## Table of Contents

## EXECUTIVE SUMMARY

In 1995, the Department of Health and Human Services (HHS) contracted with the US Bureau of the Census to produce a specially requested public use microdata file. This file supplemented information on the March 1991 Current Population Survey (CPS) public use file with Form 1040 information from each respondent's 1990 Internal Revenue Service (IRS) tax return. Because the IRS could use their data to re-identify individuals on the CPS file, great care had to be taken to mask the HHS file. In doing so, Jay Kim and William Winkler (1995) developed a very effective two-stage procedure for masking public use microdata files.

The initial step of this procedure adds randomly generated multivariate noise to the IRS data (Kim, 1986). Records are perturbed so that the means of the universe or any of its sub-domains are not biased. However, the noise is generated so that it increases the variances and covariances by a factor of $(1 + c^2)$ --- where the masking agent specifies the value of c. For relatively low values of c, this technique masks the data well without significantly diminishing its analytic utility.

Although Kim's method distorted many records beyond recognition, Winkler was able to use his matching software (Winkler, 1994, 1995a, 1995b) to re-identify an unacceptably large proportion. He then developed a supplemental routine to further mask the file. By modifying his software to swap a pre-specified percentage of re-identifiable records, Winkler could control the extent to which re-identification was possible.

Unlike Kim's method, data swapping will bias the means and variance-covariance relationships of sub-domains. Winkler's routine seeks to minimize the distortion. First, Winkler only swaps a portion of the re-identifiable records. Next, his software matches records within blocks (prescribed sets of categorical attributes). Statistical moments for these blocks are not disturbed, although moments for subs-domains within blocks may be altered. The Winkler approach also attempts to minimize this distortion. The routine swaps so that values of re-identifiable records are exchanged with those on the most similar (i.e., second best matching ) record within the same block. Both of Winkler's enhancements should greatly reduce swapping bias introduced into the sub-domain statistics.

The Census Bureau recognizes that the Kim-Winkler approach is a very powerful tool for current disclosure limitation. To employ the software effectively, the user must exhibit some expertise in the setting of certain parameters. This paper concentrates on (1) the development of a standard procedure for determining an acceptable amount of noise (i.e., value of c) for Kim's routine, and (2) the need for the Census Bureau to adopt a standard for the maximum percentage of re-identifiable "high risk" records in a publicly-released file.

In this paper, we will show that there is a critical value for c, above which Kim's method provides little additional masking. We show that this critical value is a function of the variance, top-, and bottom-codes of the distorted continuous variables. The value is also directly proportional to the maximum percentage within which an intruder would expect the values of each record in the publicly released file to differ from the corresponding record in his target file.

Developing a standard for the maximum percentage of re-identifiable "high risk" records is a more difficult problem. This paper recommends that the Census Bureau rely purely on historical work. In particular, the paper references the work of Paas and Wauschkuhn (1985); Fienberg, Makov, and Sanil (1995); and Muller, Blien, and Wirth (1995). In all three papers, an "offspring" file is created by perturbing each record in the original file. Records from this file are then matched against some of their original counterparts. Under such conditions, the research indicates that a 10 to 20 percent re-identification rate is not excessively high.

We are very enthusiastic about the potential of the Kim-Winkler approach. It incorporates the best of Kim's technique (namely, masking without biasing the means and minimally biasing the variance-covariance structure) supplemented by the major positive of Winkler's swapping technique (only the necessary number of records are swapped to ensure confidentiality). We plan continued research into investigating potential uses of the method. We hope that the Census Bureau shares our enthusiasm for this approach and will incorporate it into our arsenal of disclosure limitation techniques for microdata.

# Analysis of the Kim-Winkler Algorithm For Masking Microdata Files -- How Much Masking Is Necessary and Sufficient? Conjectures For the Development of a Controllable Algorithm

Richard A. Moore, Jr.
Statistical Research Division
US Bureau of the Census
Washington, DC 20233

## Abstract

In 1995, Jay Kim used the addition of randomly generated noise in an attempt to mask a public use microdata file. William Winkler then used extremely sophisticated matching software to attempt to re-identify records in the perturbed file. To his amazement, Winkler found that a substantial portion of the records in the "masked" file could be correctly matched to their counterparts in the original file. He then enhanced the matching software so that a portion of the re-identifiable records were swapped with "similar" records in the perturbed file.

This paper first examines Kim's and Winkler's approach separately, then examines the effect of using both in succession. It shows that the Kim-Winkler approach is a feasible masking method. The paper also explores the possibility of developing an optimal set of parameters, which generate sufficient masking with minimal distortion to the variance-covariance relationships.

## I. Introduction

In 1995, the Bureau of the Census supplemented 59,315 records from the March 1991 Current Population Survey (CPS) with income fields for each household. It obtained values for these incomes from administrative records provided by the Internal Revenue Service (IRS). The resulting microdata file was then to be released to the Department of Health and Human Services (HHS), so that it might set policy for earned income and other benefits. Having access to the actual microdata (and not just summary statistics), HHS would be able to make better decisions.

Unfortunately, this project caused very serious concerns about the confidentiality protection afforded to the respondents to the March 1991 CPS. The Census Bureau is restricted by Title 13. This law prohibits it from releasing any information by which the true identity of any individual can be determined. It was feared that the IRS might use the their data to identify particular individuals. Kim and Winkler (1995) took a two-stage approach to satisfying the confidentiality problem.

In the first stage, Kim (1986) added random noise to the file. In the second stage, Winkler (1995a, 1995b) used his matching software to identify records, which were not sufficiently protected by the previous procedure. Winkler then enhanced his matching software, so that a pre-specified percentage of re-identifications were swapped. This paper analyzes their approach.

Sections II and III describe the objective and the measurement techniques.

Sections IV through VIII evaluate Kim's approach. Sections IV and V describe the technique and the test deck. Section VI evaluates the results of the test. Section VII offers some conjectures about the results. ("Semi-proofs" of the conjectures are provided in Appendix A.) Section VIII summarizes the advantages and disadvantages of Kim's technique.

Sections IX through XII examine the Winkler swapping routine. Section IX gives an overview of the software and the technique. Section X discusses its advantages and disadvantages. Section XI chronicles conjectures for a tolerable number of re-identifications, while Section XII illustrates how to use the software to achieve this tolerance.

Section XIII provides one final conjecture on an efficient strategy for using the two techniques in tandem. Section XIV lists some topics for future research. A brief summary is provided in Section XV, while references are listed in Section XVI.

The paper also contains two appendices. Appendix A provides some technical insight into the validity of the conjectures made in Section VII, while Appendix B furnishes some brief documentation on the battery of programs used to implement the Kim-Winkler procedure.

## II. Purpose of This Study

The purpose of this research is to analyze the masking techniques of Kim and Winkler, first separately, then when the two are used in succession. Each of these procedures involve the setting of various parameters. The study has the following objectives.

(1) Explore the relationship between the value of each parameter and the amount of protection which it provides.

(2) Demonstrate that the analytical utility of the file is inversely related to the number of records which are masked.

(3) Suggest an approach which would estimate the optimal values for these parameters (i.e., those which provide the necessary protection with the least distortion).

(4) Initiate work into the development of a standard procedure to determine whether a file has been sufficiently masked for public release. This may involve research into the development of intruder model software.

The illustrations provided are based on a small-scale numerical experiment on a single data set. They serve to illustrate prospective techniques by which masking procedures may be able to be evaluated. These are only the initial steps. Before such a program can be implemented, more research will be required.

