

RESEARCH REPORT SERIES
(*Statistics #2005-09*)

**Re-identification Methods for Evaluating the
Confidentiality of Analytically Valid Microdata**

William E. Winkler

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

Report Issued: October 3, 2005

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata

William E. Winkler*, *U.S. Bureau of the Census*
Statistical Research, Room 3000-4, Washington, DC 20233-9100

KEYWORDS: re-identification, additive noise

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 [4]), to suppression (DeWaal and Willenborg [5]), and finally to sophisticated methods of data masking (Kim [12], Sullivan and Fuller [21], Fuller [11], Kim and Winkler [14], Fienberg [9]). 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, confidential 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 [5]) 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 recoding*). 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 statistical software. In a telling experiment, Bethlehem, Keller, and Pannekoek [1] 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. Newcombe (Newcombe *et al.* [17]) introduced the basic methods introduced. Geneticist Newcombe used odds ratios and decision rules. Statisticians Fellegi and Sunter [7] 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 [23], [24]; Frakes and Baeza-Yates [10]) 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 [23], Scheuren and Winkler [19]) 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 [12], Fuller [11], Kim and Winkler [14], DeWaal and Willenborg [6]), 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 [19]) 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 [11], 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 database originally analyzed by Kim and Winkler [14]. In our presentation, we examine how the different methods allow correct analyses in subdomains (Kim [13]). Being able to perform followup analyses C while not the direct intent of the data providers C 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 [15], we define the *risk of true identification* as the fraction of released records that an intruder can correctly re-identify.

3 Data and Methods

In this section we describe a variety of methods for producing confidential files using two different empirical databases. 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 [12], Fuller [11]), local and global suppression of information as performed in μ -Argus (DeWaal and Willenborg [5]), and swapping (Kim and Winkler [14]) and various modified versions of the basic methods.

3.1 Generated Multivariate Normal

We generated variables having multivariate normal distribution with mean 0 and covariance matrix the identity matrix I using the Statistical Analysis System (SAS). As in Fuller [11], 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 [11], 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 file 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; | ix) Return type; |
| ii) Adjusted gross income; | x) Number of child exemptions; |
| iii) Wage and salary income; | xi) Number of total exemptions; |
| iv) Taxable interest income; | xii) Aged exemption flag; |
| v) Dividend income; | xiii) Schedule D flag; |
| vi) Rental income; | xiv) Schedule E flag; |
| vii) Nontaxable interest income; | xv) Schedule C flag; and |
| viii) Social security income; | 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 a record may be a type of outlier in the population. For re-identification, we primarily need be concerned with higher income individuals or those with distinct characteristics that might be easily identified even when sampling rates are low.

The public-use file is used in examining tax policy and supplemental income payments. The file must allow 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.* [17]), Fellegi and Sunter [7] considered ratios \mathbf{R} of probabilities of the form

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

where $\boldsymbol{\gamma}$ 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 $\boldsymbol{\gamma} \in \Gamma$ 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 (2.11) in Fuller [11].

The decision rule is given by

If $\mathbf{R} > Upper$, then designate pair as a link.

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

If $\mathbf{R} < Lower$, then designate pair as a nonlink.

Fellegi and Sunter [7] showed that this decision rule is optimal in the sense that for any pair of fixed bounds on \mathbf{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 \mathbf{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 [7] were the first to observe that

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

could be used in determining the numerator and denominator in (1) when the agreement pattern $\boldsymbol{\gamma}$ 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 [3]) 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 [23]) determines the best set of matching parameters under certain model assumptions which are valid with the generated data and not seriously violated with the real data. In computing partial agreement probabilities for quantitative data, we make simple univariate adjustments to the matching weights such as are done in commercial record linkage software. When two quantitative items a and b do not agree exactly, we use a linear downward adjustment from the agreement matching weight to the disagreement weight according to a tolerance. Specifically, the adjustment is

