PRODUCING PUBLIC-USE MICRODATA THAT ARE ANALYTICALLY VALID AND CONFIDENTIAL

William E. Winkler*, Bureau of the Census, bwinkler@census.gov
William. E. Winkler, Bureau of the Census, Washington, DC 20233-9100

KEYWORDS: economic data, error localization

A public-use microdata file should be analytically valid. For a very small number of uses, the microdata should yield analytic results that are approximately the same as the original, confidential file that is not distributed. If the microdata file contains a moderate number of variables and is required to meet a single set of analytic needs of, say, university researchers, then many more records are likely to be re-identified via modern record linkage methods than via the re-identification methods typically used in the confidentiality literature. This paper compares several masking methods in terms of their ability to produce analytically valid, confidential microdata.

1. INTRODUCTION

With higher computing power, sophistication of software packages, and increased ability of users to develop their own software, researchers are better able to analyze microdata. These researchers (data users) are no longer content with using summary statistics produced by statistical agencies (data providers). The data users realize that, with access to appropriate microdata, they can examine issues and, indeed, find new issues that are beyond the purview and resources of the data providers. The data providers realize that they have a fundamental obligation to protect the confidentiality of data of individuals and enterprises. The data providers also realize that provision of analytically valid microdata to legitimate researchers has direct societal benefits due to improved analyses for policy purposes.

Agencies have responded by providing public-use files in which identifiers and information (variables) have been suppressed or changed in a variety of ways that the data providers (often statisticians) believe will assure confidentiality. The disclosure-limitation methods have ranged from simple suppression of names, addresses, and unique identifiers such as Social Security Number (SSN), to truncation of large values or other outliers, to data swapping (Dalenius and Reiss 1982), to suppression (DeWaal and Willenborg 1996), and finally to sophisticated methods of data masking (Kim 1986, Sullivan and Fuller 1990, Fuller 1993, Kim and Winkler 1995, Fienberg 1997). Rather than just provide publicly released microdata that have the same means and a few other properties of the confidential microdata, the sophisticated methods are intended to yield microdata that can be used for regression, loglinear modeling, or other statistical analysis even on a few important subdomains.

The ability of agencies to provide public-use microdata has been hampered by the agencies lack of resources to do the extensive extra work needed for producing such files and the view of some that their resources are better spent on their primary purpose of publishing summary statistics based on the data or letting individuals --typically sworn to abide by agency confidentiality restrictions— have direct access to microdata. Some agencies have not provided public-use data due to their belief that they cannot protect confidential data. This is particularly true with economic data. Another important consideration is the need for increased analytic and algorithmic coding skills among the computer programmers and analysts that must provide the data. Agencies have had difficulty developing the computer skills needed for sophisticated demographic, economic, and statistical analyses necessary for properly collecting, producing, and modeling their main data files. It is even more difficult doing sophisticated modeling and analyses to assure that public-use data produce similar results to what would be produced using the original, nonconfidential microdata and to perform time-consuming re-identification experiments.

Re-identification methods have predominantly involved detection of records that agree on simple combinations of keys based on discrete variables in the files (DeWaal and Willenborg 1995) or on outlier-detection techniques. When a specific combination of values of keys agree for a small set of records or for one record only, then either the specific values of some of the keys may be set to blank (*local suppression*) or different values of a key may be combined into single values (*global suppression*). These methods have the advantages that they are relatively easy for most data providers to understand and that they can be implemented in straightforward ways in computer code or via application of some statistical software. In a telling experiment, Bethlehem, Keller, and Pannekoek (1990) were able to use five quantitative income variables from the Internal Revenue Service (IRS) of the Netherlands to re-identify some individuals. They also showed how easily the records in a file could be partitioned using discrete variables such as geographic

identifiers, age, sex, demographic characteristics, and other information. The key point is that, if more information (variables that can be used as identifiers) is added to meet the needs of researchers and the files satisfy a number of analytic needs, then it is increasingly more difficult to insure confidentiality.

The methods and software of modern record linkage that can be used in re-identification experiments are very powerful. The basic methods were introduced by geneticist Newcombe (Newcombe et al 1959) who used odds ratios and decision rules. Statisticians Fellegi and Sunter (1969) provided the rigorous mathematical foundations and the means of estimating probabilities used in likelihood ratios. Implementation, however, was very slow because the means of researching and implementing record linkage have primarily involved difficult computer science and mathematical algorithms (Winkler 1994, 1995, Frakes and Baeza-Yates 1993) that are unfamiliar to most individuals at statistical agencies. Record linkage was primarily developed for unduplicating name and address lists having significant amounts of typographical variation due to transcription and keying error. Methods were extended to records having combinations of discrete and continuous variables (Winkler 1994, Scheuren and Winkler 1996) also having significant amounts of error. In other words, the normal situation in record linkage is that identifiers in pairs of records that are truly matches disagree by small or large amounts and that different combinations of the nonunique, error-filled identifiers need to be used in correctly matching different pairs of records. These modern record linkage methods are often in commercially available code that can be applied by relatively naive users in re-identification experiments. With the more sophisticated ways of producing public-use microdata (e.g., Kim 1986, Fuller 1993, Kim and Winkler 1995, DeWaal and Willenborg 1996), re-identification is considerably more difficult but possible if the individual performing the work is experienced in record linkage and able to write certain types of sophisticated computer code. At some point in the near future, it is likely that very powerful re-identification methods will be readily available in computer code. These re-identification methods (Scheuren and Winkler 1996) are primarily intended to provide a large number of analyses of sets of administrative files that have heretofore been impossible and to be performed by agencies that can keep data confidential by providing access to sworn agents at secure sites.

