Table of Contents

Executive Summary

For years, the U.S. Bureau of the Census has collected and disseminated data. It has come to realize the importance of these data products for research, analysis, planning, and policy-making. Technological advances of the 1980s have greatly increased the demand for data, particularly for the microdata products. Users now demand more detailed microdata sets than ever before. Unfortunately, the more information which the Census Bureau provides, the greater the risk that a user can determine the responses for some respondents.

Title 13 requires that the Census Bureau take the necessary steps to ensure the confidentiality of all respondents. Traditional approaches to disclosure limitation (i.e., supplying the user with microdata for only a small sample of the population, using bottom- and top-codes for continuous variables, and limiting detail-- e.g., use of ranges for some sensitive variables, providing geographical codes for only highly-populated regions) no longer provide adequate protection for some of the more sophisticated demands. As a result, the Census Bureau is constantly exploring the use and development of different disclosure limitation techniques. This paper explores the possible use of a data-swapping technique.

Data-swapping was first introduced in the late 1970s. In the early 1980s, Dalenius and Reiss proved that, when used properly, the technique provided adequate protection to a microdata file without altering marginal frequency counts. This procedure has other desirable properties, namely: (1) it removes the relationship between the record and the respondent; (2) it can be used on sensitive variables, without disturbing the non-sensitive ones; (3) it can be designed to provide protection where it is most needed; and (4) it is simple to implement.

Data-swapping also has its drawbacks. We see two major disadvantages. First, arbitrary swapping can severely distort the statistics of sub-domains of the universe. This will render the file inappropriate for research and inference. Second, the implementation of more involved swapping procedures may require a substantial amount of computer time and storage.

Rank-based proximity swapping (Section VII) appears to retain all the positive attributes (i.e., easy to implement, adequate masking, etc.), while retaining a sufficient amount of the file's analytic validity. The method can be designed to preserve (1) a sufficient proportion of the multivariate dependence/independence relationships, or (2) the means (within prescribed confidence intervals) of randomly selected subsets.

In Section VIII, we note that rank-based proximity swapping can be designed to provide sufficient control of the distortion. Let $\underline{a}$ and $\underline{b}$ be two arbitrary continuous fields, which we subject to swapping. We desire to satisfy one of the two following constraints.

(1) Let $R(\underline{a}, \underline{b})$ and $R(\underline{a}', \underline{b}')$ be the correlation between the fields before and after the swap. Suppose we desire the swap to reduce the correlation by no more than a factor of R_0, $0 < R_0 < 1$. That is, $E\ (\ R(\underline{a}', \underline{b}')\ ) = R_0 * R(\underline{a}, \underline{b})$. Suppose $N(\underline{a})$ is the number of observations between the bottom- and top-coded values for field $\underline{a}$. Assume we let $P(\underline{a})$ be the maximum percentage of the difference of the ranks (i.e., the i-th ranked

value of a can be swapped with the j-th ranked value if and only if | i - j | < P(a) * N(a)). Then

$$P(\underline{a}) = 100 * \frac{\sqrt{2 * Var(\underline{a}) * (1 - R_0)}}{(a_{topc} - a_{botc})}.$$

A similar percentage can be derived for field b.

(2) Suppose we wish to swap so that each value of field a is expected to differ from its swapped value by $\pm$ K_0 times its value, (i.e., $|E(a_{i'}) - a_i| < K_0 * a_i$). Then P(a) is a function of the mean rather than the variance, namely

$$P(\underline{a}) = 100 * \sqrt{\frac{8}{3}} * \frac{K_0 * \bar{a}}{(a_{topc} - a_{botc})}.$$

Although these relationships are proven in Appendices A, B, and C under the assumption of uniform distribution, empirical testing on five relatively highly-skewed continuous variables, indicates that a rank-based proximity swap may give good results on most unimodal distributions. These results are displayed in Section XI.

Testing also indicated that such a swap could be done quickly. The time required can be approximated by the formula

$$CPU\ for\ a_i (sec.) = N(a_i) * (\frac{P(a_i)}{4} + \frac{1}{3000});$$

$$CPU\ for\ Total\ Swap (sec.) = \sum_{i=1}^{k} N(a_i) * (\frac{P(a_i)}{4} + \frac{1}{3000}).$$

Here (1) k is the number of continuous variables swapped, (2) $N(a_i)$ is the number of values that have to be swapped for the i-th continuous variable, and (3) $P(a_i)$ is the percentage of the total number of records in the swapping interval for the i-th continuous variable.

Finally, the method can be designed to provide adequate protection against even the most sophisticated intruder. Tables 6 and 7 compare the rank-based proximity swapping approach with one that adds randomly generated multivariate noise, a technique which the Census Bureau has already employed to mask a specially requested microdata file. Rank-based proximity swapping appears to provide a similar level of protection.