## III. The Technique for the Measurement of the Amount of Masking Provided

To measure the amount of masking provided by any noise addition technique, use the procedure of Paas and Wauschkuhn (1985). Generate an offspring file from an existing microdata file by implanting random errors in each value of some continuous variables of an offspring file. Then use discrimination techniques in an attempt to correctly match each offspring record to its unperturbed parent.

This procedure can be generalized. First, construct a target file of observations with fairly unique categorical combinations, referred to as key (or block) combinations. If the continuous (or matching) variable, in conjunction with other key variables, form relatively unique combinations, then re-identification is possible. Next, to test the ability of a technique to mask a file, use the following procedure.

(1) Mask the file.

(2) By using matching software, link each record in the target file to a perturbed record of the offspring file. Allow the values of each matching variable (on the offspring file record) to differ from the corresponding values of some record of the perturbed file by less than some pre-defined percentage.

(3) When all matching fields fall within a tolerable percentage, one has a re-identification. Link these two records. If the target record is in fact linked with the correct offspring, refer to it as a correct re-identification. If it is linked to the offspring of a different record, refer to it as an incorrect re-identification.

(4) This provides two primitive measures for the amount of masking provided. First, look at the number of correctly re-identified records. Next, examine the ratio of correct to incorrect re-identifications.

## IV. Masking by the Addition of Random Noise

Kim (1986) developed a method to perturb continuous data, while preserving the means and "essentially" preserving the variance-covariance structure of the original microdata file. The method involved the random generation of n-dimensional noise vectors, centered at the origin, and generated from a scaled version of the variance-covariance structure of the original distribution. These noise vectors, when added to the original observations, perturb the data so that

$$(1)\quad Cov(\underline{a}', \underline{b}') = (1+c^2) * Cov(\underline{a}, \underline{b});$$

$$(2)\quad Var(\underline{a}') = (1+c^2) * Var(\underline{a}).$$

Here, Cov(a', b') is the bivariate covariance between the perturbed values of Fields a and b, Cov(a, b) is the original covariance, Var (a) is the original variance of Field a, Var(a') is the variance after the addition of noise, and c is the scaling factor. Thus, c is the parameter which drives the entire process. By increasing the value of c, one presumes that the additional noise better masks the data. However, the more one disturbs the fields, the more the method disperses the data. For these reasons, take care when setting the value of c. A value too low does not provide enough protection, one too high seriously distorts the variance-covariance structure.

## V. The Test Deck and the Generation of Random Noise

Initially, let's study the masking ability of the addition of random noise. For a test set, let's use the 1993 Annual Housing Survey (AHS) Public Use Microdata Sample. This file contains 64,998 records. Extract the following fields from each record:

(1) IDNUM - a unique 12-digit identifying number,
(2) REGION - region in which the housing unit was located,
(3) BEDROOMS - the number of bedrooms in the unit,
(4) BATHS - the number of bathrooms in the unit,
(5) YR_BLT - the year in which the unit was erected,
(6) INCOME - annual household income (top-coded at $100,000),
(7) HOME_VAL - the market value of the unit (top-coded at $350,000),
(8) MORTGAGE - monthly mortgage payment (top-coded at $1,800),
(9) MAINTAIN - annual maintenance cost (top-coded at $9,000),
(10) TAXES - monthly property taxes (top-coded at $62).

Use Field (1) for identification purposes, Fields (2) through (5) to determine relatively unique block combinations, and Fields (6) through (10) as matching variables. All matching variables have a bottom-code of 0. Before processing, set all bottom- and top-codes to blank. After the addition of the random noise, restore these codes. The addition of noise forces some values to exceed the top-code (or fall below the bottom-code). When this occurs, set the perturbed value to one less than the top-code. (For example, original INCOME = 98,000, Random Noise = 3,000, top-code = 100,000, perturbed INCOME = 99,999.) Use an analogous procedure when the noise causes the reported value to fall below its bottom-code.

Refer to Relationships (1) and (2) in Section IV. One can control the Non-Distortion Factor ($R_0$), equal to $1/(1+c^2)$. For a given value of $R_0$, calculate the corresponding value of c as $\{(1-R_0)/R_0\}^{1/2}$. $R_0$ is the ratio of the original (smaller) covariances to their perturbed (larger)

counterparts. Values of $R_0$ near 1 signify no distortion, near 0 much distortion. Software exists to produce noise that is centered at the origin with the desired variance-covariance structure. To convince oneself, refer to Table 2. Here values of random noise were produced for c = 0.1000. The mean for noise value is approximately 0 and its standard deviation is about one-tenth of the corresponding standard deviation in Table 1. For 64,998 values of random noise, the minimum and maximum noise occur about 4.0 to 4.5 standard deviations from the mean.

**Table 1. Statistics for the Fields Subjected to the Addition of Random Noise**

| FIELD | # OBS[1/] | TOP-CODE | BOTTOM-CODE | MEAN[2/] | STD DEV[2/] |
|---|---|---|---|---|---|
| **INCOME** | 35,717 | 100,000 | 0 | 11,369 | 13,986 |
| **HOME VALUE** | 31,394 | 350,000 | 0 | 97,698 | 67,184 |
| **MORTGAGE** | 15,908 | 1,800 | 0 | 607 | 359 |
| **MAINTAIN** | 18,374 | 9,000 | 0 | 571 | 821 |
| **TAXES** | 30,671 | 62 | 0 | 19.83 | 12.83 |

[1/] These counts exclude non-reported, top-, and bottom-coded values.

[2/] These statistics exclude non-reported, top-, and bottom-coded values.

**Table 2. Statistics for the Random Noise For c = 0.1000**

| FIELD | # OBS | MEAN | STD DEV | MINIMUM | MAXIMUM |
|---|---|---|---|---|---|
| **INCOME** | 64,998 | -6 | 1,389.0 | -6,497 | 6,016 |
| **HOME VALUE** | 64,998 | -39 | 6,716.8 | -29,065 | 27,190 |
| **MORTGAGE** | 64,998 | -1 | 35.8 | -153 | 149 |
| **MAINTAIN** | 64,998 | 0 | 81.7 | -353 | 331 |
| **TAXES** | 64,998 | 0 | 1.3 | -5 | 6 |

Because the variances and covariances are perturbed in the same direction and at the same rate,

Relationships (1) and (2) of Section IV force the following identity.

$$(3) \quad R(\underline{a}', \underline{b}') = R(a,b).$$

Table 3 illustrates that for a small value of c (c =0.1005), these coefficients are indeed not altered. Kim (1986) also has done extensive testing to ensure that Relationship (3) holds.

**Table 3. Correlation Coefficients Before and After the Addition of Random Noise For c = 0.1005 ($R_0$ = 0.99)**

| | | | Correlation Coefficients | |
|---|---|---|---|---|
| **Variable 1** | **Variable 2** | **# Obs.[1/]** | **Original** | **Post-Noise Observed** |
| **INCOME** | **HOME VAL** | 24,495 | 0.1410 | 0.1398 |
| **INCOME** | **MORTGAGE** | 12,002 | 0.0265 | 0.0263 |
| **INCOME** | **MAINTAIN** | 14,763 | 0.0427 | 0.0420 |
| **INCOME** | **TAXES** | 23,996 | 0.0910 | 0.0904 |
| **HOME VAL** | **MORTGAGE** | 15,511 | 0.6077 | 0.6073 |
| **HOME VAL** | **MAINTAIN** | 17,735 | 0.2022 | 0.2007 |
| **HOME VAL** | **TAXES** | 29,871 | 0.5763 | 0.5755 |
| **MORTGAGE** | **MAINTAIN** | 10,869 | 0.1657 | 0.1653 |
| **MORTGAGE** | **TAXES** | 15,323 | 0.5116 | 0.5101 |
| **MAINTAIN** | **TAXES** | 17,487 | 0.1667 | 0.1648 |

[1/] Number of records on which both fields contain a value. Records on which one of the fields contain a top- or bottom-coded values are excluded from the analysis.