Three key ideas are needed to clarify the focus of the presentation in this paper. We say that a public-use file is *analytically valid* if a user is able to reproduce approximately several statistical analyses that can be produced with the original confidential microdata. We say that a file is *proven* analytically valid if the statistical agency has documented the modeling and analyses in sufficient detail so that data users are assured that the public-use files will produce analytic results that are somewhat consistent with the original confidential microdata. We say that a file is *analytically interesting* if it contains a sufficient number of variables, say five discrete demographic and six continuous economic, to provide (minimally) for the needs of serious researchers.

The overall structure of our presentation is to examine the different methods in terms of their ability to produce public-use files that are analytically valid and interesting and to examine whether they yield files that are confidential. In the second section, we provide motivation and background on the methods that have been used for creating confidential files and various re-identification methods that have been developed. The third section contains specific details about the empirical data, the analytic methods, and the re-identification methods. In the fourth section, we describe in detail a simulation experiment similar to one done by Fuller (1993), describe some additional masking methods that can be easily applied to the data, and give the results from several experiments regarding analytic validity and re-identification. We do not intend to reproduce exactly Fuller's results but to show how many re-identifications occur when we use a global comparison of one entire set of pairs and contrast it to the individual comparison used by Fuller (and typically others). The fifth section compares results via a variety of methods using the large, public-use data base originally analyzed by Kim and Winkler (1995). In our presentation, we examine how the different methods allow correct analyses in subdomains (Kim 1989) and certain followup or auxiliary analyses. Being able to perform followup analyses — while not the direct intent of the data providers — is of major concern to data users . The sixth section consists of discussion and the final section is a summary.

2. MOTIVATION AND BACKGROUND

Users are concerned with the analytic validity of the public-use files. To clarify the focus of analytic validity in the applications of this paper, we say that a file is analytically valid if it (approximately) preserves means and covariances on a small set of subdomains, preserves a few margins, and (crudely) preserves at least one other distributional characteristic. A file will be analytically interesting if it provides at least six variables on important subdomains that can be validly analyzed. In other applications, it may be useful to define analytic validity in terms of preserving some ordering characteristics of the variables, a few

geometric properties of the set of variables, or a large number of terms used in loglinear analyses. It should be intuitively obvious that it is impossible to provide a public-use file satisfying a large number of analytic needs on a large number of subdomains and also being confidential. We observe that it is very straightforward to get transformations that preserve means and covariances on a variety of subdomains. What is not as straightforward is preserving means, covariances, and other distributional characteristics. We note that merely preserving means on an entire public-use file is not sufficient for demonstrating that the file is analytically valid. Agencies have an additional concern related to the analytic validity of the files that they release. If a user were to publish an analysis based on statistics in a public-use file that are not similar to corresponding statistics in the original, unmasked file, then it is the agency that must take steps to correct any erroneous conclusions that would have been reached. Such correction efforts could require substantially greater resources than the resources needed for producing a public-use file that meets additional analytic needs.

Statistical agencies are concerned with their disclosure risk if an intruder were to attack a file. Following Lambert (1993), we define the *risk of true identification* as the fraction of released records that an intruder can correctly re-identify. We note that the risk of true identification is dependent on the amount of information in the publicly released file and the amount of high quality information that the intruder would be able to use in re-identification.

3. DATA AND METHODS

In this section we describe a variety of methods for producing confidential files using two different empirical data bases. The first file contains original records generated with eight variables satisfying a multivariate normal distribution with mean 0 and covariance matrix the identity matrix. The second file is a large public-use file associated with income variables of individuals that was constructed with demographic and other discrete variables. The basic file-production methods include masking with multivariate normal noise (Kim 1986, Fuller 1993), local and global suppression of information as performed in mu-Argus (DeWaal and Willenborg, 1995), and swapping (Kim and Winkler 1995) and various modified versions of the basic methods.