$$w_{adj} = \max(\{w_{adj} - (w_{agr} - w_{dis}) \mid a - b \mid / (\tau \min(a, b))\}, w_{dis}),$$

where w_{adj} , w_{agr} , w_{dis} are the adjusted weight, full agreement weight, and full disagreement weights, respectively

and τ is the proportional tolerance for the deviation ($0 \leq \tau \leq 1$). The full agreement weights w_{agr} and disagreement weights w_{dis} are the natural logarithms of (1) that are obtained via the EM algorithm. The tolerance τ is estimated using experience and looking at matching results. For the empirical examples of this paper τ is taken to be 0.2. For re-identification experiments in which noise levels are relatively lower than those in this paper, τ might be set to 0.1. The approximation will not generally yield accurate match probabilities but works well in the matching decision rules as we show later in this paper. Because we do not accurately account for the probability distribution with the generated multivariate normal data, our probabilities will not necessarily perform as well as the true probabilities used by Fuller when we consider single pairs. 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 [23]). Specifically, we use pairs $(i,j) \in I_0$ where I_0 minimizes

$$\{\sum_{(i,j) \in I} w_{ij} \mid I \in \mathbf{J}\},$$

where $w_{i,j}$ is the comparison weight for record pair (i,j) , and \mathbf{J} is the set of index sets I in which at most one column and at most one row are present. That is, if $(i,j) \in I$ and $(k,l) \in I$, then either $i \neq k$ or $j \neq l$. The algorithm of Winkler is similar to the classic algorithm of Burkard and Derigs (see e.g., Winkler [23]) in that it uses Dijkstra's shortest augmenting path for many computations and has equivalent computational speed. It differs because it contains compression/decompression routines that can reduce storage requirements for the array of weights $w_{i,j}$ by a factor of 500 in some matching situations. When a few matching pairs in a set can be reasonably identified, many other pairs can be easily identified via the assignment algorithm. The assignment algorithm has the effect of drastically improving matching efficacy, particularly in re-identification experiments of the type given in this paper. For instance, if a moderate number of pairs associated with true re-identifications have probability greater than 0.5 when looked at in isolation, the assignment algorithm effectively sets their match probabilities to 1.0 because there are no other suitable records with which the truly matching record should be combined.

3.4 Additive Noise

Kim [12] 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. Sullivan and Fuller ([20], [21]) and Fuller [11] extended Kim's work. 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. Additive noise has the disadvantage that files may not be as confidential as with some of the other masking procedures. Kim has shown that means and covariances from the original data can be reconstructed on all subdomains using the observed means and covariances from the masked data and a few additional parameters that the data provider must produce. Fuller [11] has shown that higher order moments such as the regression coefficients of interaction terms can be recovered provided that additional covariance information is available.

3.5 Suppression

DeWaal and Willenborg ([5], [6]) describe the suppression (or masking) methodology of μ -Argus. In *global recoding*, 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. The software μ -Argus (van Gernerden, Wessels, and Hundepool [22]) 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 μ -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 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 [11] 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.

3.7 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 [16]). Kim and Winkler [14] 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 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. We refer to the combination of a small percentage of swapping with additive noise as the *second hybrid additive-noise masking technique*.

4 Results from a Simulation

Table 1 is analogous to Table 1 in Fuller [11]. 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 (2.11) of Fuller [11] 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. For instance when Fuller used eight variables, 65% of the match probabilities were above 0.5. The assignment algorithm effectively changes the match probabilities to 1.0 for the 65%. Because so many potential false matching pairs are in the 65% and effectively eliminated for the remaining 35%, the ability to correctly re-identify in the remaining 35% increases to certainty as shown in the table. 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.