## VI. The Masking Ability of Kim's Technique

Perform the following simulations. From the original file, extract a target set of 771 records with relatively unique categorical combinations. Each record is contained in mutually exclusive blocks with 7 or less records. For a given value of c, produce the random noise necessary. To compare the relationship of c to the masking ability of Kim's technique, generate perturbed sets for $R_0$ ranging from 0.900 to 0.995 in increments of 0.005. Use Winkler's software to count the number of correct and incorrect re-identifications.

Table 4 displays the results of this simulation. Column 1 contains the various values for the Non-Distortion Factors, while Column 2 contains the corresponding value of c. The number of correct and incorrect re-identifications are contained in the next two columns. The table also contains counts for questionable linkages (i.e., cases where the "best" linkage for the target record is a record from the perturbed file on which all block variables and all but one matching variable agree), and false linkages (i.e., all cases which do not fall into one of the other three categories).

**Table 4. Ability of Kim's Addition of Random Noise to Mask Data**
**771 Target Records Matched to 64,998 Perturbed Records**
**From the 1993 Annual Housing Survey Public Use File**

| | | Matching Results | | | |
|---|---|---|---|---|---|
| **Non- Dist. Factor [1/]** | **c[1/]** | **Correct Re-Ids** | **Incorrect Re-Ids** | **Questionable Links** | **False Links** |
| 0.995 | 0.0709 | 126 | 52 | 469 | 124 |
| 0.990 | 0.1005 | 115 | 54 | 411 | 191 |
| 0.985 | 0.1234 | 109 | 52 | 368 | 242 |
| 0.980 | 0.1429 | 106 | 52 | 349 | 264 |
| 0.975 | 0.1601 | 103 | 54 | 328 | 286 |
| 0.970 | 0.1759 | 99 | 53 | 320 | 299 |
| 0.965 | 0.1904 | 98 | 53 | 305 | 315 |
| 0.960 | 0.2041 | 96 | 53 | 300 | 322 |
| 0.955 | 0.2171 | 96 | 53 | 285 | 337 |
| 0.950 | 0.2294 | 93 | 53 | 280 | 345 |
| 0.945 | 0.2412 | 92 | 54 | 274 | 351 |
| 0.940 | 0.2526 | 96 | 55 | 259 | 361 |
| 0.935 | 0.2637 | 95 | 55 | 249 | 372 |
| 0.930 | 0.2744 | 95 | 54 | 249 | 373 |
| 0.925 | 0.2847 | 95 | 53 | 231 | 392 |
| 0.920 | 0.2949 | 94 | 53 | 244 | 380 |
| 0.915 | 0.3048 | 92 | 53 | 238 | 388 |
| 0.910 | 0.3145 | 92 | 54 | 235 | 390 |
| 0.905 | 0.3240 | 92 | 55 | 231 | 393 |
| 0.900 | 0.3333 | 92 | 55 | 230 | 394 |
| 0.590 | 0.4830 | 85 | 53 | 189 | 444 |
| 0.100 | 3.0000 | 82 | 54 | 174 | 461 |

[1/] Non-Distortion Factor ($R_0$) is set, then c is calculated as $\{(1-R_0)/R_0\}^{1/2}$.

As one increases the amount of distortion (i.e., the non-distortion factor decreases), the number of correct re-identifications decrease. For values of $R_0$ near 1.000, this number decreases "rapidly" with extremely small decreases to $R_0$. As $R_0$ drops below 0.970 (c = 0.1759), the decrease in the number of correct re-identifications becomes less and less obvious. As $R_0$ drops to 0.900 (c= 0.3333), the number of correct re-identifications is still 92. For extremely low values of $R_0$ (and high values of c) a "substantial" number of records can still be correctly re-identified. When the $R_0$ = 0.100, 82 records are still not "masked". In general, these are records on which most continuous fields contain missing values and the remaining fields are top- and/or bottom-coded. Although top- and bottom-coding limit the disclosure of sensitive information for certain fields, they do not prevent their use for identifying respondents.

At $R_0$ = 0.940 (c = 0.2526), the number of correct re-identifications actually increases. In these cases, there are two or more records in the perturbed universe on which the values of all matching fields are very similar to those of the target record. One is perturbed more than the other. For small values of c, this noise is not sufficient to distinguish the two records. However, as the distortion increases, the records do become sufficiently different. If the false linkage is the more perturbed, it will be “driven away” from the true linkage by larger values of c. This will cause an increase in the number of correct re-identifications. Quite frankly, I am surprised that the number increased this drastically at this value.

The number of incorrect re-identifications appear to be independent of the value of c. Throughout the simulations, this value oscillated between 52 and 55. Attribute this consistency to the number of records which contain only bottom-, top-coded, and missing values. Using a large amount of distortion does not appear to drastically change the incorrect re-identification count.

Is this re-identification criteria too strict? An intruder may count records, which agree on four of the five matching fields, as correct re-identifications. (For a larger number of matching variables, a logical model would not require all fields to match. For example, one might require the values in only 7 of 8 matching fields to agree.) Counts of linkages, which agree on only four of the matching fields, are listed in Table 3 under the column heading "Questionable Linkages". They also decrease as c increases. One can logically conjecture, "With a relaxed re-identification criteria, the rate of correct re-identification will significantly decrease, as c increases.”

Based on these simulations, very little protection is gained by using a value for the Non-Distortion Factor below 0.970 (or $c \geq 0.1759$). At this value, the Kim procedure protects all but 99 (about 13 percent) of the 771 target records. In addition, the intruder incorrectly re-identifies 53 respondents. Therefore, he is only correct on about 65 percent of the 152 re-identifications. For values of $R_0$ between 0.900 and 0.965, this percentage never goes below 62 percent.

For this set, values of c greater than 0.1759 provide little additional masking protection. This is very surprising. One assumes that he could almost completely mask any file with a large value of c to a large enough value. Yet when c is increased to 0.4830 and then to 3.0000, few extra records are masked. What is the relationship of c with its ability to mask a data file? By using

some mathematical model, can we predict the value where c "virtually" loses its masking power?

## VII. The Value Where c Loses Its Effectiveness, Mathematical Conjectures

This section states a set of conjectures which predict the value at which c provides little additional masking. I refer to these as "conjectures" rather than "theorems," since they were formulated to explain the simulation results of the previous section. Appendix A provides some technical justification for their validity. More rigorous mathematical proof work is necessary.

Conjecture 1. (An optimal value for c, when the number of matching fields is small) A data file has **n** matching fields. Assume **n** is small. A target record is correctly re-identifiable, if values from the corresponding offspring record agree (within **q** percent) on all **n** matching fields.

Let $p_0$ = the percentage of records which we can correctly re-identify after masking. Assume this value is almost 0. (In these simulations, it was difficult to get $p_0$ below 10 percent. Top- and bottom-codes restrict the lower bound on $p_0$. If one were to "blank out" the top- and bottom-coded values, $p_0$ would asymptotically approach 0.)

Suppose matching fields, $\underline{a}_i$ (with I=1, 2, ..., n) have top-, bottom-codes, and standard deviations of $A_i$, $a_i$, and STD($\underline{a}_i$), respectively. Let

$$c_i = \frac{q*(A_i - a_i)}{\sqrt{12}*STD(a_i)}.$$

The value where c loses its masking effectiveness is less than the minimum of $\{c_1, c_2, ..., c_n\}$.

Conjecture 2. If all matching fields are skewed toward the bottom-code, a value of c less than the one calculated above will protect the set.