3.1. Generated Multivariate Normal

We generated variables having mulivariate normal distribution with mean 0 and covariance matrix the identity matrix I using the Statistical Analysis System (SAS). As in Fuller (1993), we generated multivariate normal noise independently with mean 0 and covariance matrix 0.35I in a procedure we refer to as masking 1. We also generated multivariate normal noise independently with mean 0, with covariance matrix 0.35I, and with small deviations deleted in a procedure we refer to as masking 2. An original data file of 1500 records was generated. The first 150 records were masked via the two additive-noise procedures, masking 1 and masking 2. To provide comparability with Fuller (1993), we matched the two masked files of 150 records against the first 150 records in the original file. To examine re-identification in more detail, we matched the second masked files of 150 against the entire set of 1500 original records.

3.2. Data of Kim and Winkler - Large Public-Use File

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

I) Total income;
ii) Adjusted gross income;
iii) Wage and salary income;
iv) Taxable interest income;
v) Dividend income;
vi) Rental income;
vii) Nontaxable interest income;
viii) Social security income;
ix) Return type;
x) Number of child exemptions;
xi) Number of total exemptions;
xii) Aged exemption flag;
xiii) Schedule D flag;
xiv) Schedule E flag;
xv) Schedule C flag; and
xvi) Schedule F flag.

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

The public-use file is important because it is used in examining tax policy and supplemental income payments. As such, we are interested in the ability of the file to provide for analyses in subdomains in which the data providers did not specifically assure that key

statistics are preserved. We note that it is theoretically impossible for the data provider to produce public-use data that yield a moderate number of accurate analyses in a moderate number of subdomains and maintain the confidentiality of the files.

3.3. Fellegi-Sunter Model of Record Linkage

A record linkage process attempts to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.*, 1959), Fellegi and Sunter (1969) considered ratios **R** of probabilities of the form

$$\mathbf{R} = \mathbf{Pr}(\gamma \in \Gamma \mid \mathbf{M}) / \mathbf{Pr}(\gamma \in \Gamma \mid \mathbf{U}) \qquad (1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur or deal with different types of comparisons of quantitative data. The fields compared (surname, first name, age) are called *matching variables*. The numerator in (1) agrees with the probability given by equation (2.13) in Fuller (1993).

The decision rule is given by

If $R > Upper$, then designate pair as a link.

If $Lower \le R \le Upper$, then designate pair as a possible link and hold for clerical review.

If $R < Lower$, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that this decision rule is optimal in the sense that for any pair of fixed bounds on **R**, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds, *Upper* and *Lower*, are determined by the error bounds. We call the ratio **R** or any monotonely increasing transformation of it (typically a logarithm) a *matching weight* or *total agreement weight*. Likely re-identifications, called matches, are given higher weights, and other pairs, called nonmatches, are given lower weights.

In practice, the numerator and denominator in (1) are not always easily estimated. The deviations of the estimated probabilities from the true probabilities can make applications of the decision rule suboptimal. Fellegi and Sunter (1969) were the first to observe that

$$\mathbf{Pr}(\gamma \in \Gamma) = \mathbf{Pr}(\gamma \in \Gamma \mid \mathbf{M})\,\mathbf{Pr}(\mathbf{M}) + \mathbf{Pr}(\gamma \in \Gamma \mid \mathbf{U})\,\mathbf{Pr}(\mathbf{U}) \qquad (2)$$

could be used in determining the numerator and denominator in (1) when the agreement pattern γ consists of simple agreements and disagreements of three variables and a conditional independence assumption is made. The left hand side is observed and the solution involves seven equations with seven unknowns. In general, we use the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) to estimate the probabilities on the right hand side of (2). To best separate the pairs into matches and nonmatches, our version of the EM algorithm for latent classes (Winkler 1994) determines the best set of matching parameters under certain model assumptions which are valid with the generated data and not seriously violated with the real data. In computing partial agreement probabilities for quantitative data, we make simple univariate adjustments to the matching weights such as are done in commercial record linkage software. Because we do not accurately account for the probability distribution with the generated multivariate normal data, our probabilities will not necessarily perform as well as the true probabilities used by Fuller when we consider single pairs. To force 1-1 matching as an efficient global approach to matching the entire original data sets with the entire masked data sets, we apply an assignment algorithm due to (Winkler 1994). When a few matching pairs in a set can be reasonably identified, many other pairs can be easily identified via the assignment algorithm. The assignment algorithm has the effect of drastically improving matching efficacy, particularly in re-identification experiments of the type given in this paper.

3.4. Additive Noise

Kim (1986) introduced independent additive noise with the same covariance as the original data X so that $Y = X + \epsilon$ is the resultant masked data. He showed that the covariance of Y is a multiple of the covariance of X and gave a transformation to another variable Z that is masked and has the same covariance as X. He also showed how regression coefficients could be computed and how estimates could be obtained on subdomains. His work has been extended by Sullivan and Fuller (1989, 1990) and Fuller (1993). In this paper, we will consider the basic additive noise $Y = X + \epsilon$ as was also considered by Fuller. Masking via additive noise has the key advantage that it can preserve means and covariances. As shown by Fuller (1993), arbitrary distributions can be transformed to normality, masked via additive noise, and then transformed back to the original scale. The two transformations also induce some bias. Additive noise has the disadvantage that files may not be as confidential as with some other the other masking

procedures. Kim has also shown that, if additive noise methods are used properly, then means and covariances from the original data can be reconstructed on all subdomains using the observed means and covariances from the masked data and a few additional parameters that the data provider must produce. Fuller (1993) has additionally shown that higher order moments such as the regression coefficients of interaction terms can be recovered provided that additional covariance information is available and specialized error-measurement software is applied.

