Modeling and Quality of Masked Microdata

William E. Winkler, william.e.winkler@census.gov [1]
U.S. Bureau of the Census, Room 3000-4, Washington, DC 20233-9100

Abstract

Statistical organizations collect data via survey forms and other methods. The microdata are valuable for modeling and analysis. To produce a public-use file, the organizations mask the data in a manner that may prevent re-identification of data associated with individual entities. The public-use microdata may allow one or two sets of analyses that approximately reproduce analyses that could be performed on the original microdata. This paper describes a general method of creating models of data that is related to methods of creating appropriate aggregates of data that are needed for sufficient statistics in general classes of models (Moore and Lee 1998, DuMouchel et al. 2000, Owen 2003). If the aggregates can be approximately reproduced, then the masked microdata may allow one or more analyses that correspond to analyses on the original, non-public microdata. It will typically not yield data suitable for general analyses.

Keywords: Data Mining; Likelihood, Loglinear, Multivariate

1. Introduction

Statistical agencies collect data via surveys and other sources that are used in producing aggregate statistics that are used in publications. Increasingly, there is demand for public-use data files that can be used for a variety of analyses that are not covered by the published aggregates. There are two issues. The first is that the public-use data should be suitable for several analyses that approximately reproduce analyses that are performed on the original, non-public data. The strengths and the limitations of the public-use data should be described in detail. The second is that the privacy of data associated with individual entities should be maintained. In this paper, we will be primarily concerned about the former issue but will indicate situations where re-identification of information with individuals may be straightforward.

There are two ways in which an analysis of microdata might be described. In the first, more difficult, method, we might do detailed modeling of the distributional characteristics of the data that are used for evaluation and prediction. In the second method, which might be considered as a special case of the first method, we might specify a number of aggregates that are needed for an analysis. For instance, we might specify the particular moments of continuous variables that are needed for analysis. In a regression, we might specify that the first two moments need to be preserved with very high accuracy. In loglinear modeling we might specify the marginal totals or other quantities that must be preserved for an analysis to be accurate.

In producing a public-use file **Z**, we assume that the original, non-public data **X** are of high quality and then are systematically distorted (i.e., *masked*) so that we cannot re-identify information with individuals. By high quality, we mean that the list of entities is complete, unduplicated, and representative of the desired population and that values associated with the individual fields correspond to some underlying reality. For instance, the income of an individual in a record might correspond accurately to the actual income of the individual. The marital status might accurately correspond to actual marital status. The date-of-birth would be accurate and could be used for computing age at different dates. In this possibly idealized situation the original data **X** could be used for several analytic purposes.

Our key concept of quality is that the data are "fit for use" in the sense of reproducing one or more analyses. The "fit for use" and other definitions of quality such as the Eurostat set (Haworth et al. 2001) are too general and vague to be applied in a straightforward manner. We will be primarily concerned with accuracy and comparability for which we can specify certain aggregates such as the first two moments that are needed in a regression analysis. As an example, we might like the first two moments that are computed on the masked data **Z** to approximately agree with the first two moments on the original data **X**. The approximate agreement should be sufficiently close that regression coefficients almost agree and that variance of the regression coefficients either agrees or increases slightly according to which masking procedure has been used.

The goal of the paper is to supply a few ideas for systematic methods that, given a specific analytic need, can be applied to determine the accuracy of an analysis on masked data **Z**. This can be somewhat straightforward when we show how closely a set of aggregates computed on **Z** is to the corresponding set of aggregates computed on **X**. If we want to show that a regression can be done on an entire set **Z** and on several subdomains of **Z**, then the aggregates place substantial restrictions of the specific records in **Z**. If the set of restraints is sufficiently large in the sense of exactly preserving a large number of aggregates from the original data **X**, then the masked data **Z** may need to agree with the original microdata **X**. If the restraints only need to be preserved within some small epsilon, then portions of some records in **Z** may necessarily be close to some records in **X**. As a specific instance, combinations of fields in some records in **Z** may have values in the tails of certain distributions causing them to be much more easily re-identified.