Let's examine how the conjecture and corollary predict c for our simulation. Refer to Table 4. There was almost no masking improvement for values of c above 0.1759. The acceptable percent difference used was **q** = 0.10. Table 5 displays the values of $c_i$ suggested by Conjecture 1.

**Table 5. Theoretical Values at Which c Loses Its Masking Power**

| MATCHING VARIABLE | RANGE, $A_i - a_i$ | STANDARD DEVIATION | PCT DIFF ALLOWED, q | OPTIMAL VALUE OF c |
|---|---|---|---|---|
| **INCOME** | 100,000 | 13,986 | 0.10 | 0.2064 |
| **HOME VAL** | 350,000 | 67,184 | 0.10 | 0.1504 |
| **MORTGAGE** | 1,800 | 359 | 0.10 | 0.1447 |

| | | | | |
|---|---|---|---|---|
| **MAINTAIN** | 9,000 | 821 | 0.10 | 0.3165 |
| **TAXES** | 62 | 12.83 | 0.10 | 0.1395 |

The minimum value of c on this table is 0.1395. This suggests that our conjecture and its corollary underestimates (rather than overestimates) the "optimal" value of c. Why did this occur? Refer to table 4. Notice that the rate, at which the number of re-identifiables diminishes, decreases somewhat near c = 0.1429, 0.2041, and 0.3000. Assume a value of c = 0.1395 is used. A record is definitely masked, if and only if, the value of TAXES can be perturbed. TAXES are present on only 30,671 records. Therefore, little masking to TAXES will occur beyond c = 0.1395, since all non-masked records have missing, top-, or bottom-coded values for TAXES. However, some of these records may contain information in one of the other continuous fields. Higher values of c may be required to mask these fields. What are the alternatives? First, one may desire to restrict his analysis to matching fields where a substantial percentage of records contain values which are subject to noise. INCOME has the most values present. The optimal c for INCOME is 0.2064, a value much closer to the observed "optimal" value. A second alternative is to use the maximum of the $c_i$. This will guarantee the needed protection.

Return to the proof of Conjecture 1. It requires that all values in one matching field be perturbed outside the **q**-percent range. Is this too stringent? When there are a large number of matching variables, is it not likely that the file will be protected when only a few severe perturbations are found in each field? The next conjecture addresses this problem.

Conjecture 3. (An optimal value of c for a large number of matching fields) Assume one has a large number of matching variables (e.g., $\mathbf{n} \geq 9$). One obtains the same optimal value of c as that found in Conjecture 1, if he requires only one field to be outside of the **q**-percent range.

Let's finish this section by considering the following problem. Suppose an intruder had **n** matching fields. Suppose he would allow a record to match on only **n-k** fields to be considered a re-identification. How would one choose a suitable value of c?

Conjecture 4. (An optimal value for c, when only **n-k** of the **n** fields are required to match in the **q** percent range) To be a re-identification, assume an intruder requires only **n-k** fields to match. To solve for the problem, calculate $c_i$ (in the same manner as Conjecture 1) for all matching fields for which there are a substantial percentage of non-missing, non-top-coded, and non-bottom-coded responses. Ignore the lowest **k** values calculated. Set c to the next lowest value.

To illustrate this: An intruder allows one field to not match. INCOME is the field with the most responses present. Many other records have values for 2 or more of the matching fields. Some may be top- or bottom-coded. How many records can be masked? Look in the last column of Table 4. For c = 3.000, the intruder would be unable to re-identify 461 records.

When c = 0.2041 (near the suggested cut-off for INCOME), the process has masked 299. This number steadily increases through c = 0.2637 and beyond. At this value, 372 cases are re-

identifiable. Soon after this point, the number of additional re-identifiable records slows to a crawl. At c = 0.3333, there are only 310 such cases. Assuming INCOME is the best variable to use as a cut-off, use the next highest value of c, 0.3165. It corresponds to MAINTAIN.

This illustration is flawed for several reasons. First, MAINTAIN is probably not a suitable second field. It contains only 18,374 observations; INCOME contains 35,717. Second, one is interested in only correct re-identifications. To do the analysis correctly, one would have to segregate the "Questionable Linkages" into "Correct Questionable Linkages" and "Incorrect Questionable Linkages" categories. Then you would compare the sums of the "Correct Re-identifications" and "Correct Questionable Linkages" for each value of c. Nevertheless, this example illustrates that the method suggested by Conjecture 3 may be a reasonable way to set an upper bound on c.

## VIII. The Advantages and Disadvantages of Kim's Masking Technique

Kim's (1986) noise addition technique is quite attractive. The mask itself is done in such a way that two key qualities of the microdata file are preserved.

(1) Noise is generated (and added) in an unbiased manner. The expectation of the mean for each continuous variable is not changed by the swap. This includes not only means for the entire universe, but also the means of any subset.

(2) The amount of bias introduced to the covariances and correlation coefficients are predictable. They can be controlled by manipulating the value of c in the term $(1 + c^2)$. Again the biases apply not only to the entire universe, but also to any subset.

The noise addition technique does have certain disadvantages. Two are listed below.

(1) Highly visible values (such as, large values of INCOME) receive, on average, the same amount of noise as values which lie extremely close to the bottom-code. If the re-identification routine is based on a percentage difference, the highly visible values are more likely not be to be masked.

(2) As seen in Table 4, Kim's routine does not guarantee that the file is masked. Even large values of c, which significantly diminish the utility of the perturbed file, do not provide enough protection for some records.

Because of these two disadvantages, microdata files masked by this technique may be unable to be released. Re-identification software must be used to determine (1) how well the file has been masked and (2) how much additional protection is required. If more protection is necessary, can the noise-added file be perturbed to provide the additional protection? Or should we use a different method of perturbation on the original file? Winkler's software provides the answer to these questions and concerns.

## IX. Winkler's Matching Software, An Overview

Following any masking procedure, the public use file probably contains records which sophisticated intruders may be able to re-identify. This was an especially touchy situation. The HHS special request microdata file contained information which the IRS could match back to their files to identify respondents. The IRS could then obtain the CPS data for these individuals. To determine whether Kim's approach sufficiently masked the file, Winkler's software was used to determine the extent to which re-identification was possible.

Winkler's software (1994, 1995a) is among the most sophisticated in existence. Records are first assigned to equivalence classes or blocks. Blocks are generally constructed from information that may be readily available to any intruder. The less readily available quantitative (or matching) variables need not match exactly. Program parameters allow the user to decide the maximum percentage which he will allow the target and universe values to differ. For each record in the target file, pairs are formed with records in the same block of the universe file. The Fellegi-Sunter (1969) routine is used to assign each pair a score. This score is an aggregate sum of weights assigned when each matching variable is compared to the value of the corresponding field in the paired universe. The Estimation Maximization (EM) algorithm for latent classes is used to calculate the optimal discriminating weights. The pair with the highest score is linked. The higher the score the more likely the linkage is correct. Based on pre-specified cut-offs, this score determines definite mismatches, questionable linkages, and definite re-identifications.

When Winkler applied his software to the file masked by Kim, he re-identified a large portion of the records. The Census Bureau had strong apprehensions about releasing such a file. Winkler's software was so sophisticated that no technique, which preserves analytic validity, adequately protected this file. More must be done. Could the re-identification software be used to mask the file?