3.5. Suppression

The suppression (or masking) methodology of mu-Argus is described by DeWaal and Willenborg (1995, 1996). In *global recoding (or global suppression)*, several categories of a variable are combined to form new categories. For instance a geographic code such as State abbreviation may have a subset of code values replaced by different code such as NorthEast U.S. In this way, the number of variables agreeing on the code (or variable) is increased. *Local suppression* sets certain values of individual variables to missing. The purpose of local suppression is to increase the set of records that agree on a combination of code (or key) values. DeWaal and Willenborg (1995) discuss a method in which the information-theoretic loss to local suppression can be minimized. The software mu-Argus (van Gemerden, Wessels, and Hundepol 1997) contains facilities to allow a user to determine combinations of key variables that place a record at risk of re-identification, give the user tools so that the user can quickly globally recode a file and analyze the results, and to locally suppress a file automatically. We note that the risk of re-identification used by mu-Argus is the risk when simple combinations of key variables are used in matching. The risk does not refer to re-identification via arbitrary means.

3.6. Swapping

Swapping is a method in which certain fields in a record are switched with the corresponding fields in another record. While it is a good way to assure confidentiality, it typically distorts distributions and key statistics severely (Little 1993). Kim and Winkler (1995) used a modified swapping procedure that was restricted so that means and covariances were preserved in certain subdomains. They applied their swapping procedure to a small percentage (<1%) of the records that the additive noise procedure could not effectively protect from disclosure. On specified subdomains, means and covariances could be preserved. On a few important subdomains, the means and covariances were often only slightly distorted because the percentage of swapping was very low. If we analyze variables in a subdomain with significantly different properties than other subdomains, then we need to be careful that the swapping does not seriously distort statistics in the subdomain. For instance, if we analyze a subdomain of individuals owning stock, then we want to assure that the swapping does not distort dividend and other stock-related income. When a small percentage of swapping is combined with additive noise, we will refer to it as the second hybrid additive-noise masking technique.

3.7. Fuller's Hybrid Masking Technique

Because quite a high proportion of the records could be easily re-identified with the additive noise procedure and simulated data of his main example, Fuller (1993) added two procedures to improve confidentiality protection. In the first, he only used noise vectors in a modified ϵ that had caused deviations in norm above a certain bound. This assures that fewer masked records are close to the corresponding unmasked records in norm. In a second procedure, Fuller adjusted the ϵ associated with the first and second best matches in situations where there was a high probability of re-identification. In our simulations, we also used Fuller's first adjustment for small deviations. It does not seriously affect covariances. The deviations over successive realizations of the random number generation process exceed the deviations caused by the adjustment from removing small deviations.

4. RESULTS FROM A SIMULATION

Table 1 is analogous to Table 1 in Fuller (1993). The first two columns of numbers are taken from Fuller's paper. The last three are produced via the procedures of this paper in which we generate multivariate normal data with zero mean and identity matrix for covariance. The probability (2.13) of Fuller (1993) is used for the first two columns of numbers and is optimal when matching single records in isolation. The results of the last three columns use estimated probabilities (crude general approximations) such as might be computed in commercial record linkage software and are quite suboptimal. The means of forcing 1-1 matching are what account for the dramatic improvement in the results exhibited in the last three columns of quantitative data. If the analyst were to model and use probabilities as in Fuller, then it is likely that the 4-variable-Winkler column would have almost as high match rates as the 6-variable-Winkler column. We note that the high correct match rates are consistent with Bethlehem et al (1990) who observed high accuracy when using five Internal Revenue Service of the Netherlands variables for matching.

Table 2 takes its first two columns from Table 3 of Fuller (1993). To mask variables further, Fuller removed small deviation noise and adjusted the noise associated the first and second best matches until the two match probabilities were approximately the same.

From examination of the two columns, Fuller (1993) concluded that the data were effectively masked. He also noted the correlations in the observed data differed by less than one standard deviation from the correlations in the unmasked data. Our examination of

Table 1. Distribution of our match probabilities for known vectors of different dimensions in a released data set of size 150. (Entries are percentages).