The rank-based proximity swapping method is easy to implement, has the ability to control the amount of distortion, and masks the data well. We strongly urge the U.S. Bureau of the Census to consider implementation of this technique to limit the risk of disclosure in future public use microdata files.

CONTROLLED DATA-SWAPPING TECHNIQUES FOR MASKING PUBLIC USE MICRODATA SETS

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

Abstract

For many years, the U.S. Bureau of the Census has collected and disseminated data. The Bureau has come to realize the importance of such data for research, analysis, planning, and policy-making. In 1963, the Bureau released a one-in-a-thousand sample file for the 1960 Decennial Census. Today, microdata files are an important part of our decennial census and demographic surveys program. The advent of the technological revolution in the early 1980s and the accessibility of personal computers to small businesses and individuals has greatly increased the demand for such data. Users now demand large and extremely detailed data sets. Unfortunately, the more information which the Census Bureau provides, the greater the risk that a user can determine the responses for some respondent.

Title 13 requires that the Census Bureau take the necessary steps to ensure the confidentiality of all respondents. Since today's data users are more sophisticated than those of the past, it is becoming difficult to provide the anonymity which the law requires. As a result, the Bureau is not always able to completely fulfill all requests. This paper gives a brief overview of the evolution of data-swapping techniques and presents a more sophisticated technique than found in the existing literature.

I. Introduction

For many years, the U.S. Bureau of the Census has recognized its obligation to collect and supply the nation's data user community with meaningful products. Some of these products include public use microdata sample files. Access to such files allow researchers to conduct important studies quickly, inexpensively, and efficiently. Without them, all specially requested tabulations would have to be conducted through a contract with the Census Bureau. This would put a tremendous strain on the programming and computer support staffs, since much of their time is committed to the day-to-day operations of producing the standard products.

On the other hand, the Census Bureau not only has an obligation to its data users but also to its data suppliers. Title 13 requires that it disseminate no product from which specific information about any particular respondent can be derived. In the case of a microdata file, this implies that data users are not able to query the file for the purpose of identifying individuals. To reduce the risk of respondent identification, it must subject each file to the appropriate disclosure limitation techniques. Many of these techniques have been in use since 1963, when the Census

Bureau released its first public use microdata set.

II. Current Disclosure Limitation Techniques

The Census Bureau currently uses several standard techniques to mask microdata sets. The first is a release of data for only a sample of the population. Intruders (i.e., those who query the file for the sole purpose of identifying particular individuals with unique traits) realize that there is only a small probability that the file actually contains the records for which they are looking. The Bureau currently releases three public use samples of the decennial census respondents. One is a 1 percent sample of the entire population, the second a 5 percent sample, and the third a sample of elderly residents. Each is a systematic sample chosen with a random start. None of these files "overlap," so there is no danger of matching to each other. Most demographic surveys are 1-in-1000 and 1-in-1500 "random" samples. Generally the public use file for each survey contains records for each respondent.

The second technique involves the limitation of detail. The Census Bureau releases no geographic identifiers which would restrict the record to a sub-population of less than 100,000. It also "recodes" some continuous values into intervals and combines sparse categories. Intruders must have extremely fine detail for other highly sensitive fields in order to positively identify targets.

The third technique protects the detail in sensitive responses in continuous fields. It is referred to as top/bottom-coding. This method collapses extreme values of each sensitive field into a single value. For example, the record of an individual with an extremely high income would not contain his exact income but rather a code showing that the income was over $100,000. Similarly the low-income records would contain a code signifying the income was less than $0. In this example $0 is a bottom-code and $100,000 a top-code for the sensitive or high visibility field of income.

III. The General Data Swapping Disclosure Limitation Technique

The accessibility of personal computers and the accompanying technology (modems and electronic data transfer) have allowed data users to handle larger and more detailed data sets than ever before. They also allow the user to compare individual records from Census Bureau-released records with those from other files available to the general public (e.g., real estate databases, information released by local governments, etc.). Today's users are demanding larger and more detailed microdata files than ever before. Users are also demanding larger samples, finer geographic detail and relaxation of the top- and bottom-codes. Disclosure limitation techniques of the 1960s may no longer sufficiently mask the anonymity of the respondents. When used in conjunction with the techniques listed above, data swapping may be a feasible procedure to adequately mask data files while providing users with more information.

One of the first references to data swapping is found in Reiss (1980). In his article, the author

describes interchanging values of individual records within a highly visible field. The swap protects the univariate distribution of that variable. Since its value belongs to some other respondent, each respondent's anonymity is protected. Consider the example below, where income is to be the sensitive field protected.

Example. A microdata file contains the age and income for 6 respondents. In order to protect the anonymity of the respondents, income values are randomly swapped among the records. Incomes on the first and sixth records, those on the second and third, and those on the fourth and fifth are pairwise swapped.