Winkler enhanced the record matching software. The user now has the ability to specify a percentage, **p**, of the correct re-identifications be swapped. The software not only keeps track of 0.the linked record with the largest score, but also the one with the second largest score. The software first checks to ensure that the linkage is a definite re-identification (i.e., the largest score is above some user-specified threshold). If this is the case, it then compares the identification numbers. When these agree, one has a correct re-identification. Finally, the routine draws a random number between 0.0000 and 1.0000. If this number is less than the specified swapping percentage, **p**, a swap is performed. Each value from each matching field is swapped with the corresponding value on the second most likely linkage. This will force at least **p** (and possibly 2***p**) percent of the “previously correct re-identifications” to now be “incorrect re-identifications”.

Example. Refer to Table 3. Suppose Kim's technique with c = 0.1234 is used. Winkler's software reveals that 109 cases can be correctly re-identified, and 52 cases incorrectly re-identified. Invoke the software to swap, for example, 10 percent of the correctly re-identified cases. Expect it to swap about 11 cases. Following the swap, about 98 cases would be correctly and 63 incorrectly re-identified.

## X. The Advantages and Disadvantages of a Swap Procedure

Winkler's technique has two very obvious advantages.

(1) Regardless of the previous masking techniques used (top-coding, Kim's noise additive approach, etc), one can swap as many re-identifications as desired. Using a swapping percentage of 100 percent (**p** = 1.0000), one can mask all re-identifications.

(2) The Winkler swap preserves means and variance-covariance structure of the entire universe. Suppose ($\underline{a}$, $\underline{b}$) are the paired values of any two fields before the swap and ($\underline{a}'$, $\underline{b}'$) the paired values after the swap, then

(a) $\overline{\underline{a}'} = \overline{\underline{a}}$,

(b) $\overline{\underline{b}'} = \overline{\underline{b}}$, and

(c) $R(\underline{a}', \underline{b}') = R(\underline{a}, \underline{b})$.

These three properties are identities, not just expectations. Regardless of the number of swaps, one still preserves them. Should not we swap all re-identifications (**p** = 1.0000)? Winkler's approach has several drawbacks, so small values of **p** are desirable. Its major weaknesses are listed below.

(1) Unlike Kim's approach, Winkler's swapping does not guarantee the preservation of means and the variance-covariance structure for arbitrary subsets.

The more records swapped, the more potential for subset statistics to be distorted. The amount of distortion increases with the number of swaps, so avoid a massive swap.

Winkler's software does match within blocks (i.e., specific key combinations). The swapping will not affect statistics of subsets constructed by combinations of blocks. The swap is done in such a way that a record is swapped with its most similar (i.e., similar using the second best match from the EM method). This should cause less distortion than a random swap.

(2) Winkler's approach is based on the ability to reasonably construct an intruder model. Re-identification is dependent on (a) the construction of a target set, (b) the setting of the amount of perturbation which an intruder allows, and (c) defining appropriate cut-offs for

re-identifications, questionable linkages, and definite false linkages. Does one have an appropriate understanding of the sophisticated intruder to set these parameters correctly? Are we giving the intruder too much credit? To how much information does he have access? How reliable is his information from alternative sources?

## XI. Conjectures for a Re-Identification Tolerance Level, Historical Evidence

The previous section states some of the caveats of direct data-swapping. Too much may diminish the credibility of subset statistics derived from this file. Yet, agencies should be compelled not to release data for which certain cases are "high disclosure risks". The expanding capacity of computers and the constant improvement of matching software makes total elimination of disclosure risk an impossibility. Sullivan and Fuller (1989) developed a noise-additive technique which creates a file where "[t]he probability that an intruder with some information on an individual can correctly identify the record of that individual is considerably less than one."

How can one determine what is "considerable?" First, examine the results of Paas and Wauschkuhn (1985). They constructed an intruder model with a target set and matching software (See Section III.). Records in the target were matched to the mask file and the number of correct and incorrect re-identifications tallied. How do their results compare with those in Table 4? Paas and Wauschkuhn correctly re-identified 20 percent of their target file. They concluded that the current software is so sophisticated that it is virtually impossible to mask any microdata file.

Believing that Paas and Wauschkuhn were too liberal in assessing the amount of information available to the most sophisticated intruder, Muller, Blien, and Wirth (1995) questioned this conclusion. To attain such a high matching rate, an intruder must be certain of the following.

(1) Every individual in the target file definitely has a record in the microdata file.

(2) The data in the target file is consistent with the original data in the microdata file. Non-sampling errors must be consistent between the two files.

(3) The intruder has very specific knowledge of the perturbation technique used.

Never do microdata files contain all individuals. The intruder's best scenario is to create target file of records that are unique to the universe. Whenever matches to the microdata file are found, they are very likely to be correct re-identifications. For public use microdata samples with low sampling rates (e.g., 1 percent or 1 in 1500), most of the targets would not match.

Suppose an intruder does "re-identify" a small number of targets. Are the files consistent on the key and matching fields? Data collected from human respondents are always subject to response and transcription errors. In addition, the intruder must assume that records in both files were subjected to similar edit and imputation procedures.

Assume that the intruder is also assured that all sources use extremely meticulous methods to obtain and process this data. In order to be certain of his matches, he must have some knowledge of the perturbation process. If he has no exact knowledge, his matching software may not be suitable to detect re-identifications. (The Winkler software is less susceptible to this than most matching software.) If he is aware of the technique, but not aware of the extent of the perturbation, he may not be able to set certain matching parameters correctly.

Muller, et al point out that in Paas and Wauschkuhn's model, the intruder had all of the three above concerns tilted to his favor. Their intruder knew that all targets were in the microdata file. He knew the fields were consistent, since the target file was an electronically transferred version of the original file. He tailored his software to undo the perturbation introduced by the masking. Under these extreme conditions, Muller, et al concluded that a 15 to 20 percent re-identification rate was not considerable.

Return now to the HHS public use microdata file. Suppose the IRS obtains a copy of this file and produces a target set from the file. It could then match this set against the entire 1040 universe, which contains unique identifiers and the unperturbed values for many of the matching fields. The IRS also has very sophisticated computers and the finances to obtain sharp matching contractors. In this case, the IRS could re-identify significant portions of the universe, except for the fact that it does not know the parameters which were used. Realizing this, Kim and Winkler had to use additional care in choosing these parameters.

Fienberg, Makov, Sanil (1995) developed an intruder model. Its approach is similar to the one of Paas and to that of this study. In this model, the intruder not only considers random noise (error added to mask the data), but also a percentage bias caused by the sensitivity of the attribute (e.g., income, moral views) and the subjective feelings of the respondent. Suppose the intruder had some knowledge of the sensitivity of each respondent. In this model, he could regulate this bias between respondents. Suppose $x_{ij}$ is the target value of the j-th field for the i-th respondent, the corresponding perturbed value would be $x_{ij}' = (p_{ij} * x_{ij}) + \epsilon_j$. Fienberg, et al concluded that (1) the random noise, $\epsilon_j$, alone made re-identification difficult; (2) respondent bias, $p_{ij}$, added little additional protection. In a simulation, Fienberg used a large amount of noise ($c \underline{\sim} 1.00$, where he uses c to define a noise distribution in the same manner as Kim) in conjunction with <u>no</u> top- or bottom-codes. The model still re-identified 43 (or 6.5 percent) of 662 records correctly.

Most intruder models are very similar to that of Paas. Unfortunately, there is no quick and inexpensive method for constructing a target file which is independent of the perturbed file. Nor is there any method of determining how much perturbation an intruder suspects has been introduced. Consequently, our only choice is to assume that the intruder does have precise information and that our information is also precise. The intruder's targets must be so unique, that if they are in the public use file and they are not masked, they will be re-identified. Finally, the intruder realizes that the releasing agency does not want to seriously limit the utility of the file, so it imposes some restrictions on the amount of allowable perturbation. This allows him to make a reasonable guess as to the amount of perturbation used.