Table 2 takes its first two columns from Table 3 of Fuller [11]. To mask variables further, Fuller removed small deviation noise and adjusted the noise associated with the first and second best matches until the two match probabilities were approximately the same. From examination of the two columns, Fuller [11] 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 the two columns of numbers produced by Fuller cause us to believe that the data are not effectively masked if 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 \geq 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. For the 64% of the records that were associated with truly matching pairs with probabilities above 0.5 when looked at in isolation, the assignment algorithm effectively sets their match probabilities to 1.0. The remaining 36% of the records are associated with truly matching records that are generally farther away than the closest records that are not true matches. With only small deviation noise removed, covariances were preserved up to a small multiplicative adjustment factor as used by Kim [12]. 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 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 -		-- Winkler --		
	Four	Eight	Four	Six	Eight
0.0-0.1	49	9	42	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	0	0
0.99-1.0	0	11	58	97	100

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 -		--- Winkler ---			
	Four	Eight	Four	Six	Six*	Eight
					?	
0.0-0.1	51	2	42	4	12	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	6	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	0	0
0.8-0.9	3	11	0	0	0	0
0.9-1.0	1	0	58	96	64	100

*/ Match against 1500 instead of 150.

We close this section by quoting two sentences from Fuller ([11], 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. Fellegi [8] has already noted the lack of control on privately held credit files and the ready access to them.

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 [14]. We begin by masking the file in two different ways suggested by the current version of μ -Argus software (van Gernerden, Wessels, and Hundepool [22]). 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.

5.1 Naïve application of μ -Argus

We used a subset of the variables in the database of 59315 records used by Kim and Winkler. The discrete variables are IRS form type, State code, age, race, and sex. The continuous variables are total income, adjusted gross income, wage, taxable interest, nontaxable income, rental income, social security income, dividends, and CPS wage.

We applied μ -Argus as a naive user might. We used μ -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 μ -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 many local suppressions in the second type of recoding, it is likely that the analytic validity of the masked file is compromised.

5.2 More advanced masking procedures

In this section, we compare results from using two procedures. The first uses a file in which additive noise has been used to mask the quantitative income variables according to the procedures of Kim [12]. In the second, we perform a swapping of quantitative data in a manner similar to Kim and Winkler [14] but use software that gives more control of the swapping rates applied in different portions of the files. 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 and 4, we describe re-identification rates from two 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 having total income above 80000 and a 0.05 proportion of records having total income 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 a 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.

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 [14] 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 has 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.

Use of the additive noise procedure of Kim [13] 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 5 illustrates that correlations are accurately

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

Match	80k+			60k-80k			60k-		
	Wgt	True	Fal Prob	True	Fal Prob	True	Fal	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	80k+			60k-80k			60k-		
	Wgt	True	Fal	Prob	True	Fal	Prob	True	Fal
-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	.

Table 5.
 Correlations in a Subdomain
 Where Swapping is Controlled

	Raw	Masked	Masked & Swapped	
		Only	(5%)	(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

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 [14]. 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 6, 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

Table 6.
Correlations in a Subdomain
Swapping is Not Controlled
Form Type C

	Raw	Masked Only	Masked & Swapped (5%)	(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

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 7). Swapping was not controlled at the State level. The size of the subdomain associated with Table 7 is 600 while the sizes of the subdomains associated with Tables 5 and 6 are 5900 and 7800, respectively.

Table 7.
Correlations in a Subdomain
Swapping is Not Controlled
State Code = 46

	Raw	Masked Only	Masked & Swapped (5%)	(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 [12], Sullivan and Fuller [20], Kim [13], Sullivan and Fuller [21], and Fuller [11]) have taken care to demonstrate that it provides a few recoverable analytic properties on subdomains. As the analysis of Kim and Winkler [14] and this paper show, moderate amounts of additive noise do not yield files that are completely free of disclosures. Both Fuller [11] and Kim and Winkler [14] have observed that large amounts of additive noise destroy the analytic validity of files. The empirical results of Fuller [11], Kim and Winkler [14], 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. Although the results of section 5 are for files in which more than twenty variables were used, the analytic variations observed on certain subdomains and re-identification probabilities would not have changed if as few as ten variables were used. These ten variables are age, race, sex, State code, and any six quantitative variables. With only seven, eight or nine variables, obtaining re-identification probabilities as high as the probabilities in this paper is dependent on the specific variables and the specific subdomains that are used. With six or fewer variables, it should be possible to produce files that are both analytically valid and allow a negligible percentage of re-identifications in most situations.