The outline of this paper is as follows. Following this introductory section, we provide additional background on some of the methods for masking data **X** to put it in a form suitable for public-use data **Z**. The masking methods are often easily implemented methods that have not been justified in terms of providing a file with analytic properties or preventing re-identification. The exceptions for analytic properties are suitably adjusted additive noise (Kim 1986, 1990; Fuller 1993; Yancey et al. 2002) and synthetic microdata (Raghunathan et al. 2003; Reiter 2002; Kennickell 1999; Abowd and Woodcock 2002, 2004). The clear exception for not being re-identifiable is k-anonymity (Samarati and Sweeney 1998, Sweeney 2002, Bayardo and Aggrawal 2005) for which no analytic properties have ever been established. In the third section, we provide some methods that might be used for defining analytic properties and quality of public-use microdata. The primary intuition is that the microdata should be suitable for an intended use. The third section provides some comments and research problems. In the final section, we provide some concluding remarks.

2. Background and Definitions

This section is divided into three subsections. In the first section, we provide general background on some of the general methods that relate to analytic properties of microdata. In the second section, we give more detail of the specific methods used in statistical agencies. We quickly reject many of the methods because they have never been justified as providing valid analytic properties in the masked microdata. In the third section, we describe a crude, but specific, definition of quality that might be applied in evaluating analytic properties of microdata.

2.1. General Analytic Properties of Masked Microdata

Users of public-use microdata are primarily concerned with the analytic properties of data. Given that the original, non-public microdata are of high quality, then the masking procedures used in producing a public-use file will reduce the quality. The masked files may only allow zero (see section 2.2 for examples) or one analyses. Because we are primarily concerned with methods that have demonstrated analytic properties, we begin with an example that involves regression and additive noise.

Kim introduced the idea of additive noise in a form such that data **X** might be transformed to data $\mathbf{Y} = \mathbf{X} + c\,\varepsilon$ that is in turn linearly transformed to data **Z** (also see Tendick and Matloff, 1994). The independent noise ε has mean zero and the same covariance Σ as **X**. Data Y is the same mean as data **X** and $\text{cov}(\mathbf{Y}) = (1+c)\,\Sigma$. The positive constant c is between 0 and 0.5. The data **Z** has the same means as **X** and approximately the same covariances and correlations as **X**. Regression analyses that are possible on **X** can be approximately reproduced on **Z** in the sense that regression coefficients are the same (very slight bias) but with slightly inflated coefficients of variation. Kim (1986, 1990) and Kim and Winkler (1995) demonstrated that some analyses such as regression were still possible on some subdomains (that were sufficiently large and with some additional properties) but that most subdomain analyses on **Z** did not correspond to subdomain analyses of **X**.

Both Fuller (1993) and Lambert (1993) indicated that a masking procedure used in producing public-use data **Z** from data **X** should approximately preserve means, covariances, and one additional analytic property. Brand (2002) observed that few additional statistical properties were preserved in the files **Z** due to the effect that the basic additive noise procedure had on the detailed distributional properties of the original data **X**. Fuller (1993) noted that it might be possible re-identify a small proportion of records in **Z** that are masked via the basic additive noise procedure. Yancey et al. (2002) demonstrated that a mixture of additive noise approach could still provide the same analytic properties as the Kim approach while significantly reducing re-identification risk. Mixtures of additive noise (e.g., Yancey et al. 2002) would have somewhat greater deleterious effects of the distributions than the effects of basic additive noise.

Several authors have suggested creating synthetic microdata **Z** that reproduces some of the aggregates obtainable from the original data **X** or allows reproduction of a model **M** that could be built on **X**. Raghunathan et al. (2003) provided the theory for estimating the variances of certain univariate estimates of synthetic data obtained via a valid probabilistic modeling procedure. Reiter (2002) had earlier demonstrated that certain properties of the original data **X** that were not explicitly included in the model **M** would never be in the synthetic data **Z**. Reiter also showed that, due to lack of data, some properties that were in the model **M** would not necessarily be in the synthetic data **Z**. Various authors have demonstrated that, if a sufficiently detailed model **M** of the data **X** is developed, then some of the records in the synthetic data **Z** will necessarily be close to some of the real original data records **X** (Fienberg 1997, Reiter 2002, Winkler 2004).

Certain distributional outliers (that are often easy to detect) may be needed for accurately estimating multiple moments. Kim and Winkler (1995, see appendix to longer web report) provided controlled distortion procedures for changing outliers while (approximately) preserving sets of moments in sufficiently large domains. If only a few outliers need to be changed because they cause records to be easily re-identified, then a large number of aggregates may need to be modified. These types of modifications will yield moderate or substantial changes in a very large number of records.