Original Responses **Responses After Swap #1**

#	Age	Income	#	Age	Income
1	21	20,000	1	21	15,000
2	24	30,000	2	24	30,000
3	35	30,000	3	35	30,000
4	36	25,000	4	36	55,000
5	45	55,000	5	45	25,000
6	50	15,000	6	50	20,000

Records 2 and 3 appear unchanged. Record 2 was swapped with a record having the same income. For these respondents, the swap has provided no masking. The probability that a swap has masked a particular record is inversely proportional to the frequency of its value appearing in the file. For large data files, this is acceptable. An income which appears frequently in a microdata file does not as easily identify the respondent as one which appears very rarely.

The releasing agency may also decide to swap more sensitive fields. In the example above, age could also be deemed as a highly visible identifier. In this case, age values would be swapped in the file "Responses After Swap #1". Suppose the age value on Record 1 is swapped with the one on Record 2, that on Record 3 with the one on Record 4, the age values on Record 5 with Record 6. The following file results.

Responses After Swap #1

#	Age	Income
1	24	15,000
2	21	30,000
3	36	30,000
4	35	55,000
5	50	25,000
6	45	20,000

The random swap of age can differ from the random swap of income. Independent multiple random data swaps can be used in succession. This further ensures that the resulting file has adequately masked accurate information about each respondent.

IV. Advantages of a Data Swap

Any data swapping procedure has the following benefits and advantages.

1. Data swapping masks accurate information about each respondent.

2. If performed on all potential key variables (i.e., variables whose values when taken together may contribute to the linking of a record with a respondent), swapping removes any relationship between the record and its respondent.

3. This procedure is extremely simple and requires nothing more than a microdata file and a random number generating routine to implement. The programming is very straight-forward.

4. The swapping procedure can be used on a select set of one (or more) variables, without disturbing the responses for non-sensitive and non-identifying fields.

5. Swapping of continuous variables provides protection when it is most necessary. Rare and unique responses are generally used to identify respondents. These values are very likely to be changed. Frequently recurring responses are less likely to be of value to an intruder and less likely to be altered by the swap. (See the income example in Section III.)

6. The procedure is not limited to continuous variables, categorical variables (such as race, sex, occupation) can also be swapped. Care must be used when swapping categorical variables; otherwise one can greatly decreases the usefulness of the file by losing the true information and creating a large number of strange combinations (such as male secretaries).

V. Disadvantages of a Data Swap

The advantages of a data swap are listed above, but any data swapping procedure has several disadvantages.

1. One disadvantage was briefly mentioned in item 6. Arbitrary categorical swaps can produce a large number of records with unusual combinations. Arbitrary swaps on continuous variables may do the same. For example, a clerk's income may be swapped with that of a brain surgeon.

2. Another is a function of the number of records in the file and the number of variables that are to be subjected to a swap. It may take a significant amount of time and computer resources to swap and store the original file and the swapped version.

3. Data swapping may significantly weaken the microdata file's analytical value. Although swapping does not affect univariate analysis on the entire population, it will affect analysis on any sub-domain (e.g., calculation of means and variances for the income of janitors). Swapping can also destroy multivariate relationships (such as, regressions and correlations between two or more variables).

VI. The Dalenius and Reiss Data Swap

Reiss (1980) realized that data swapping would mask most microdata files. However, such a procedure could destroy the analytical value and utility of the released file. Researchers, analysts, planners, policy-makers, and other data users would draw inaccurate inferences from the file. Recent research in this area has concentrated on controlling the data swap so that there would be some control over the amount of distortion introduced into the resulting file. Dalenius and Reiss (1982) defined a t-order frequency count as follows:

> Let $x_1, \ldots, x_t$ be t pre-specified variables. For any values $a_1, \ldots, a_t$, define $N(a_1, \ldots, a_t)$ to be the number of observations with $x_1 = a_1, \ldots, x_t = a_t$. The set $\{N(a_1, \ldots, a_t)\}$ over all ordered combinations $\{(a_1, \ldots, a_t)\}$ is the set of t-order statistics for $x_1, \ldots, x_t$.

The authors then proceed to show that any data swap which preserves a pre-specified set of t-order statistics will significantly reduce the risk of disclosure. Such a swap is possible with the following algorithm.

1. Two observations, x and x', are (t-1, x_1) equivalent if $x_2 = x_2'$, $x_3 = x_3'$, …, $x_t = x_t'$. The t-1 indicates that they agree on t-1 of the t key variables; x_1 indicates that records may disagree on the field x_1. Form all (t-1, x_1) equivalence classes.

