregation than with additive noise. This is due to the more straightforward (and structured) way that rounding and single-variable micro-aggregation induce changes in the masked file. With each individual in the masked file, it is quite straightforward to reconstruct approximately the distribution of the original, confidential values of the variables. The approximately reconstructed distributions can, in turn, be used to created metrics that significantly increase re-identification rates. We note that the metrics measure how far a value of a masked value of a variable can deviate from the other masked values of the same variable and from the values in the original distribution. As we showed earlier, the re-identification can, in fact, be performed manually (but re-identification is often much faster with appropriate software).

Yancey et al. (2002) showed that new metrics developed by Kim and Winkler (1999) allowed higher re-identification rates (still small) than what Kim and Winkler (1995) had originally obtained with different metrics. The first point is that sampling can only protect records that are in the interior of point clouds where the point clouds are only based on a few variables. The second point is that more sophisticated re-identification metrics on some variables will allow additional re-identification. It is almost impossible to anticipate the multiple methods that an intruder will use in re-identifying on a particular file.

4.2. Masking Methods with Justified Analytic Properties

In the following, by *justified* analytic properties, we mean the ability to support one or two sets of analyses that are supported by the original microdata. The authors that have justified the analytic properties of their microdata have also been careful to note that it can often only be used for a pre-specified set of analyses. The have typically noted specific limitations on the set of analyses.

Methods for producing analytically valid synthetic data from valid models have been clearly been justified in most literature. These include Latin Hypercubes (Dandekar et al. 2002), Blank-and-Impute (Kennickell 1997; Woodcock 2002, 2004) and pure synthetic data (Reiter 2002, 2005; Raghunathan et al. 2003). Various methods for creating partially synthetic data have also been justified (Muralidhar and Sarathy 2002, 2006a,b, 2007; Little and Liu 2002, 2003; Reiter 2005).

Other justified methods are additive noise (Kim 1986), mixtures of additive noise (Yancey et al. 2002), general perturbation (Muralidhar and Sarathy 2002, 2006a,b) and the Post Randomization Method (Gouweleeuw et al. 1998; Van Den Hout and Van der Heijden 2002; De Wolf 2006). The Post Randomization method is more difficult to

implement and its ability to provide valid analytic properties appears to hold in a narrow range of situations.

4.3. Two Principles of Masking Methods

We can summarize our main two points about masked, public-use microdata as follows.

1. Individuals should first justify the analytical properties of a public-use file **X1**.

2. With an analytically valid public-use file **X1,** individuals should then apply effective re-identification methods to assure that risk is within acceptable levels.

4. Concluding Remarks

This paper provides examples that demonstrate that many of the widely used, easy-to-implement masking procedures do not yield microdata that preserve analytic properties from original, confidential microdata.

1/ This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress). Any views expressed on (statistical, methodological, technical, or operational) issues are those of the author(s) and not necessarily those of the U.S. Census Bureau. The author acknowledges a number of comments by Philip Steel and Rolando Rodriquez that led to improvement in the exposition.

References

Abowd, J. M., and Woodcock, S. D. (2002), "Disclosure Limitation in Longitudinal Linked Data," in (P. Doyle et al, eds.) *Confidentiality, Disclosure, and Data Access*, Amsterdam, The Netherlands: North Holland.

Abowd, J. M., and Woodcock, S. D. (2004), "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases 2004*, New York: Springer.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Burkhauser, R. V., Feng, S., and Jenkins, S. P. (2007), "Using the P90/P10 Index to Measure U.S. Inequality Trends with Current Population Data: A View from Inside the Census Bureau Vaults," Center of Economic Studies Discussion Paper CES 07-17, available at http://www.ces.census.gov/index.php/ces/cespapers .

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

Dandekar, R., Cohen, M., and Kirkendal, N. (2002), "Sensitive Microdata Protection Using Latin Hypercube Sampling Technique," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, New York: Springer, 117-125.

De Wolf, P.-P. (2007), "Risk, Utility and PRAM A Comparison of Proximity Swap and Data Shuffle for Numeric Data," in (J. Domingo-Ferrer, ed.) *Statistical Data Protection 2006*, Springer: New York, N.Y.

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002), "Practical Data-Oriented Microaggregation for Statistical Disclosure Control," *IEEE Transactions on Knowledge and Data Engineering*, 14 (1), 189-201.

Feng, S., Burkhauser, R. V., and Butler, J.S. (2006), "Truncation Bias and the Measurement of Income Inequality," *Journal of Business and Economic Statistics*, 6 (3), 335-337.

Fienberg, S. E. (1997), "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, commissioned by Committee on National Statistics of the National Academy of Sciences.

Fienberg, S. E. and MacIntyre, J. (2005), "Data Swapping: Variations on a Theme of Dalenius and Reiss," *Journal of Official Statistics*, 21 (2), 309-323, also http://www.jos.nu/Articles/abstract.asp?article=212309 .

Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, 9**,** 383-406 (http://www.jos.nu/Articles/abstract.asp?article=92383).

Gouweleeuw, J.M., Kooiman, P., Willenborg, L. C. R. J., and De Wolf, P.-P. (1998), "Post Randomisation For Statistical Disclosure Control: Theory and Implementation," *Journal of Official Statistics*, 14, 463-478, also http://www.jos.nu/Articles/abstract.asp?article=144463 .