Match Probability	Dimension of known vector - Fuller - Four	Eight	-- Winkler -- Four ?	Six	Eight
0.0-0.1	49	9	39	3	0
0.1-0.2	23	8	0	0	0
0.2-0.3	9	6	0	0	0
0.3-0.4	4	6	0	0	0
0.4-0.5	3	6	0	0	0
0.5-0.6	2	8	0	0	0
0.6-0.7	1	9	0	0	0
0.7-0.8	0	8	0	0	0
0.8-0.9	4	7	0	0	0
0.9-0.99	5	22	0	97	0
0.99-1.0	0	11	61	0	100

the two columns of numbers produced by Fuller cause us to believe that the data are not effectively masked if additional record linkage procedures such as forcing 1-1 matching are used. The last four columns of Table 2 present our results from generating masked data in which no small deviation noise was used as in Fuller. Unlike Fuller, however, we did not adjust the match probabilities of the best two matches for each record. The primary reason that we did not is that the 1-1 matching procedure will easily overcome adjustments of the first few of the highest probability matches for a record. The secondary reason was that we were unsure exactly how Fuller adjusted the match probabilities to minimize the distortions in the correlations. The '?' indicate situations where I was not able to exactly compute matching probabilities because of the 1-1 matching. The most revealing results are in the next-to-last column of numbers in which we use six matching variables and match a file of 150 records against a file of 1500 records. Even in that situation, the 1-1 matching procedure yields a reasonably high correct match rate. With only small deviation noise removed, covariances were preserved up to a small multiplicative adjustment factor as used by Kim (1986). The deviations between the covariances in the masked data and the covariances in the unmasked were less than 0.1 of the standard deviation.

Table 2. Distribution of our match probabilities for known vectors of different dimensions in a modified masked released data set of size 150. (Entries are percentages).

Match Probability	Dimension of known vector - Fuller - Four	Eight	--- Winkler --- Four ?	Six	Six* ?	Eight
0.0-0.1	51	2	39	4	6	0
0.1-0.2	21	5	0	0	8	0
0.2-0.3	13	2	0	0	10	0
0.3-0.4	4	3	0	0	0	0
0.4-0.5	1	7	0	0	0	0
0.5-0.6	2	20	0	0	0	0
0.6-0.7	1	23	0	0	0	0
0.7-0.8	3	27	0	0	4	0
0.8-0.9	3	11	61	0	3	0
0.9-1.0	1	0	0	96	69	100

*/ Match against 1500 instead of 150.

We close this section by quoting two sentences from Fuller (1993, p. 393). "The analysis rested on the assumption that the intruder had information on a single target and used only this information in constructing a prediction." " The match probabilities are no longer valid if the intruder is able to use the information on a number of individuals to increase the probability of correctly matching a target to a released record. " Our results show that forcing 1-1 matching can significantly improve matching efficacy just as Fuller suggested might be possible. With the ready availability of credit files and other files and the possible availability of certain types of files containing health information, we can no longer assume that the knowledgeable intruder will look at records in isolation. The lack of control on privately held credit files and the ready access to them has been noted by Fellegi (1997).

5. RESULTS WITH A LARGE PUBLIC-USE FILE

In this section, we examine various additional masking methods using a large public-use file created by Kim and Winkler (1995). We begin by masking the file in two different ways suggested by the current version of mu-Argus software (van Gemerden, Wessels, and Hundepol 1997). We then proceed to a more detailed examination of matching and analytic results than the one produced by Kim and Winkler using procedures that are almost the same as Kim-Winkler and a version that we call enhanced mu-Argus.

5.1. Naive application of mu-Argus

We used a subset of the variables in the data base of 59315 records used by Kim and Winkler. The discrete variables are IRS form type, State code, age, race, and

sex. The continuous income variables are total income, adjusted gross income, wage, taxable interest, nontaxable income, rental income, social security income, dividends, and CPS wage.

We applied mu-Argus as a naive user might. We used mu-Argus on a file containing only the five discrete variables. It suggested collapsing on the age variable. We did this in two ways: (1) global recode of age to 999 and (2) global recode on age to ranges 1: 1-30, 2: 31-60, and 3: 61- followed by a pass to allow mu-Argus locally suppress (set to missing) certain values of variables. With each suppressed file, we were able to re-identify 59315 records when we used all five discrete and all nine continuous variables during matching. Because of the high re-identification rate, we did not examine analytic properties of the files. Due to the many local suppressions in the second type of recoding, it is likely that the analytic validity of the masked file is compromised.

As another naive application of mu-Argus, we recoded quantitative variables by rounding variables less than 80000 to the nearest 100 and variables greater than 80000 to the nearest 1000 and ran the resultant file containing 14 discrete variables through mu-Argus. We tried global suppressions on several variables but were unable to get the current beta version mu-Argus to produce files that looked to be modified correctly according mu-Argus methodology. We suspect that mu-Argus has difficulty with large numbers of variables, particularly when some variables have many value-states.