The Kim-Winkler procedure, like the more general DuMouchel et al. (1999) procedure (see section 3.1), depends on the subdomains where the controlled distortion procedures are applied to be quite large. The values of fields in individual records are the unknowns that must be solved so that the moment equations are satisfied. If there are

insufficient records in the subdomain, the equations cannot be solved.

We can summarize this section as follows. Masking is a procedure that is intended to create (partially) synthetic data **Z** that satisfies one or two analytic characteristics and prevents re-identification. If we require more analytic restraints on the data **Z**, then the data **Z** are likely to be close to the original data **X**, particularly in the tails of the distributions of certain variables. Any further masking of the outlier(s) that preserves most of the original analytic properties of the data **X** is likely to reduce the ability to perform auxiliary analyses. Unlike some of the synthetic-data generation methods, masking (with, say, small amounts of additive noise) is intended to allow some auxiliary analyses that have not been specified a priori. Whether additional analytic properties are preserved should be checked.

2.2. Specific Masking Methods Traditionally Used in Statistical Agencies

Statistical agencies have typically adopted masking methods because they are *easy to implement*. The easiest-to-implement methods seldom, if ever, have been justified in terms of preserving one or two analytic properties or in preventing re-identification. In extreme situations, the crude application of masking methods may yield a file that cannot be used for analyses and yet still allows some re-identification.

If a statistical agency needs to produce a public-use file, then the agency should first consider the analytic properties that might be needed in the masked data **Z**. The best situation is when a single group of users with very focused analytic needs (say for only one set of analyses) will be using the data.

Before beginning with some details, we need preliminary notation. We denote the original microdata by **X** and individual data records by $r_i = (x_{i1}, x_{i2}, x_{i3}, ….., x_{in})$ where value x_{ij} is associated with the jth value of the i^{th} field. Individual fields (or variables) are denoted by $\mathbf{X}_i$. The variables $\mathbf{X}_i$ can be continuous or discrete. For instance, $\mathbf{X}_i$ might be a geographic identifier such as a State Code. Or $\mathbf{X}_i$ might be a continuous variable such as income. Generally, we will assume that masked data **Y** has the same number of fields n and same number of records m as the data **X**. This does not need to be the case. These assumptions will also hold for any data **Z** that are obtained from **Y** because the data **Z** may be easier to analyze.

The idea of the masking is that each masked record $s_k \in \mathbf{Z}$ will be difficult to re-identify with an original records $r_i \in \mathbf{X}$ for some i. As an example, assume that data $\mathbf{X} = (\mathbf{X}_1, …., \mathbf{X}_n)$ have the form (ZIP, Profession, Education, Age, Sex, Race, Income) where Profession is given 1000 categories, Education is broken into 14 categories, and Income is rounded into thousands of dollars consists of 10,000 records for a particular state. It is possible that two variables alone may cause a re-identification. For instance, if there is only one surgeon in a particular ZIP code, then it may be possible to associate the name of the surgeon with the income in the file. There are several straightforward things that might be done. The professional categories might be broken into 50 collapsed categories in which all types of physicians are given the category doctor. In the collapsed data **Y**, there may be several physicians in the ZIP code and re-identification might not be easy. If the doctor has exceptionally high income, then the income might be replaced with a topcoded income of $400K dollars making it more difficult to re-identify an income with the particular doctor. If Education and Race on blanked (given a single neutral value that effectively removes the Education and Race variables from the file), then it may still be possible to re-identify the doctor using ZIP, truncated income, age, and sex. Although it is not possible to find the exact income of the doctor, the fact that we can reasonably state the doctor's data (in a severely altered form) is present in the masked data **Z** may still represent a disclosure.

From the above example, we can see that masking can place severe limitations on the analytic properties of microdata. Even in the situations where the masking is severe, it may still be possible to re-identify some information in some records. The following easy-to-implement masking methods have never been justified in terms of preserving analytic properties.