Let's assume the intruder does everything well. It is not unreasonable for him to re-identify a large portion of his target set. Consequently, under perfect conditions, any intruder will correctly re-identify at least 10 percent of a public use microdata file. It is unreasonable to assume that he does do everything almost perfectly. As the quality of one or more of these conditions deteriorate, the intruder loses much of this ability. This leads to the following conjecture.

Conjecture 5 (Re-identification Tolerance, Approach #1). Under ideal conditions, an intruder will be able to correctly re-identify at least **Q** percent of his target file. Winkler's matching software operates under these conditions. Assume data is swapped, so that Winkler's software will correctly re-identify only **Q** percent of its targets. Then the public use file is adequately masked. Assume a value of **Q** = 15 percent provides "considerable" masking.

Semi-proof. **Q** = 15 percent is based, strictly on simulated results of Paas and Wauschkuhn (1985); Muller, Blien, and Wirth (1995); Fienberg (1995); and of this study (See Table 4.).

Refer to Table 4. When an extremely high value of distortion is used (c = 3.000, $R_0$ = 0.1000), one can correctly re-identify 82 records and incorrectly re-identify 54. This ratio is approximately 3 to 2 (or 0.60 to 0.40). This is an extreme amount of distortion. When less is used, expect the ratio to be larger. Consequently,

Conjecture 6 (Re-identification Tolerance, Approach #2). In order to sufficiently mask a public use microdata file, one should shoot for a ratio of 2 correct to 1 incorrect re-identification.

## XII. Using Winkler's Approach to Achieve Re-Identification Tolerance

Winkler's swapping software is potentially a very powerful disclosure limitation tool. Regardless of the method used to mask the data, it can be used to ensure that certain disclosure limitation standards have been met. The beauty of the software is that it is independent of the masking technique used. Use any technique to mask the file, then use his software to re-identify the records in some portion of the file. If necessary, swap some records. This ensures an extremely low probability that an intruder will be able to accurately use the file for re-identification purposes. One only needs to solve the question, "How much swapping is necessary?"

Let's illustrate how to use Winkler's software to attack these problems. Use the parameters suggested in Conjectures 5 and 6. In addition, solve the problem using slightly different parameters. Such parameters are still subjective. In order to include this software in any disclosure limitation routine, the Census Bureau must adopt some standards.

Examples. Suppose one uses Kim's noise-additive technique to mask a microdata file, then attempts to re-identify 771 target records. Assume a value of c = 0.1005 is used. From Table 4, one sees that we have 115 correct and 54 incorrect re-identifications.

Problem. Is swapping necessary? If so how much?

Solutions. The answers to these questions lie in the criterion used from Section XI.

Criterion 1. The file is protected, if 15 percent or less of the targets are re-identifiable.

Solution 1. $0.15 * 771 \simeq 117$ correct re-identifications. This method masks all but 115. This is less than 117, therefore no swapping is necessary.

Criterion 2. The file is protected, if the ratio of correct to incorrect re-identifications is 2 : 1 (0.67 : 0.33).

Solution 2. There are 169 re-identifications. One wants $0.67 * 169 \simeq 112$ correct re-identifications. One has 115 such cases. We need to distort 3 correct re-identifications. Set the software to perform 2 or 3 swaps.

## XIII. Conjecture on an Optimal Masking Strategy

Since Kim's noise-addition technique preserves (in mathematical expectation) the mean of each continuous variable, it appears to be advisable to use Kim's technique to mask as large a portion of the file as possible. This would involve setting the Non-Distortion Factor ($R_0$) as low as tolerable (and c as large as tolerable). If analysis (similar to that in Section VII) were to show a smaller value of c masks the file almost as well, use the smaller value. Following this masking procedure, use Winkler's matching software to determine the number of additional swaps necessary to adequately mask the file, and swap this number of records. This swap may bias the moments of some subsets. However, since c was taken to be as large as tolerable (or efficient), this bias should be minimal.

## XIV. Future Research

The ideas listed in this section are solely those of the author. They in no way reflect the views of the US Bureau of the Census or its Statistical Research Division. These ideas are intended only to suggest possible areas of future research.

1. Development of Formal Proofs for the Conjectures of Section VI. The conjectures of Section VI were based on very vague assumptions. These assumptions need to be specified more precisely. The conjectures could then be proven using rigorous epsilon-delta arguments. Assuming uniform distribution on the interval between the bottom- and top-codes seems to give good estimates. Could this be generalized to give better estimates?

2. Perform More Simulations With Finer Iterations. Simulations were performed for iterations of $R_0$ of 0.005. It was difficult to determine the sensitivity of the correct re-identification rate to changes in $R_0$. The program should be re-executed with iterations of 0.001 (or less) in the range $0.975 < R_0 < 0.999$. This should shed more insight on the relationship of these two variables. In addition, refer to Table 5. Three of the five values for c are extremely close; for TAXES, the c

value is 0.1395; for MORTGAGE, 0.1447; for HOME VALUE, 0.1504. In Table 4, values of c jump from 0.1429 to 0.1601. Finer re-iterations may allow us to analyze the masking component of each of these variables separately.

3. Using Kim's Technique To Relax Top- and/or Bottom-Codes. Corollary 1 (Section VI) shows that the maximum ability of $c_i$ to distort was a function of the top- and bottom-codes. Assume we are going to use Kim's technique to mask a data set. Could we fix $c_i$ , then use this formula to relax the top- and or bottom-code? Kim's technique worked well on this data file. Would it have worked as well if the data had not been top- and bottom-coded?

4. Development of a Standard Set of Parameters for the Kim and Winkler Routines. The Census Bureau has already used the techniques in the paper to mask a public use file. Before employing them again, it should study feasible values for all parameters involved. These should include:

(a) the optimal value of c, the amount of distortion allowable, for the Kim routine;

(b) the allowable percentage difference between the values of the target set and perturbed universe in order for the matching variables to be considered to be in agreement;

(c) the cut-off aggregate scores for pairs to be considered definite re-identifications, questionable linkages, and definite false linkages; and

(d) the maximal percentage of re-identifications (and/or the target ratio of correct to incorrect re-identifications) allowable for a public use file to be considered as having a "limited risk of disclosure".

5. Development of an Intruder Model. Some standards or guidelines should be developed for simulation of the actions of a typical intruder. These include recommendations for the following concerns. Some of these concerns are closely related to those expressed in Item 4 above.

(a) What are a reasonable set of block variables? What categorical information should we include in the block? (i.e., How should we determine what information may be readily available to the most aggressive intruder?)

(b) How few observations should each block in the target set contain? (i.e., How close to unique does an individual have to be, in order for an intruder to search for him?)

(c) What are a reasonable set of matching variables? (i.e., What continuous variables would an intruder use to aid him in the identification of his targets?)

(d) How much noise does an intruder expect between continuous variables in his target set and those in the released microdata file? Is this a function of, from where, and how the intruder obtained his information?

## XV. Conclusions

The Kim-Winkler approach is a very powerful method for masking data files. This technique combines the positive attributes of two separate methodologies. The Kim method adds random noise to a microdata file. It generates noise so that the means of the universe and every subset are preserved, in mathematical expectation. In addition, the variance-covariance structure of the file is diminished by a known factor. The user can control this factor, if he desires to constrain the extent to which the multivariate relationships are disturbed. The Winkler method swaps continuous variable information to limit the ability of the file to be used for re-identification purposes. This swap will not harm the means or the variance-covariance structure of the universe. Values are swapped with records that usually have similar values. However, there is no guarantee that a subsets statistics are not significantly altered by this swap. As a result, it is generally a good idea to swap as few records as necessary.