2. Within each (t-1, x_1) equivalence class, randomly swap the values of x_1.

3. Repeat Steps 1 and 2 for (t-1, x_2), (t-1, x_3), …, (t-1, x_t) one field at a time.

Example.

> A data file contains the fields, Race, Occupation, and Salary. Construct a swap which preserves the 2-order frequency counts Race x Occupation (x_1 = race, x_2 = occupation, t = 2).
>
> First, swap race codes within each occupation class. The races of Clerks 1 and 4 are swapped; as well as those of Executives (Exec) 6 and 8.

Next, swap occupations within each race class. Individual 2's occupation got swapped with individual 6's, while individual 3's occupation got swapped with individual 5's.

The table below allows the reader to follow the consequences of this procedure.

Original Data					**After Race Swap**				**After Occupation Swap**			
#	Race	Occup	Salary	#	Race	Occup	Salary	#	Race	Occup	Salary	
1	W	Clerk	10000	1	B	Clerk	10000	1	B	Clerk	10000	
2	W	Clerk	12000	2	W	Clerk	12000	2	W	Exec	12000	
3	B	Clerk	11000		3	B	Clerk	11000	3	B	Exec	11000
4	B	Clerk	11000	4	W	Clerk	11000	4	W	Clerk	11000	
5	B	Exec	70000	5	B	Exec	70000	5	B	Clerk	70000	
6	B	Exec	75000	6	W	Exec	75000	6	W	Clerk	75000	
7	W	Exec	65000	7	W	Exec	65000	7	W	Exec	65000	
8	W	Exec	80000	8	B	Exec	80000	8	B	Exec	80000	

Dalenius and Reiss did not require swapping to occur for all records in each (t-1, x_i) equivalence class. They only required that records be swapped for 2 records within each class. The authors then proceeded to calculate the protection provided by such a swap, when (1) the file contains "N" individuals, (2) there are "t" key variables, (3) there are "r" individuals in each (t-1, x_i) equivalence class, and (4) "k" of these r individuals have their observations swapped.

Although frequency counts are preserved, statistics for variables outside of the pre-specified ones need not resemble the original data. In the example above, we have swapped the occupation of a clerk with that of an executive. Statistics for race-occupation have not been changed; but the distribution for occupation by income has been drastically altered. We'd like to do more than preserve frequency counts for some pre-specified set of t variables.

VII. A Rank-Based Proximity Swapping Algorithm

In an unpublished manuscript, Brian Greenberg (1987) introduced a data swap procedure for masking ordinal field data. This procedure can be used to swap values for any continuous variable. The rank-based proximity swap procedure differed from Reiss' original procedure in that it restricted the range for which each value could be swapped. Any data swap with this enhancement would definitely limit the distortion. Greenberg assumed that the ranges could be made so stringent that any statistic obtained from the resulting set should be a good estimate of the corresponding statistic obtained from the original set. Below is the algorithm which he suggested.

1. Start with a data file of size N and order responses by a single variable, a. That is, index

responses to a, by i = 1, 2, …,N; where $a_i \leq a_j$, if $i < j$.

2. Determine a value P(a), with $0 \leq P(a) \leq 100$. The intent of the procedure is to swap the value of a_i with that of a_j, so that the percentage difference of the indices, i and j, is less than P(a) of N. That is $|i - j| < P(a) * N/ 100$.

3. Initialize all ranks with a_i set to a top- or bottom-code as "swapped". Also initialize the ranks of all imputed and blank values to "swapped". Initialize all other ranks as "unswapped".

4. Let j be the lowest unswapped rank, randomly select a record with an unswapped rank from the interval [j + 1, M] where M = min {N, j + (P(a)*N/100)}. Suppose the randomly selected record has rank k.

5. Swap the values a_j and a_k. Set the labels on these to ranks to "swapped".

6. Return to Step 4 and continue until all ranks are labelled "swapped".

7. Suppose one swaps on several additional fields, b, c, … . Return to Step 1 and repeat the procedure one field at a time. First use field b, then field c, … . P(b) need not equal P(a).

8. When the swap is complete, calculate and compare multivariate statistics. If they are not within a suitable range, repeat the procedure using smaller values for P(a), and/or P(b), … .

Greenberg stopped short of guaranteeing that such a swap would preserve statistics within an acceptable error. Methods, proposed in this paper, extend rank-based proximity swap idea. We construct suitable sets from which each swap can occur. The resulting set preserves multivariate statistics within a suitable statistical error.