5.2. More advanced masking procedures

In this section, we compare results from using two procedures. Both begin with files in which additive noise has been used to mask the quantitative income variables according to the procedures of Kim (1986). In the first, we perform a swapping of quantitative data in a manner similar to Kim and Winkler (1995) but use software that gives more control of the swapping rates applied in different portions of the files. In the second, we use mu-Argus to suppress data. In this case we follow the suggestion that age be globally recoded (set to a fixed value). Since we did not have the resources to perform matching against several source files containing more than 100 million records, we make simplifying assumptions that allow us to compute absolute re-identification probabilities as is done in other papers. We begin by determining the probability of matching a record in the masked file of sample records against the original unmasked file of sampled records. Our assumptions allow us to compute the absolute probability of matching the masked sample file against an unmasked file of more than 100 million records. If a record has a total income less than 60000, we assume that the record has 1/1000 chance of being in a sample for a source file containing all records. If a record has a total income above 60000 and less than 80000, we assume that the record has 1/10 chance of being in a sample for a source file containing all records. If a record has a total income above 80000, we assume that the record has 1/1 chance of being in a sample for a source file containing all records. The assumptions are reasonable because (1) we are only using a subset of the variables that can be used for matching and (2) records having total incomes above 80000 are often associated with characteristics that make them outliers in the entire population, not just in the sample.

In Tables 3, 4, and 5, we describe re-identification rates from three matching passes. In the first, we match a file that has only been masked according to the additive noise procedure of Kim against the original unmasked file. Prior to the second pass, we swap all of the quantitative income data in records that total income above 80000 and a 0.05 proportion in records below 80000. We only swap in a subset of records that agree on keys consisting generally of IRS form type, age, race, sex, and State code. In situations where there are not sufficient number of items agreeing on a set of keys (less than 50 items), we collapse some of the combinations of keys. In the second matching pass, we match the masked/swapped file against the original unmasked file. Prior to the third matching pass, we use mu-Argus to determine a suppression strategy in which all ages are collapsed in a single age. A number of the resultant subsets in which matching is done (i.e., those agreeing on keys IRS form type, sex, race, and State code) have 1000 or more records. Because of the collapsing on age, no analyses involving age are possible in the masked file used in the third pass.

The results in Table 3 show that we can accurately match a high proportion of masked records having total income above 80000. Due to the facts that records having total income above 80000 have a few identifying characteristics somewhat different from other records having income above 80000 and that we have many matching variables, additive noise allows more than 1000 re-identifications. When higher levels of additive noise were used, Kim and Winkler (1995) observed a significant deterioration in the accuracy of correspondences of correlations of pairs of variables. The combination of swapping and additive-noise procedures used in creating the file used in the second pass have the advantage that easily re-identified records in the masked-only file are generally non-re-identifiable and that means and covariances are approximately preserved on the entire set of pairs and on important subdomains. We observe (Table 4) that the re-identification rate is effectively negligible in the file used in the second pass. On the other hand, the file

Table 3. Matching Counts and Truth Probabilities By Total Income Category Identification Pass, Masked File

Match Wgt	80k+ True	80k+ Fal	80k+ Prob	60k-80k True	60k-80k Fal	60k-80k Prob	60k- True	60k- Fal	60k- Prob
-5	0	1	0.00	0	15	0.00	0	578	0.00
-4	66	9	0.88	208	20	0.91	16E3	1901	0.90
-3	73	1	0.99	111	19	0.85	3095	694	0.82
-2	74	6	0.93	150	19	0.89	1780	766	0.70
-1	68	10	0.87	109	28	0.80	1500	1055	0.59
0	77	5	0.94	96	41	0.70	949	1072	0.47
1	71	5	0.93	68	41	0.62	605	976	0.38
2	79	7	0.92	96	41	0.70	594	1045	0.36
3	81	9	0.90	95	40	0.70	665	1213	0.35
4	91	8	0.92	91	49	0.65	693	1041	0.40
5	99	15	0.87	110	53	0.67	708	1115	0.39
6	109	11	0.91	125	64	0.66	744	1255	0.37
7	122	4	0.97	142	62	0.70	783	1309	0.37
8	149	9	0.94	131	54	0.71	846	930	0.48
9	181	12	0.94	155	58	0.73	836	649	0.56
10	195	6	0.97	153	53	0.74	886	478	0.65
11	213	7	0.97	187	36	0.84	847	297	0.74
12	221	5	0.98	159	11	0.94	609	110	0.85
13	222	6	0.97	171	8	0.96	496	66	0.88
14	223	0	1.00	112	4	0.97	292	24	0.92
15	147	0	1.00	50	1	0.98	106	5	0.95
16	67	0	1.00	3	0	1.00	8	0	1.00
17	24	0	1.00	2	0	1.00	0	0	.
18	8	0	1.00	0	0	.	0	0	.
19	1	0	1.00	0	0	.	0	0	.