Kim and Winkler took the proper approach to mask a public use file for the Department of Health and Human Services. Kim's methodology was used first. This preserved the means of the universe and all subsets, while expanding the variance-covariance relationships by a factor of $(1 + c^2)$, where Kim chose the value of c to limit the distortion. Winkler then used his matching software to identify records which Kim's methodology may not have provided enough protection. Winkler then swapped a percentage of these records to further ensure that the confidentiality of the individuals on the file was not violated.

There is a critical value for c. Above this value, Kim's method provides little additional protection. The critical value is a function of the matching variables, their top- and bottom-codes, and their standard deviations between these codes. It is also a function of the maximum percentage within which an intruder expects the values of each record in the publicly released file to differ from those of the corresponding record in his target file. Consequently, given a known set of matching variables and this percentage, one can calculate the critical value of c.

Following the addition of noise, some additional matching is required to provide "considerable" masking. Based on published results by Paas and Wauschkuhn (1985); Muller, Blien, and Wirth (1995); Fienberg, Makov, and Sanil (1995); as well as the results of this study, we conclude that a 15 percent re-identification for an offspring target file is not excessive. To the casual reader, this rate may seem excessive. However, keep in mind that the intruder cannot be sure that (1) all his target records are in the publicly released file, (2) the values in the fields in the target and the microdata files are consistent, and (3) the method and extent to which the data was perturbed. Any one of these factors alone significantly diminishes the accuracy of re-identification software.

The author recommends that this method be used more often to mask microdata files. As more research is done in this area, we will not only be able to determine the critical value of c, but also, perhaps, to use the techniques of Kim and Winkler to relax top- and bottom-codes. If we can determine the psychological mind-set of the sophisticated intruder, we can also refine previous estimates on an acceptable tolerance for the percentage of re-identifications in an arbitrary offspring target file, as well as new innovative methods of creating such files.

## XVI. References


1. Felegi, I. P. and Sunter, A. B. (1969). A Theory of Record Linkage, Journal of the American Statistical Association, **64**, 1183-1210.

2. Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1995). A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for continuous Data. Bureau of the Census Contract No. 50-YABC-2-66025, Task Order No. 8. Washington, DC.

3. Kim, J. J. (1986). A Method For Limiting Disclosure in Microdata Based on Random Noise and Transformation, Proceedings of the Survey Research Methods Section , American Statistical Association, 370-374.

4. Kim,J. J. and Winkler, W. E. (1995). Masking Microdata Files, Proceedings of the Survey Research Methods Section, American Statistical Association, to appear.

5. Muller, W. , Blien, U., and Wirth, H. (1995). Identification Risks of Microdata, Sociological Methods and Research, **24**, 131-157.

6. Paas, G. And Wauschkuhn, U. (1985). Datenzugang, Datenschutz, und Anonymisierung, Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten. Munich and Vienna: R. Oldenbourg.

7. Subcommittee on Disclosure Avoidance Techniques (1994). Statistical Policy Working Paper No. 22: Report on Statistical Disclosure Limitation Methodology. Statistical Policy Office, Office of Information and Regulatory Affairs, office of Management and Budget, Washington, DC.

8. Sullivan, G. And Fuller, W. A. (1989). The Use of Measurement Error to Avoid Disclosure. Proceedings of the Section on Survey Research Methods, American Statistical Association, 435-439.

9. Winkler, W. E. (1994). Advanced Record Linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association, 467-472.

10. Winkler, W. E. (1995a). Matching and Record Linkage, in Business Survey Methods (B.G. Cox, ed.), New York: J. Wiley, 355-384.

11. Winkler, W. E. (1995b). Reidentification and Swapping Software. (Unpublished Documentation), Statistical Research Division, US Bureau of the Census.

## Appendix A. The Value Where c Loses Its Effectiveness, A Technical Approach

This appendix is meant to provide mathematical support for the claims of Section VII. Because the "proofs" require very specific circumstances, I prefer to refer to these as "conjectures," rather than "theorems". This appendix is not meant to provide rigorous proofs, but rather to illustrate that there is reason to believe that extremely large values of distortion provide no better masking than much smaller values.

Conjecture 1. (An optimal value for c, when the number of matching fields is small) A data file has **n** matching fields. Assume **n** is small. Call a target record correctly re-identifiable, if values from the corresponding offspring record agree (within **q** percent) with the corresponding values in all **n** matching fields and all block variables.

Let $p_0$ = the percentage of records which we can correctly re-identify after masking. Assume this value is almost 0. (In the simulation of Section IV, it was difficult to get $p_0$ below 10 percent. Top- and bottom-codes restricted the lower bound on $p_0$. If one were to "blank out" the top- and bottom-coded records $p_0$, would asymptotically approach 0.)

Suppose matching fields, $\underline{a}_i$ (with i=1, 2, ..., n) have top-, bottom-codes, and standard deviations of $A_i$, $a_i$, and $STD(\underline{a}_i)$, respectively. Let

$$c_i = \frac{q*(A_i - a_i)}{\sqrt{12}*STD(a_i)}.$$

Then, the value where c loses its masking effectiveness is less than the minimum of $\{c_1, c_2, ..., c_n\}$.

Semi-proof. A re-identification requires the value of each matching field from a perturbed record to agree with its offspring target by less than **q** percent. Since $p_0$ is almost 0, and **n** is small, there probably exists at least one field where almost all values are perturbed more than **q** percent. Let's construct a value for c, which will cause every value in at least one field to fall outside of this **q** percent range.

Let f(y) be the probability distribution for field $\underline{a}_1$.

Let $p_1$ be the probability that the perturbed values of $\underline{a}_1$ fall outside of the **q** percent range. The noise is multivariate normal with standard deviation, $c_1*STD(\underline{a}_1)$, thus

$$0 = 1 - p_1 = \int_{a_1}^{A_1} \int_{(-q*y)}^{(q*y)} f(y) * \frac{e^{\frac{-x^2}{2*c_1^2*STD^2(\underline{a_1})}}}{c_1 * STD(\underline{a_1}) * \sqrt{2*B}} dx\ dy.$$

Assume all perturbed values of $\underline{a}_1$ fall outside of the range, $p_1$ is approximately 1, then the left-hand-side of the above equation is approximately 0.

Now assume f(y) is uniform distributed, thus $f(y) = 1/(A_1 - a_1)$, for every value of y in $\underline{a}_1$. Replace the numerator with its Taylor series expansion. Truncate this series at $x^2$. Integrating and solving, one finds

$$c_1 = \frac{q*(A_1 - a_1)}{\sqrt{12}*STD(\underline{a_1})}.$$

Calculate similar values of $c_i$ for Fields $\underline{a}_2$, ..., $\underline{a}_n$. Let $c = c_j$ = min of $\{c_i\}$. This obviously protects almost all the values in the j-th field. Therefore, it protects each record.

Each of the above fields are not uniformly distributed, but skewed toward the bottom-code. Therefore, smaller values of $c_i$ will protect the i-th matching variable.

Corollary 2. If all matching fields are skewed toward the bottom-code, a value of c smaller than the one calculated in Conjecture 1 will protect the set.

Return to the proof of Conjecture 1. It requires that all values in one matching field be perturbed outside the **q**-percent range. Is this too stringent? When there are a large number of matching variables, is it not likely that the file will be protected, with only a few severe perturbations are found in each field? Conjecture 3 addresses this problem.

Conjecture 3. (An optimal value of c for a large number of matching fields) Assume one has a large number of matching variables (**n** $\geq$ 9). The same optimal value (i.e., the same as calculated in Conjecture 1) of c will be obtained, if only one field is required to be outside of the **q** percent range.