1. *Blanking values in certain fields in a selectively chosen number of records.* This causes missing data (a type of item nonresponse) that makes the masked file unsuitable for use in a statistical or data mining software package. This method is also referred to as *local suppression*.
2. *Swapping values from certain fields across records.* This is easy to implement. If only a few values of a few fields are swapped, then the resultant masked file has no analytic properties beyond the means of individual fields that are preserved. No correlations, regressions, or loglinear modeling properties are preserved.
3. *Rank swapping values in records*. Single-variable rank swapping (Moore 1995) involves sorting the values in individual fields (continuous variables) and swapping values p% of the ranked distance. If the swapping percentage p exceeds a small value (say 1-2%), then the correlations can be severely affected. As a special case of more general re-identification methods developed for microaggregation (see below), re-identification is quite straightforward.
4. *Global recoding of values in certain fields.* This involves collapsing the number of values that a variable can assume. For instance, in a geographic identifier, 50 State codes might be replaced by 4 codes representing for regions. Individual ages might be replaced by aggregates consisting of 5-year age ranges.

Efficient algorithms (Bayardo and Agrawal 2005, Lefevre et al. 2005) for global recoding to achieve k-anonymity currently exist. A file is k-anonymous if every record $r \in \mathbf{Z}$ has *k-1* records that are identical to it. The analytic properties of global recoding (and k-anonymity) are currently research problems.

5. *Microaggregating records*. Some authors have suggested sorting individual fields and replacing groups of *k*-values with an aggregate such as the average or median (called single-variable *k*-microggregation). Generally *k* is taken to be 3 or 4 because simple analytic properties such as correlations degrade rapidly as *k* increases. Winkler (2002) showed how to build new metrics for comparing fields into re-identification (record linkage) software quickly and easily based on the observed distributions of the *k*-microaggregates associated with individual fields. Two uncorrelated variables were often sufficient to obtain moderately high re-identification rates. Some authors (Domingo-Ferrer et al. 2002) have suggested using several variables at a time in a clustering procedure in which the values of records associated with each cluster are replaced by an aggregate such as the centroid. As they suggest and can be easily checked, such as *n*-variable, *k*-aggregation can quickly degrade simple statistics such as correlations.

There has been additional work that demonstrates one or two analytic properties of masked microdata.. We mention a few that are representative of several others. Kennickell (1997) extended multiple imputation ideas to the generation of 'synthetic' microdata by successively blanking values in fields and filling in the data according to a (multiple-imputation) model. The ideas have also be effectively applied and extended by Abowd and Woodcock (2002, 2004). The authors have made substantial effort to demonstrate the analytic properties that are preserved and some of the limitations of the microdata. The issue of whether the Kennickell-types of methods allow re-identification is still open. Dandekar et al. (2002) demonstrated that synthetic data that are generated via a Latin Hypercube methodology could preserve some analytic properties.

Dandekar's work and the earlier work on additive noise very strongly indicate that if analytic properties need to be preserved on subdomains, then the modeling needs to be done on the individual subdomains. There are two requirements on the subdomain modeling. The subdomains must partition the entire set of records. Each subdomain must be reasonably large. For very simple analytic properties, the subdomain will likely need a thousand or more records. For a moderate number of aggregates, the subdomain may need many tens of thousands of records (DuMouchel et al. 1999, Kim and Winkler 1995). If any analyses need to be performed on the set of subdomains and on the entire file, then each record must satisfy two sets of analytic restraints. In the simple case of regression (which also preserves correlations), two sets of moment conditions need to be satisfied by all of the data records (with either additive-noise-masked data or synthetic data).

2.3. Quality

If the original data **X** has values in individual fields that correspond to some underlying reality, then the data will satisfy all aggregate restraints (even on subdomains) and be suitable for many analyses. We are unaware (e.g., Winkler 2003) of documents that clearly demonstrate the quality of original data **X** in terms of being able to allow many accurate analyses.

The quality of the original data **X** affects the quality of the masked data **Z**. If certain aggregates in the original data **X** are in error and the masked data **Z** reproduces the aggregates, then any analysis based on the file **Z** may reproduce certain errors or lead to erroneous conclusions. If the original data **X** are not in error, then the distortions in the masked data **Z** may also lead to erroneous conclusions. This is a sensitive issue. How much detail can a statistical agency provide in terms of the limitations of the masked data **Z**? If the details are sufficient to describe the limitations of the microdata **Z**, will the users of the pubic-use file be able to use the file **Z** and also understand its limitations?