Kennickell, A. B. (1999), "Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 248-267 (available at http://www.fcsm.gov under Methodology reports).

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 370-374 (http://www.amstat.org/sections/SRMS/Proceedings/papers/1986_069.pdf).

Kim, J. J., and Winkler, W. E. (1995), "Masking Microdata Files," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 114-119 (http://www.amstat.org/sections/SRMS/Proceedings/papers/1995_017.pdf , longer report http://www.census.gov/srd/papers/pdf/rr97-3.pdf) .

Kim, J. J., and Winkler, W. E. (1999), "Multiplicate Noise for Masking Continuous Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, CD-ROM, also http://www.census.gov/srd/papers/pdf/rrs2003-01.pdf .

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*, 9**,** 313-331 (http://www.jos.nu/Articles/abstract.asp?article=92313).

Lane, J. (2007), "Optimizing the Use of Microdata: An Overview of the Issues," *Journal of Official Statistics*, 23 (3), 299-317 (http://www.jos.nu/Articles/abstract.asp?article=233299).

Little, R. J. A., and Liu, F. (2002), "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, CD-ROM, also http://www.bepress.com/umichbiostat/paper6/ .

Little, R. J. A., and Liu, F. (2003), "Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, CD-ROM.

Little, R. J. A. and Rubin, D. B. (2002), *Statistic Analysis with Missing Data (2nd Edition)*, New York, N.Y.: John Wiley.

Moore, R. (1995), "Controlled Data Swapping Techniques For Masking Public Use Data Sets," U.S. Bureau of the Census, Statistical Research Division Report rr96/04, (available at http://www.census.gov/srd/www/byyear.html).

Muralidhar, K., and Sarathy, R. (2003), "A Theoretical Basis for Perturbation Methods," *Statistics and Computing*, 13 (4), 329-335.

Muralidhar, K., and Sarathy, R. (2006a), "Data Shuffling – A New Masking Approach to Numerical Data," *Management Science*, 52 (5), 658-670.

Muralidhar, K., and Sarathy, R. (2006b), "A Comparison of Multiple Imputation and Perturbation for Masking Numerical Variables," *Journal of Official Statistics*, 22 (3), 507-524, also http://www.jos.nu/Articles/abstract.asp?article=223507 .

Muralidhar, K. and Sarathy, R. (2007), "'Easy to Implement' is Putting the Cart before the Horse: Effective Techniques for Masking Numerical Data," Federal Committee on Statistical Methodology Research Conference, to appear on CD-ROMl.

Nin, J., Herranz, J., and Torra, V. (2007a), "Rethinking Rank Swapping to Decrease Disclosure Risk," *Data & Knowledge Engineering*, 64, 346-368.

Nin, J., Herranz, J., and Torra, V. (2007b), "On Method-specific Record Linkage for Risk Assessment," UNECE Work Session on Statistical Data Confidentiality, Manchester, UK, also http://www.unece.org/stats/documents/2007/12/confidentiality/wp.7.e.pdf .

Nin, J., Herranz, J., and Torra, V. (2008), "Attribute Selection in Multivariate Microaggregation," *EBDT 2008*, to appear.

Palley, M. A., and Simonoff, J. S. (1987), "The Use of Regression Methodology for the Compromise of Confidential Information in Statistical Databases," *ACM Transactions on Database Systems*, 12 (4), 593-608.

Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, 19, 1-16, also http://www.jos.nu/Articles/abstract.asp?article=191001 .

Reiter, J.P. (2002), "Satisfying Disclosure Restrictions with Synthetic Data Sets," *Journal of Official Statistics*, 18, 531-543, also http://www.jos.nu/Articles/abstract.asp?article=184531 .

Reiter, J.P. (2003a), "Inference for Partially Synthetic, Public Use Data Sets," *Survey Methodology*, 181-189.

Reiter, J.P. (2005), "Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study, *Journal of the Royal Statistical Society, A*, 168 (1), 185-205.

Van Den Hout, A., and Van Der Heijden, P. G. M. (2002), "Randomized Response, Statistical Disclosure Control, and Misclassification: A Review, *International Statistical Review*, 70 (2), 269-288.

Winkler, W. E. (1998), "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," *Research in Official Statistics*, 1, 87-104, http://www.census.gov/srd/papers/pdf/rrs2005-09.pdf .

Winkler, W. E. (2004), Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Database*, Springer: New York, 231-247, also http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf .

Winkler, W. E. (2005), "Modeling and Quality of Masked Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Method*, CD-ROM, also http://www.census.gov/srd/papers/pdf/rrs2006-01.pdf .

Winkler, W.E. (2007a), "General Discrete-data Edit/Imputation Software with Applications to Statistical Matching and Synthetic-data Generation under a Variety of Constraints," documentation of software.

Winkler, W. E. (2007b), "Analytically Valid Discrete Microdata Files and Re-identification," technical report, presented at the 2007 Annual Meeting of the American Statistical Association, available at http://www.census.gov/srd/www/byyear.html .

Winkler, W.E. (2008), "General Methods and Algorithms for Modeling and Imputing in Discrete Data under a Variety of Constraints," technical report.

Yancey, W.E., Winkler, W.E., and Creecy, R. H. (2002) "Disclosure Risk Assessment in Perturbative Microdata Protection," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, New York: Springer, 135-151, (also http://www.census.gov/srd/papers/pdf/rrs2002-01.pdf).