VIII. Enhancing the Rank-Based Proximity Swap Algorithm

The remainder of this paper concentrates on enhancing the Rank-Based Proximity Swap Algorithm. The research focuses on finding a methodology for specifying suitable values for P(a), P(b), … prior to the initial swap. Before proceeding, determine which statistics must be preserved, then concentrate on the preservation of the following two conditions:

> 1. Preservation of Multivariate Dependence/Independence. Let R(a, b) = original correlation between the values in fields a and fields b. Let R(a', b') = the correlation between the two fields after swapping. Given an $0.0 < R_0 < 1.0$, swap so that
>
> $$E[\, R(a', b')] = R_0 * R(a, b).$$
>
> Even under controlled conditions, a random swap will destroy some of the natural dependence between any two fields. Hence, R_0 must be less than 1.0. One definitely does not want $R_0 < 0.0$, otherwise he has reversed the correlation between the fields

(i.e., positively correlated fields would appear to be negatively correlated and vice-versa).

2. Preservation of Means of Subsets Which Contain a Large Number of Observations. A second desirable property of the swap is to preserve most univariate statistics (e.g., means, variances, skewness). This is particularly important for subsets which contain a large number of observations. Researchers may draw conclusions based on subsets of the microdata file. They are skeptical of inferences based on a small number of observations. This skepticism diminishes as the size of the subset increases. The easiest such statistic to preserve is the mean. Given $K_0 > 0.0$, we would like to construct a swap so that if $a_{i'}$ is swapped with a_i, then

$$E[\ |a_{i'} - a_i|/a_i\ ] = K_0.$$

For example, if $K_0 = 0.10$, then on average the swap would alter the value of each a_i by 10 percent. By the Central Limit Theorem, one would expect that a 95 percent confidence interval for the mean of a subset with N observations would be

$$(\bar{X} - \frac{2*K_0*\bar{X}}{\sqrt{N}}, \bar{X} + \frac{2*K_0*\bar{X}}{\sqrt{N}})$$

The algorithm defined below involves calculating a few initial statistics for each swap field. These are then used to set the corresponding percentages. This guarantees the swapped file will approximate the original file. It requires only the very simplistic assumption:

Let a_{botc} = value of the bottom code for field $\underline{a}$, and a_{topc} = the value of the top-code for field $\underline{a}$. Then the a_i in the interval (a_{botc}, a_{topc}) are approximately uniformly distributed.

At first glance, this assumption seems too simplistic. However, an almost identical assumption was used by Dalenius and Hodges (1959) to derive the "Cumulative Root f Rule" for an optimal construction of strata. Our empirical work with 5 skewed variables seems to indicate that such an assumption works well here also. Theorem 3 gives an estimate for $P(\underline{a})$ which preserves multivariate dependence/independence. Theorem 4 gives an estimate for $P(\underline{a})$ which preserves the univariate means of large subsets.

Lemma 1. Let $\underline{a}$ and $\underline{b}$ be two continuous fields. Suppose each value a_i is swapped with the value $a_{i'}$ and each value b_j is swapped with $b_{j'}$. Let $R(\underline{a}, \underline{a}')$ = the correlation of the values in field $\underline{a}$ before the swap to the values in field $\underline{a}$ after the swap. Define $R(\underline{b}, \underline{b}')$ similarly. Let $R(\underline{a}, \underline{b})$ = the correlation of the values in the field $\underline{a}$ to those in field $\underline{b}$ before the swap and $R(\underline{a}', \underline{b}')$ = the correlation after the swap. Appendix A shows that

$$E[\ R(\underline{a}', \underline{b}')\ ] = R(\underline{a}, \underline{a}') * R(\underline{b}, \underline{b}') * R(\underline{a}, \underline{b}).$$

Corollary 2. Given an $0.0 < R_0 < 1.0$, assume one construct swaps on fields a and b so that $R(a, a') = R(b, b') = R_0^{1/2}$; then

$$E[\, R(a', b') \,] = R_o * R(a, b).$$

Theorem 3. Given an $0.0 < R_0 < 1.0$, one can construct a swap so that $R(a, a') = R_0^{1/2}$.

Let a_{topc} and a_{botc} equal the top- and bottom-codes for field a. Let Var(a) = variance of the remaining values in field a. Appendix B shows that a reasonable estimate for P(a) is

$$P(a) = 100 * \frac{\sqrt{2 * Var(a) * (1 - R_0)}}{(a_{topc} - a_{botc})}$$

Suppose rather than preserve multivariate covariances, one desires to preserve univariate means of large subsets. The theorem below shows the relationship between P(a) and K_0.