As data **X** are used for additional analytic purposes the file often needs additional 'clean-up' prior to its use in the additional analyses. Such additional analyses often occur when economists or demographers use data **X**. Unresolved (research) issues with the 'clean-up' are whether it sufficiently improves the original data **X** for the additional analysis and whether the 'clean-up' hurts the file **X** for other analyses.

A general definition of data quality is that a file (or set of files) be "fit for use." We more narrowly define data quality as the property that a file has *suitable quality for a particular analysis*. This means that a file allows the reproduction (possibly approximately) of a large number of aggregates (such as suggested by DuMouchel et al. 1999, Moore and Lee 1998) needed for a particular analysis or model. If a file has data quality sufficiently high for two analyses, then it will approximately reproduce two sets of aggregates with limitations that may be described by the statistical agency. The data may have significant other limitations that make it unsuitable for analyses other than the specified one or two analyses.

3. Methods

This section is divided into three components. In the first subsection, we describe a method of analyzing data and a general template that can be used for determining some of the quality characteristics of a set of data. The methods apply to both masked data **Z** and original data **X**. In the second subsection, we go into more detail about some of the

limitations of additive noise. Although many other masking methods can have greater limitations (in senses that are specific to a given analysis), additive noise has been studied more extensively. In the third subsection, we describe two general methods of re-identification for microdata **Z**.

3.1. A General Method of Analysis

Some of the best methods of analysis (e.g. Fienberg 1997, Reiter 2002, Raghunathan et al. 2003) involve a number of components that model detailed conditional distributions that form the basis of the likelihood of the data and a model. In this section, we describe analyses in terms of a number of aggregates that should be easier than the more general synthetic-data-generation methods. We believe that the overall method can serve as a template for how to do some analyses and that the accuracy of the aggregates needed in the analyses determines how well data **Z** allow an analysis that correspond to an analysis on original data **X**. In particular, users of the public-use data **Z** could ask a statistical agency to re-run the software for an analysis on the original data **X**.

Our template requires an agency to determine one or two analyses (possibly via consultation with an appropriate set of potential users) that might be performed on a public-use file. For a specific analysis, the template is

1. delineate the aggregates that are needed for the analysis,
2. produce the public-use data **Z** from original data **X**,
3. determine how well the aggregates from **Z** correspond to the aggregates from **X**,
4. determine limitations of data **Z** in terms of supporting the specific analysis, and
5. refine the masking procedures (if necessary) to provide data **Z** that better support the single analysis.

We observe that, if the aggregates are delineated, then the remaining procedures are quite straightforward.

DuMouchel et al. (1999) provide a systematic generalization of the methods in the template. They showed how to create a set of aggregates consisting of a large number of moments needed in an approximation of likelihoods used in creating particular types of models for a set of continuous data. Moore and Lee (1998) show how to create a large number of approximations of aggregates needed for the loglinear modeling of discrete data. In each of these situations, the authors are interested in taking a file consisting of millions of records and providing a large set of numbers corresponding somewhat to numbers that could be obtained from the original data. A file of millions of records might be replaced by a file of thousands or tens of thousands of aggregated 'records.' The aggregated information or 'records' would approximately reproduce key aggregates in the original file. In our context, we are only interested in how the analytic constraints affect the forms that the microdata can assume in terms of restricting masked microdatata to be close to the original microdata. Because key insights provided by DuMouchel et al. (1999) are representative of similar insights provided by Moore and Lee (1998), we only describe the DuMouchel et al. work.

The procedure of DuMouchel et al. (1999) consists of three components. We are only concerned with their procedure for maintaining moments on subdomains. In DuMouchel et al. (1999), they are concerned with the log-likelihoods

$$\sum_{i=1}^{M} w_i \log(\quad f (B_{i1},..., \quad B_{iC} , Y_{i1},..., \quad Y_{iQ} ; \theta) =$$

$$\sum_{i=1}^{N} \log(\quad f (A_{i1},..., \quad A_{iC} , X_{i1},..., \quad X_{iQ} ; \theta) . \qquad (1)$$

Here f is the likelihood, B's and A's are discrete variables, Y's and X's are continuous data of dimension Q. The B's and Y's are associated with artificial (synthetic) data that might preserve certain aggregates needed in approximating the log-likelihoods. In our situation, we have no discrete variables B's or A's. DuMouchel et al. assume that the log-likelihoods are sufficiently smooth functions that they can be approximated by moments of the continuous variables Y's and X's. The original data might have N=1,000,000 records that we wish to approximate with M=10,000 or M=100,000 records. It was their purpose to create data Y by solving a system of equations corresponding to (1) and satisfying a specified set of moment equations. The new data Y would be suitable for use in data mining or statistical software packages that can handle data Y with weights w_i where the original data X was too large for most software.