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 [12] and Fuller [11] 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 [11]) and very limited swapping (Kim and Winkler [14]) are applied, then disclosure risk is significantly reduced and analytic properties are somewhat compromised.

*This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion. The author thanks an anonymous referee for comments on an earlier version of this paper. This paper originally appeared in *Research in Official Statistics*, 1, 87-104 in 1998.

References

- [1] J. A. Bethlehem, W. J. Keller, and J. Pannekoek, "Disclosure Control of Microdata," *Journal of the American Statistical Association*, **85** (1990) 38-45.
- [2] U. Blien, U. Wirth, and M. Muller, "Disclosure Risk for Microdata Stemming from Official Statistics," *Statistica Neerlandica*, **46** (1992) 69-82.
- [3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, **B**, **39** (1977) 1-38.
- [4] T. Dalenius, and S. P. Reiss, "Data-swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference*, **6** (1982) 73-85.
- [5] A. G. DeWaal, and L.C.R.J. Willenborg, "Global Recodings and Local Suppressions in Microdata Sets," *Proceedings of Statistics Canada Symposium 95, 1995*, pp. 121-132.
- [6] A. G. DeWaal, and L.C.R.J. Willenborg, "A View of Statistical Disclosure Control for Microdata," *Survey Methodology*, **22** (1996) 95-103.
- [7] I. P. Fellegi, and A. B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64** (1969) 1183-1210.
- [8] I. P. Fellegi, "Record Linkage and Public Policy - A Dynamic Evolution," in W. Alvey and B. Jamerson (eds.), *Record Linkage Techniques 1997*, Federal Committee on Statistical Methodology, Washington DC, USA, 1997, pp. 3-12.
- [9] S. E. Fienberg, (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.
- [10] W. Frakes, and R. Baeza-Yates, "Information Retrieval - Data Structures and Algorithms," Prentice-Hall: Upper Saddle River, N.J., 1992.
- [11] W. A. Fuller, "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, **9** (1993) 383-406.
- [12] J. J. Kim, "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation,"

- American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 1986, pp. 303-308.
- [13] J. J. Kim, "Subdomain Estimation for the Masked Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 1990, pp. 456-461.
 - [14] J. J. Kim, and W. E. Winkler, "Masking Microdata Files," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 1995, pp. 114-119.
 - [15] D. Lambert, "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*, **9** (1993) 313-331.
 - [16] R. J. A. Little, "Statistical Analysis of Masked Data," *Journal of Official Statistics*, **9** (1993) 407-426.
 - [17] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James, "Automatic Linkage of Vital Records," *Science*, **130** (1959) 954-959.
 - [18] G. Paas, "Disclosure Risk and Disclosure Avoidance for Microdata," *Journal of Business and Economic Statistics*, **6** (1988) 487-500.
 - [19] F. Scheuren, and W. E. Winkler, "Regression Analysis of Data Files that are Computer Matched – Part II," *Survey Methodology*, 1997, pp. 157-165).
 - [20] G. Sullivan, and W. A. Fuller, "The Use of Measurement Error to Avoid Disclosure," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 1989, pp. 802-807.
 - [21] G. Sullivan, and W. A. Fuller, "Construction of Masking Error for Categorical Variables," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 1990, pp. 435-439.
 - [22] L. van Gemerden, A. Wessels, and A. Hundepool, "µ-Argus Users Manual, Version 2," Statistics Netherlands, Document TM- 1/D, 1997.
 - [23] W. E. Winkler, "Advanced Methods for Record Linkage," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 1994, pp. 467-472.
 - [24] W. E. Winkler, "Matching and Record Linkage," in B. G. Cox (ed.) *Business Survey Methods*, New York: J. Wiley, 1995, pp. 355-384.