Theorem 4. If one desires to control the swap so that each value of Field a is expected to differ from its swap value by $\pm K_0$ times its value, using the notation of Theorem 3, Appendix C shows that a reasonable estimate for P(a) is

$$P(a) = 100 * \sqrt{\frac{8}{3}} * \frac{K_0 * \bar{a}}{(a_{topc} - a_{botc})}.$$

Equate the right-hand-side expressions for P(a) in Theorems 3 and 4 to find that R_0 and K_0 spawn the following relationships:

$$R_0 = 1 - \frac{4}{3} * \frac{\bar{a}^2}{Var(a)} * K_0^2 \quad ;$$

$$K_0 = \sqrt{\frac{3}{4} * (1 - R_0)} * \frac{\sqrt{Var(a)}}{\bar{a}}.$$

Once one has specified K_0, he has also specified a value of R_0, and vice-versa. For example, A distribution has a mean of 1000 and a variance of 1500^2. A controlled ordinal swap is used with $K_0 = 0.10$. One expects

$$R_0 = 1 - (4/3) * (0.10)^2 * (1000/1500)^2 = 0.995.$$

Similarly, if a rank swap yields an R_0 of 0.99, then one expects

$$K_0 = (3/4)^{1/2} * (1 - 0.99)^{1/2} * (1500/1000) = 0.131.$$

IX. Testing the Enhanced Data Swap Theory, The Test Deck

All research was performed on the 1993 Annual Housing Survey Public Use File. From each of the 64,998 individual records, the following 10 fields were extracted:

(1) IDNUM : A 12-character identification number
(2) REGION: Region in which the housing unit was located (1, 2, 3, or 4)
(3) BEDROOMS: Number of bedrooms in the unit (0 through 10)
(4) BATHS: Number of bathrooms in the unit (0 through 10)
(5) YR_BLT: Year the unit was built (80 through 92)
(6) INCOME: Income of the household (0 through 100,000)
(7) HOME_VAL: Value of the housing unit (0 through 350,000)
(8) MORTGAGE: The monthly mortgage payment (0 through 1,800)
(9) MAINTAIN: Annual cost of maintenance (0 through 9,000)
(10) TAXES: Property taxes (0 through 62).

The IDNUM was used to link swapped and unswapped values. REGION, BEDROOMS, BATHS, and YR_BLT were used as categorical key variables. INCOME, HOME_VAL, MORTGAGE, MAINTAIN, and TAXES as continuous key variables.

The rank-based proximity swapping procedure was performed on each of the 5 continuous fields. Values, which were either not reported or which failed the top- (bottom-) code edits, were not subjected to the swap. All remaining values were. Table 1 below shows some relevant statistics for each of the continuous fields.

Table 1. Statistics for the Fields Subjected to Swapping

FIELD	# OBS[1/]	TOP-CODE	BOTTOM-CODE	MEAN[2/]	STANDARD DEVIATION[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.

X. Objectives of the Rank-Based Proximity Swap

Retention of the Covariate Relationships. This study's primary objective is to show that one is able to control the swap. The masking agent wants to swap values so that the univariate and covariate properties of the universe are retained. He uses a swapping procedure which retains all of the universe's univariate distributions. Unfortunately, this procedure destroys some of the intrinsic bivariate relationships. One hypothesizes that for a given value of R_0, $0 < R_0 < 1$, he can control the ranges from Fields a and b, on which he performs the random swaps. He constructs these ranges so that the expected post-swap correlation of values in a and b is assumed to be R_0 times the value of the original correlation, (i.e., R(a', b') = R_0 * R(a, b)). The goal is to demonstrate that, by using this method, one can come close to his correlation, R(a', b').).

Control Within Each Continuous Field. The project has several secondary objectives. If one has achieved his target correlation, R(a', b'), has he done so by controlling the distortion within each field? The test must confirm three conjectures. First, there exists a relationship of the percentage of the total number of records in each swapping interval, P(a), with the desired correlation of the swapped value with its original, R(a, a'). Second, there is also a relationship of P(a) with K_0, the average expected absolute percentage difference of the swapped value with the original. Third, R(a, a') is a good predictor of K_0, and vice-versa.

Feasibility of Implementation. Another objective examines the logistical feasibility of implementing such a swap. The procedure must be relatively easy to program. Programs must be written so that they can be readily modified for different variables and microdata files. The programs must also execute in a reasonable amount of time.

Masking Ability. A final objective examines the amount of distortion necessary to adequately mask the data. How is the amount of protection related to R_0 or K_0? We can use matching software developed by Winkler (1995) to determine the percentage of records in any masked file, which can be re-identified.

XI. Results of the Testing

Retention of the Covariate Relationships. Testing reveals that the swapping method generally yields covariances within an acceptable range of their targets. In Table 2, the pre-swap original, post-swap target, and post-swap observed correlation coefficients for the 10 bivariate combinations of fields swapped are listed. Values correspond to the factor R_0 = 0.975. One obtains the values in "POST-SWAP EXP" column by multiplying the corresponding value in the "PRE-SWAP" column by a factor of 0.975. Compare these to the actual post-swap correlation coefficients displayed in the final column.