Table 4. Matching Counts and Truth Probabilities By Total Income Category Re-Identification Pass, Masked/Swapped File

Match Wgt	80k+ True	80k+ Fal	80k+ Prob	60k-80k True	60k-80k Fal	60k-80k Prob	60k- True	60k- Fal	60k- Prob
-5	0	4	0.00	0	2	0.00	0	763	0.00
-4	11	8	0.58	16	15	0.52	2470	3252	0.43
-3	20	8	0.71	15	7	0.68	394	697	0.36
-2	22	11	0.67	23	16	0.59	244	903	0.21
-1	18	15	0.55	21	27	0.44	286	1642	0.15
0	25	23	0.52	20	35	0.36	197	1706	0.10
1	10	36	0.22	9	43	0.17	96	1274	0.07
2	4	62	0.06	8	61	0.12	106	1992	0.05
3	8	81	0.09	16	79	0.17	140	2966	0.05
4	8	96	0.08	17	108	0.14	160	2246	0.07
5	8	115	0.07	26	107	0.20	177	2484	0.07
6	8	130	0.06	31	149	0.17	240	3386	0.07
7	8	156	0.05	34	186	0.15	262	4993	0.05
8	13	178	0.07	47	216	0.18	338	4329	0.07
9	11	215	0.05	56	288	0.16	390	3185	0.11
10	9	251	0.03	64	323	0.17	434	3257	0.12
11	10	244	0.04	81	334	0.20	471	2527	0.16
12	5	242	0.02	74	232	0.24	329	1294	0.20
13	7	247	0.03	91	177	0.34	290	964	0.23
14	3	223	0.01	60	101	0.37	187	423	0.31
15	5	143	0.03	36	28	0.56	65	110	0.37
16	0	68	0.00	1	2	0.33	6	4	0.60
17	0	24	0.00	1	1	0.50	0	0	.
18	0	8	0.00	0	0	.	0	0	.
19	0	1	0.00	0	0	.	0	0	.

that has been masked via additive-noise and the mu-Argus procedure (Table 5) only allows a couple hundred re-identifications with probability above 0.25. Some individuals would argue that the third file is effectively masked.

Table 5. Matching Counts and Truth Probabilities By Total Income Category Enhanced mu-Argus Pass, Masked/Argus File

Match Wgt	80k+ True	80k+ Fal	80k+ Prob	60k-80k True	60k-80k Fal	60k-80k Prob	60k- True	60k- Fal	60k- Prob
-1	46	120	0.28	20	70	0.22	381	2426	0.14
0	58	111	0.34	18	66	0.21	252	2461	0.09
1	31	78	0.28	15	45	0.25	167	1478	0.10
2	50	84	0.37	19	57	0.25	160	2061	0.07
3	53	98	0.35	38	80	0.32	201	2924	0.06
4	39	109	0.26	37	95	0.28	243	2275	0.10
5	55	141	0.28	39	111	0.26	266	2493	0.10
6	51	140	0.27	43	144	0.23	327	3273	0.09
7	45	148	0.23	60	181	0.25	377	4652	0.07
8	50	137	0.27	58	167	0.26	457	3899	0.10
9	51	128	0.28	76	189	0.29	500	2852	0.15
10	51	138	0.27	76	309	0.20	593	2854	0.17
11	46	119	0.28	114	267	0.30	633	2243	0.22
12	30	69	0.30	107	186	0.37	453	1087	0.29
13	28	57	0.33	122	161	0.43	394	772	0.34
14	27	43	0.39	84	79	0.52	254	370	0.41
15	17	13	0.57	40	24	0.63	94	63	0.60
16	14	2	0.88	2	0	1.00	8	0	1.00
17	8	0	1.00	1	0	1.00	0	0	.
18	3	1	0.75	0	0	.	0	0	.

Use of the additive noise procedure of Kim (1989) allows us to recover means and correlations of important statistics. Swapping, on the other hand, can only assure that means and correlations are preserved in domains specified (controlled) by the individual doing the swapping. Table 6 illustrates that correlations are accurately preserved in a subdomain determined by Form Type. In the second-to-the-last column, 5% of all records are swapped as in Kim and Winkler (1995). In the last column, 5% of records with incomes below $80000 and all records with incomes above $80000 are swapped. The more complete set of swapping assures that the more easily identified large income individuals are not likely to be re-identified as is shown in Table 4. In Table 7, we show how correlations may not be preserved in the subdomain of records having some of their information taken from IRS Schedule C. Since we did not control record swapping in that subdomain and the individuals in the subdomain have characteristics that are distinctly different from the population as a whole, we see that certain key statistics are severely distorted. For instance, the swapping procedure severely distorts the correlation between wage and dividend. The reason is that the subdomain determined by IRS Schedule C corresponds to (partially) self-employed individuals having higher incomes and much higher dividend income than the entire population. In a similar manner, we see that, if we restrict to a subdomain consisting of a single State, then correlations may also be distorted (Table 8). Swapping was not controlled at the State level. The size of the subdomain associated with Table 8 is 600 while the sizes of the subdomains associated with Tables 6 and 7 are 5900 and 7800,

respectively.