Semi-proof. For a record to be unprotected, each matching variable must be inside the range. For the -I-th variable, this probability is $1 - p_i$. Assuming independence

$$(1-q) = \prod_{i=1}^{n} (1-p_i).$$

Assume that each value of $p_i$ is approximately the same.

$$1-p_i = (1-q)^{\frac{1}{n}}.$$

Since **n** is large the right-hand side is approximately 1, for every i. Hence each $p_i$ is approximately 0. (**n** was small in Conjecture 1, forcing $p_i$ to be approximately 0, for some i.)

Proceed by calculating each $c_i$ in the same manner as in Conjecture 1. You will get the same values as in Table 5. This shows that the value for large **n**, the following are equivalent: (1) finding a value of c which perturbs all values in any one field, or (2) finding a value of c which perturbs at least one value on any field of each record.

Suppose an intruder had **n** matching fields. Suppose he would allow a record to match on only **n-k** fields to be considered a re-identification. How would one choose a suitable value of c?

Conjecture 4. (An optimal value for c, when only **n-k** of the **n** fields are required to match in the **q** percent range. Assume an intruder only allows **n-k** fields to match. To solve for the problem of Conjecture 1, calculate $c_i$ for all matching fields for which there are a substantial percentage of non-missing, non-top-coded, and non-bottom-coded responses. Ignore the lowest **k** values calculated. Set c to the next lowest value.

Semi-proof. Assume c is taken to be the **k-th** lowest. This will cause the record to fail on that field. It will also cause the record to fail on the **k-1** fields that have lower corresponding $c_i$'s. It will not make any guarantees that a value in one of the other fields will fail. Thus, the record may meet the re-identification criteria. The **(k+1)-th** lowest will make this guarantee.

## Appendix B. Algorithm for Using Kim-Winkler Method to Protect Microdata

I. Definitions

A. Identifying Variable - a field which is unique to each record on the public use file.

B. Block Variable - any field for which an intruder could easily obtain information for any given target record(s). The value may be obtained through common knowledge (e.g., a geographic code) or through public records (e.g., the number of bedrooms and baths obtained from local tax records).

C. Block - all records for a given combination of block variables.

D. Matching Variable - any variable for which an intruder would desire an exact value. This data is deemed too sensitive to be released as part of public record. It is assumed that the intruder can make reasonable estimates for some (possibly all) of the matching variables. Assume that through the use of fairly sophisticated matching software, the intruder will attempt to accurately re-identify his target.

E. Random Noise - values added to each matching variable, to ensure that the intruder will not get exact values for a given target, even if the target is correctly re-identified.

F. Swapping - a technique which swaps the value of each matching variable on a re-identifiable record with the corresponding value on the "most similar" non-matching record. This forces sophisticated matching software to incorrectly identify the target.

II. General Approach

A. Add random noise to the value of each matching variable. The noise will be added in a way that preserves the univariate means and disturbs the univariate standard deviations and bivariate covariance relationships in a predictable manner.

B. All blocks in the public use file will be required to have at least "n" observations. Identify all blocks with less than this tolerance level. Extract these records to a separate file.

These records will be matched against the entire universe. The blocks will be constructed by combining the initial blocks across geography. For example, (initial block structure: baths-bedrooms-year built-region, new block structure: baths-bedrooms-year built). Re-identifiable records within the new block will be swapped to preserve anonymity.

C. Following the addition of random noise and the swapping, all records in blocks (created with the initial block structure) with "N" ($n << N$) or less observations will be matched against the entire sample. This will give us some measure whether a well-informed intruder with sophisticated software can re-identify a large portion of his targets.

III. The Computer Programs

All programs are found on the Sun Unix (srdsbn1:/home/rmoore/agstock/mask_soft). The programming languages are given by the extensions used (*.SAS for SAS, *.f for FORTRAN, *.C for C++).

A. Preparing the Data Set

Step 1.SAS: Create a SAS dataset with the blocking, matching, and identification variables.

Step 2.SAS: Determine the top- and bottom-codes for the matching variables.

Step 3.SAS: Create a modified set with all top- and bottom- codes set to blank.

Step 4.SAS: Calculate the variance-covariance structure of the modified set.

B. Generate and Add Random Noise

Step 5.F: Calculate a set of multivariate random noise with the (only scaled) variance-covariance structure of the modified dataset.

Parameters to be set: (1) Number of Records
(2) Number of Variables
(3) The Value of c
(4) The Random Start

F. Step 6.SAS: Evaluate the noise dataset. Ensure that the means and standard deviations of the noises and the absolute noises are consistent with those of the modified dataset and the scaling factor.

G. Step 7.SAS: Add the random noise to the non-blank values in the modified set. All Perturbed values must fall between the top and bottom codes.

H. Step 8.SAS: Calculate the variance-covariance structure of the modified set after the addition of noise.

I. Step 9.SAS: Perform post-noise-addition univariate analysis.

(1) Calculate the univariate biases for the means of each continuous variable.
(2) Check univariate frequency distribution changes.

C. Determine the Amount of Re-identification After Noise Addition

Step 10.SAS: Restore the top- and bottom-coded values.

Create an intruder target set. Set the tolerance level for the minimum number of observations per block. Extract all records that are in sub-tolerance level blocks.

Form new blocks for the extracted set and the original sample (before noise) by combining the original blocks across geographies. Exclude all records where the geographic codes have been suppressed. (The program is designed to handle less than 2,000 observations per block. Inclusion of these records may cause the block to contain more observations than can be handled.)

Step 11.SAS: Create the parameter files for calculating the "weights" for each matched and non-matched matching field. The "weights" are calculated using the estimation-maximization algorithm. This process is iterative and requires the user to provide a set of reasonable starting weights. I have always had good luck with "0.80 and 0.05" for the initial guesses for each matching variable.

This parameter file also contains the block structure.

Step 12.C: Provide counts for the number of records that satisfy each of the $2^k$ matching combinations, where k = the number of variables (e.g., (match on var. 1, match on var. 2, non-match on var. 3, ...). The next step uses the output of this program.

Check the file "sumcb.dat" to ensure the program has run correctly.

Step 13.F: Calculate the optimal weights using the EM algorithm.. These will be contained in the file "initi.dat".

Step 14.SAS: Transfer the optimal weights from "initi.dat" to the cards section of this routine. When the program is run, the parameter file will be updated with the correct weights.

This file also contains a record for the re-identification, questionable linkage, and incorrect linkage cut-offs.

Step 15.C: Swap 0 percent of the initial re-identifiable targets.

Parameters: (1) PROP_DIFF 0.pqrs. Here "pqrs" is the acceptable difference between values on the target and perturbed records.
(2) SMP_PROB 0.0000. SMP_PROB is the probability that a re-identified record is swapped.

The file summ.dat is a summary of this matching operation. The file contains the following important output.

(1) arecs = number of records in target file,
(2) brecs = number of records in perturbed file,
(3) skipa = number of records in target with no corresponding block in perturbed file,
(4) skipb = number of records in perturbed file with no corresponding block in the target,
(5) match pairs = number of re-identifications,
(6) true matches above HC = number of correct re-ids,
(7) clerical pairs = number of questionable linkages,
(8) unmatched A records = number of false linkages.

D. Swap the Necessary Number of Records

Determine the number of correct re-identifications to be swapped, $r_c$. Set SMP_PROB $\simeq$ 0.67 * $r_c$ / match pairs.

Step 16.C: Swap SMP_PROB percent of the initial re-identifiable targets.

The file summ.dat is a summary of this matching operation. The file contains the same information as in Step15.SAS plus

(9) swap matches above HC = correct re-ids swapped.