As DuMouchel et al. (1999) show, each record in **X** may represent multiple records in **Y**. The degrees of freedom in solving the equations are represented by the number of records M associated with **Y**. As the number of moment constraints increases, the reduction factor in going from N to M can go from a factor $F > 1$ down to 1. In other words, as the number of moment constraints increases, the size M of data **Y** will increase until $M=N$.

Superficially, the procedure of DuMouchel et al. (1999) is similar to Kim and Winkler's (1995) controlled distortion procedure. The number of records and number of variables approximately represent the degrees-of-freedom in the computation. Kim and Winkler (1995) were concerned with moving easily identified outliers (specific values of certain fields in certain records) into the interiors of distributions while preserving the first two moments of the data. As a value of an outlier was changed, a possibly large number of values in fields in other records need to change. If there were not sufficient records in a domain where moments needed to be preserved, then the set of equations associated with the moments could not be preserved.

As an alternative to the DuMouchel et al. (1999) methods, Owen (2003) observed that, in some situations, it is more straightforward to draw a random sample and use empirical likelihood to preserve some of the analytic restraints.

Whereas the DuMouchel et al. methods may require special programming, there is high quality generalized software for empirical likelihood.

3.2. More on Limitations of Masked Microdata

Only a few of the masking methods have been demonstrated to provide one or two analytic properties. In this section we give insights why we cannot generally expect many other analytic properties to be valid. To begin, we provide a comparison of the analytic information that is available in a set of continuous variables using transformed additive noise (Kim 1986) and data generated according to the moment conditions of DuMouchel et al. (1999). We note that the masked data will only preserve the first two moments of the continuous data. We simplify even further by assuming that we only preserve the moments on the entire domain. We do not preserve moments on any large subdomains. Although this is possibly the easiest situation that can be realistically dealt with, it will provide substantial insight.

Kim (1986) provided a method of adding noise ε to data **X** where the noise ε had the same covariance structure as the data X. The new data had $\mathbf{Y} = \mathbf{X} + c\,\varepsilon$ had covariance $(1+c)$ cov(**X**). He further provided a linear transform that created data **Z** having the same means as **X** and approximately the same covariances and correlations as **X**. Part of the approximate agreement of the covariances is due to the nature of the random number generation process associated with epsilon. With approximately 10 variables (10×10 covariance matrix), the random number generation process does not stabilize until sample sizes are reasonably large (above 500). In the random-number generation process, we need to generate a set of n-dimensional vectors whose n×n covariance matrix is the identify matrix *I*. The off-diagonal entries do not become consistently close to zero until sample size is reasonably large. Via the Cholesky transform (Kim 1986) or the singular value decomposition (Yancey et al. 2002), the matrix *I* can be transformed to $\varepsilon = c^2 \Sigma^{1/2} I\, \Sigma^{1/2}$. The array $\Sigma^{1/2}$ is the square root of the cov(**X**) from Cholesky or SVD procedures.

For files masked with additive noise, Kim (1986), Fuller (1993), Roque (2000), and Yancey et al. (2002) observed that regression can be performed that produces unbiased estimates of the coefficients of the independent variables. If *c* is kept small (below 0.1), then the variance increase associated with the noise addition will be quite small. The linear transform that takes $\mathbf{Y} = \mathbf{X} + \varepsilon$ to **Z** will approximately reduce the variances of **Z** to the variances of **X**. With additive noise, the hope is that one might be able to perform additional analyses. Fuller (1993) demonstrated that it is possible to perform slightly more sophisticated regressions in which pairs of independent variables are used provided that the constant *c* was kept small. Whereas analytic properties are better preserved by small *c*, he also noted re-identification might be possible if sophisticated methods were used.