Table 6. Correlations in a Subdomain Where Swapping is Controlled

	Raw	Masked Only	Masked & Swapped (5%)	Masked & Swapped (5%) Large
wage-divid	.027	.030	.030	.030
wage-tax int	.108	.100	.100	.100
divid-ss	.155	.162	.162	.162
tax int-rent	.172	.156	.156	.156
divid-rent	.040	.044	.044	.044
ntax-ss	.056	.056	.056	.056

Table 7. Correlations in a Subdomain Swapping is Not Controlled Form Type C

	Raw	Masked Only	Masked & Swapped (5%)	Masked & Swapped (5%) Large
wage-divid	.631	.634	.080	.060
wage-tax int	.190	.190	.188	.122
divid-ss	.153	.151	.125	.136
tax int-rent	.198	.199	.124	.121
divid-rent	.129	.127	.061	.052
ntax-ss	.106	.103	.086	.051

Table 8. Correlations in a Subdomain Swapping is Not Controlled State Code = 46

	Raw	Masked Only	Masked & Swapped (5%)	Masked & Swapped (5%) Large
wage-divid	.057	.061	.061	.074
wage-tax int	-.088	-.082	-.082	-.012
divid-ss	.144	.150	.149	.088
tax int-rent	.181	.154	.151	.130
divid-rent	.033	.033	.033	.029
ntax-ss	.139	.130	.125	.172

6. DISCUSSION

The reason that we prefer additive noise as the starting point for a masking methodology is that authors (Kim 1986, Sullivan and Fuller 1989, Kim 1990, Sullivan and Fuller 1990, and Fuller 1993) have taken care to demonstrate that it provides a few recoverable analytic properties on subdomains. As the analysis of Kim and Winkler (1995) and this paper show, moderate amounts of additive noise do not yield files that are completely free of disclosures. Both Fuller(1993) and Kim and Winkler (1995) have observed that large amounts of additive noise destroy the analytic validity of files. The empirical results of Fuller(1993), Kim and Winkler (1995), and this paper strongly suggest that only a very few analytic properties of the original files may be recoverable at the costs of using specialized software and much larger variances for higher order statistics.

7. SUMMARY

This paper examines a variety of methods for masking files that are intended to provide analytically valid public-use files in which disclosures are limited. It corroborates that the additive-noise methods of Kim (1986) and Fuller (1993) can produce masked files that allow a few analyses that approximately reproduce a few analyses on the original, unmasked data. It also shows that, if additional masking procedures such a probability adjustment (Fuller 1993) and very limited swapping (Kim and Winkler 1995) are applied, then disclosure risk is significantly reduced and analytic properties are somewhat compromised.

*The views expressed in this paper are those of the author and do not necessarily represent those of the U.S. Bureau of the Census.

REFERENCES

Bethlehem, J. A., Keller, W. J., and Pannekoek, J., (1990) "Disclosure Control of Microdata," *Journal of the American Statistical Association*, **85**, 38-45.

Blien, U., Wirth, U., and Muller, M. (1992),"Disclosure Risk for Microdata Stemming from Official Statistics," *Statistica Neerlandica*, **46**, 69-82.

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

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

De Waal, A. G., and Willenborg, L.C.R.J. (1995), "Global Recodings and Local Suppressions in Microdata Sets," Proceedings of Statistics Canada 95, 121-132

DeWaal, A. G., and Willenborg, L.C.R.J. (1996), "A View of Statistical Disclosure Control for Microdata," *Survey Methodology*, **22**, 95-103.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American*

Statistical Association, **64**, 1183-1210.
Fellegi, I. P. (1997), "Record Linkage and Public Policy - A Dynamic Evolution," *Proceedings of the Record Linkage Workshop 1995, National Academy of Sciences*, to appear.
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.
Frakes, W. and Baeza-Yates, R. (1992), "Information Retrieval - Data Structures and Algorithms," Prentice-Hall: Upper Saddle River, N.J.
Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, **9**, 383-406.
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*, 303-308.
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.
Lambert, D. (1993), "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*, **9**, 313-331.
Little, R. J. A. (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, **9**, 407-426.
Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science,* **130**, 954-959.
Paas, G. (1988), "Disclosure Risk and Disclosure Avoidance for Microdata," *Journal of Business and Economic Statistics*, **6**, 487-500.
Scheuren, F., and Winkler, W. E. (1996), "Recursive Merging and Analysis of Administration Lists," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 123-128 (presently available on http://www.amstat.org in the Section on Government Statistics).
Sullivan, G., and Fuller, W. A. (1989), "The Use of Measurement Error to Avoid Disclosure," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 802-807.
Sullivan, G., and Fuller, W. A. (1990), "Construction of Masking Error for Categorical Variables," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 435-439.
Van Gewerden, L., Wessels, A., and Hundepol, A. (1997), "Mu-Argus Users Manual, Version 2," Statistics Netherlands, Document TM-1/D.
Winkler, W. E. (1994), "Advanced Methods for Record Linkage, American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 467-472.
Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.