Table 2. Comparison of the Observed with Expected Correlation Coefficients For a Data Swap with Factor, $R_0 = 0.975$

			CORRELATION COEFFICIENTS		
FIELD a	**FIELD b**	**# OBSERVATIONS[1]**	**PRE-SWAP**	**POST-SWAP EXP.**	**POST-SWAP OBS.**
INCOME	**HOME VAL**	23,318	0.150	0.146	0.141
INCOME	**MORTGAGE**	11,225	0.034	0.032	0.024
INCOME	**MAINTAIN**	14,058	0.050	0.049	0.045
INCOME	**TAXES**	22,791	0.095	0.093	0.087
HOME_VAL	**MORTGAGE**	15,514	0.607	0.592	0.595
HOME_VAL	**MAINTAIN**	17,735	0.202	0.197	0.202
HOME_VAL	**TAXES**	29,871	0.576	0.562	0.567
MORTGAGE	**MAINTAIN**	10,872	0.166	0.162	0.164
MORTGAGE	**TAXES**	15,326	0.511	0.499	0.500
MAINTAIN	**TAXES**	17,487	0.167	0.163	0.168

[1/] Number of Records on which FIELD a and FIELD b both contained a reported value that fell between the corresponding bottom- and top-code range.

In Table 2, all expected covariances differ from the observed post-swap by less than 0.008. There are only three combinations (HOME_VAL/MORTGAGE, HOME_VAL/TAXES, and TAXES/MORTGAGE) which have a correlation over 0.500. The swapping algorithm does an excellent job of hitting its target correlation for these combinations. These three combinations are of the most interest to data-users, since it seems frivolous to do regression analysis on independent variables. The swap achieves its objective. It preserves bivariate independence and reduces the correlation of highly dependent variables by a pre-defined factor of R_0.

Use Table 3 to compare the expected values to observed post-swap correlations for the HOME_VAL/MORTGAGE, HOME_VAL/TAXES, and MORTGAGE/TAXES combinations. The listed values correspond to $R_0 = 0.975$, 0.950, and 0.900. Expected post-

swap correlations are also calculated for these combinations when a swap is constructed to yield an average absolute percentage difference of 10 percent (i.e., $K_0 = 0.10$). In the latter instance, the target factor R_0 is calculated by multiplying the correlation (corresponding to $K_0 = 0.10$ (See Theorem 4.)) for FIELD a with the corresponding correlation for FIELD b (i.e., $R_0 = R(a, a') * R(b, b')$).

Table 3. Observed versus Expected Correlation Coefficients of Highly Correlated Combinations of Continuous Variables For Various Target Values of R_0

				CORRELATION COEFFICIENTS		
FIELD a	**FIELD b**	**# OBS.[1/]**	**R_0**	**PRE-SWAP**	**POST-SWAP EXP.**	**POST-SWAP OBS.**
HOME_VAL	**MORTGAGE**	15,514	0.975	0.607	0.592	0.595
		15,514	0.950	0.607	0.577	0.590
		15,514	0.900	0.607	0.541	0.560
		15,514	0.973[2/]	0.607	0.590	0.591
HOME_VAL	**TAXES**	29,871	0.975	0.576	0.562	0.567
		29,871	0.950	0.576	0.548	0.555
		29,871	0.900	0.576	0.518	0.538
		29,871	0.970[2/]	0.576	0.556	0.558
MORTGAGE	**TAXES**	15,326	0.975	0.511	0.499	0.500
		15,326	0.950	0.511	0.486	0.487
		15,326	0.900	0.511	0.460	0.467
		15,326	0.971[2/]	0.511	0.495	0.496

[1/] Number of Records on which FIELD a and FIELD b both contained a reported value that fell between the corresponding bottom- and top-code range.

[2/] These factors are calculated by multiplying the predicted factors for the two fields which corresponds to $K_0 = 0.10$.

Table 3 indicates that our swapping procedure adequately predicts the post-swap correlation for

large values of R_0 (namely, $R_0 \geq 0.950$). This precision diminishes as R_0 drops from 0.950 to 0.900. One now has to address the reason for the diminished accuracy of the estimator. Attribute a large portion of this problem to the loss of control of the swap within one or more of the continuous variables (i.e., R(a, a') differs significantly from $R_0^{1/2}$.).

Control Within Each Continuous Field. The original problem states,

> "Given R_0, can we determine appropriate values of P(a) and P(b), (i.e., the maximum number of observations in each swapping set)?"

One "solves" this problem by constructing appropriate percentages based on R_0. If the expected value of the post-swap correlation, R(a', b'), differs significantly from the original correlation diminished by a factor of R_0 (i.e., R_0 * R(a, b)), then the problem may occur in one of the following three assumptions:

(1) $E[\ R(a', b')\ ] = R_0 * R(a, b)$;

(2) Theorem 3 does not adequately estimate P(a) for certain values of R_0; or

(3) Theorem 3 does not adequately estimate P(b) for certain values of R_0.