To reduce re-identification risk while still preserving regression properties, Roque (2000) introduced mixtures of additive noise. Yancey et al. (2002) provided efficient computational procedures for adding mixtures of additive noise ε' where each component in the mixture is biased away from zero. As shown by Roque (2000), the mixtures can reduce re-identification rates by a factor of ten. Yancey et al. (2002) observed that the mixtures of additive noise did not yield stable covariances until domain size approached 1000.

In a general survey article, Brand (2002) observed that the mixture distribution associated with $\mathbf{Y} = \mathbf{X} + c\,\varepsilon$ or with **Z** could yield distributions that deviated substantially from the original distribution of **X**. In particular, if **X** is univariate normal and the noise ε is univariate normal, then the distribution of $\mathbf{Y} = \mathbf{X} + c\,\varepsilon$ $(0 < c < 0.1)$ differs substantially from the univariate normal distribution. Her work very strongly suggests that we should not expect any other analytic properties in data masked by additive noise beyond the properties demonstrated by Kim (1986) and Fuller (1993).

There are several observations we can make. If we are only preserving the first two moments needed for linear regression, then it is likely that no other statistical properties of the data will be preserved. Mera (1998) originally applied general software to transform data in a manner that preserved means and covariances. No other statistical properties were preserved. The controlled distortion procedure of Kim and Winkler (1995) was intended to move certain outliers that were on the boundaries of a multivariate point cloud into the interior of the distribution while still preserving means and covariances. The purpose was to cause the outlier to be difficult to identify while still preserving the regression relationships. The corresponding changes in a large number of records that were needed to preserve means and covariances were likely to cause other analytic properties of the transformed data to deteriorate. It is likely that the DuMouchel et al. (1999) methods of preserving moments will only allow a few analyses just as the synthetic-data-generation methods of Reiter (2002) create synthetic data that is only suitable for a one or two analyses.

3.3. Re-identifying Masked Microdata

The most straightforward way of re-identifying microdata is to directly match masked data **Z** with original data **X**. The statistical agency can extrapolate matching rates downward if it believes that an intruder would not have data of comparable quality and amount to data **X**. The advantage of this type of re-identification is that it is straightforward provided suitable record linkage software is available.

The other main method of re-identification is to use analytic techniques based on the understanding of the population **X** and the characteristics of the variables in **X** to do re-identification (Lambert 1993). There are two

observations that we can make based on the work of DuMouchel et al. (1999) and Moore and Lee (1998). First, for any model represented by likelihoods, we need to preserve a substantial number of aggregates that may be of the forms of moments of continuous variables or of closely approximated sums. If there are a greater number of analytic restraints on the data, then there will be a greater number of moments that need to be preserved (with continuous data). Second, DuMouchel et al. demonstrated that as the number of moments that need to be preserved increases, then the reduction factor decreases in going from the number N_1 of original records to the number N_2 of masked records. With only a moderate increase in the number of moments, the only data **Y** that will satisfy the restraints is the original data **X**. This is consistent with Fienberg (1997), Reiter (2002), and Mera (1998) who all showed that if there are a sufficient number of restraints on the synthetic microdata, then the some of the microdata records must be very close to the original, non-public microdata records.

A template for re-identification using analytic restraints is:

1. for each analysis, determine the specific aggregates and other restraints (such as positivity and additivity) that hold for a set of microdata **Z**,
2. determine the number of fields, the number of value-states of the fields, and the number of records, and
3. using the equations of step 1 and the unknowns of step 2, determine whether there are sufficient degrees of freedom on the masked data **Z** to assure that some of the data cannot be re-identified using analytic methods.

It is often very difficult to determine all of the restraints on data. If the number of degrees of freedom is twice as great as the number of restraints, then it seems possible that some information may be re-identified (using the tails of distributions).

4. Comments and Research Problems

The quality of original data **X** is seldom evaluated in terms of very specific aggregates (or sufficient statistics) that are needed for analyses. In this paper, we have assumed that the underlying data **X** are of high quality and the error induced in a public-use file is only due to masking. If an analysis must be performed on masked data **Z**, then certain aggregates that are part of the analysis must be produced with high accuracy. We can surmise that, if one or two sets of aggregates are produced with high accuracy, then it is quite possible – even likely – that other aggregates will not be accurate. Synthetic data generation (Reiter 2002) places substantial limitations on the analyses that can be performed. With general masked data **Z**, we can only assure that the analyses corresponding to aggregates such as a set of moments may be performed. These ideas lead to several research questions.

The first two research questions involve a set of moments (aggregates) with continuous data. How accurate must be the aggregates be for the analysis on the synthetic data **Z** to correspond reasonably to an analysis that might be performed on the corresponding general data **X**? If we have created accurate aggregates needed for one analysis, what other types of analyses might be possible with a given set of masked data? Statistical agencies are often unfamiliar with elementary forms of loglinear modeling. When are the sufficient statistics associated with a very elementary loglinear model suitably accurate to allow reproduction of an analysis from the original data **Z**?

5. Concluding Remarks

The only current masking methods that have been demonstrated to preserve one or two analytic properties in public-use microdata files are linearly transformed additive noise (Kim 1986), a multiple-imputation based method (Kennickell 1999), and certain types of synthetic data (Reiter 2002, Raghunathan et al. 2003, Dandekar et al. 2002). This paper provides some aggregates needed for many elementary models that must be preserved in the masked data. At present, the required accuracy in a set of aggregates needed for analyses of continuous data and in a set of sufficient statistics for reproducing loglinear models of discrete data is an open problem.

1/ This report is released to inform interested parties of ongoing research and encourage discussion. The views are those of the author and not necessarily those of the U.S. Census Bureau.

6. 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.

Bayardo, R. J., and Agrawal, R. (2005), "Data Privacy Through Optimal *K*-Anonymization," *IEEE 2005 International Conference on Data Engineering*.

Brand, R. (2002), "Microdata Protection Through Noise Addition," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, New York: Springer, 97-116.

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.

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.

DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C., and Pregibon, D. (1999), "Squashing Flat Files Flatter," *Proceedings of the ACM Knowledge Discovery and Data Mining Conference*, 6-15.

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.

Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, 9**,** 383-406.

Haworth, M., Bergdahl, M., Booleman, M., Jones, T., and Magaleno, M. (2001). "LEG chapter on Quality Framework," Proceedings of Q2001, Stockholm, Sweden, May 2001, CD-ROM.

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.

Kim, J. J. (1990), "Subdomain Estimation for the Masked Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 456-461.

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

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

LeFevre, K., DeWitt, D. and Ramakrishnan, R. (2005), "Incognito: Efficient Full-Domain K-Anonymity," *ACM SIGMOD Conference*, .

Mera, R. (1998), "Matrix Masking Methods That Preserve Moments," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 445-450.

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).

Moore, A. W., and Lee, M. S. (1998), "Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets," *Journal of Artificial Intelligence Research*, 8, 67-91.

Owen, A. (2003), "Data Squashing by Empirical Likelihood," *Data Mining and Knowledge Discovery*, 7 (1), 101-113.

Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, 19, 1-16.

Reiter, J.P. (2002), "Satisfying Disclosure Restrictions with Synthetic Data Sets," *Journal of Official Statistics*, 18, 531-543.

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

Roque, G. M. (2000), "Masking Microdata Files with Mixtures of Multivariate Normal Distributions," Ph.D.Dissertation, Department of Statistics, University of California at Riverside.

Samarati, P., and Sweeney, L. (1998), "Protecting Privacy when Disclosing Information: *k*-anonymity and Its Enforcement through Generalization and Cell Suppression," Technical Report, SRI International.

Sweeney, L. (2002), "Achieving *k*-Anonymity Privacy Protection Using Generalization and Suppression," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10 (5), 571-588.

Tendick, P., and Matloff, N. (1994), "A Modified Random Perturbation Method for Database Security," *ACM Transactions on Database Systems*, 19, 47-63.

Winkler, W. E. (2002), "Single Ranking Micro-aggregation and Re-identification," Statistical Research Division report RR 2002/08 at http://www.census.gov/srd/www/byyear.html.

Winkler, W. E. (2003), "Methods for Evaluating and Creating Data Quality," *Proceedings of the ICDT Workshop on Cooperative Information Systems*, Sienna, Italy, January 2003, longer version in *Information Systems* (2004), 29 (7), 531-550.

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.

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 Statistical Research Division report RR 2002/01 at http://www.census.gov/srd/www/byyear.html).