See Appendix A for a proof that Assumption (1) holds. The results of Tables 2 and 3 also support this. Let's concentrate attention to the validity of P(a) and P(b). Recall that the construction of P(a) hinges on one simple, but very crucial, assumption, "The values of Field a are uniformly distributed between the bottom- and top-code for a." How uniform are these distributions?"

All five of these continuous variables are skewed. In uniform distributions, the standard deviation would be approximately 28 percent of the interval length. Use Table 1 to calculate these ratios for each field. They are approximately

INCOME …… 14 percent,
HOME VALUE … 18 percent,
MORTGAGE ……. 20 percent,
MAINTAIN ……. 9 percent, and
TAXES …… 21 percent.

The most highly skewed fields are INCOME and MAINTAIN; while the other three are relatively uniform. Expect P(a) to be a worse approximation of interval length for MAINTAIN than for TAXES. Table 4 confirms this. Suppose one desires a target correlation of the swapped to the unswapped value of 0.987. For MORTGAGE, TAXES, and HOME VALUE, this method yields correlations which fall between 0.982 and 0.985. Notice the observed post-swap correlations for the other variables. INCOME has a post-swap correlation

of 0.955, while MAINTAIN has one of 0.924. One cannot dismiss the importance of a distribution's skewness. Should he abandon this hypothesis, since it was based on an erroneous assumption?

For "near uniform" distributions, the theory does a good job of estimating P(a) for large values of R_0. Yet it fails as R_0 decreases. Examine the proofs in Appendix B. It is very important that the distribution be nearly uniform on each swapping interval. The range of the swapping interval increases as R_0 decreases. For values of R_0 near 1.00, the ranges will be very small and

Table 4. Results of the Rank-Based Proximity Swap Test

			CORRELATION BETWEEN SWAPPED AND ORIGINAL VALUE, R(a, a')		AVERAGE ABSOLUTE PERCENTAGE CHANGE, 100*|a-a'|/a	
FIELD	**PCT OF RECS IN SWAP INT**	**CPU TIME (MIN: SEC)**	**EXP.**	**ACTUAL**	**EXP.**	**ACTUAL**
INCOME	3.1	5:15	0.987	0.955	16.79	19.70
	4.4	7:26	0.974	0.930	23.89	27.56
	6.2	10:18	0.948	0.895	33.58	39.52
	1.8	3:07	0.996	0.975	10.00	11.43
HOME VALUE	4.4	5:15	0.987	0.982	9.44	11.62
	6.3	7:17	0.974	0.967	13.35	16.79
	8.8	9:54	0.948	0.943	18.88	23.24
	4.5	5:56	0.986	0.981	10.00	13.21
MORTGAGE	4.4	1:32	0.987	0.985	8.10	8.84
	6.3	2:09	0.974	0.973	11.45	12.84
	8.8	2:57	0.948	0.953	16.20	17.45
	5.5	1:59	0.984	0.980	10.00	13.18
MAINTAIN	2.0	0:59	0.987	0.924	19.74	8.84
	3.1	1:32	0.974	0.893	27.93	11.34
	4.0	1:52	0.948	0.847	39.48	15.39
	1.1	0:48	0.996	0.983	10.00	4.72

			CORRELATION BETWEEN SWAPPED AND ORIGINAL VALUE, R(a, a')		AVERAGE ABSOLUTE PERCENTAGE CHANGE, 100*|a-a'|/a	
FIELD	**PCT OF RECS IN SWAP INT**	**CPU TIME (MIN: SEC)**	**EXP.**	**ACTUAL**	**EXP.**	**ACTUAL**
INCOME	3.1	5:15	0.987	0.955	16.79	19.70
	4.4	7:26	0.974	0.930	23.89	27.56
	6.2	10:18	0.948	0.895	33.58	39.52
	1.8	3:07	0.996	0.975	10.00	11.43
HOME VALUE	4.4	5:15	0.987	0.982	9.44	11.62
	6.3	7:17	0.974	0.967	13.35	16.79
TAXES	4.6	5:52	0.987	0.983	8.89	10.75
	6.5	8:09	0.974	0.970	12.56	14.85
	9.2	11:11	0.948	0.949	17.78	20.68
	5.2	6:36	0.982	0.983	10.00	11.79

distribution will be "almost" uniform of most swapping intervals. As R_0 decreases, the intervals become bigger and less uniform.

For highly-skewed distributions, P(a) can be accurately determined only at relatively high values of R_0. For INCOME, $R_0 = 0.975$ ($R(a, a') = R_0^{1/2} = 0.987$) is too low. For less-skewed distributions, such as TAXES, $R_0 = 0.900$ and $R(a, a') = 0.948$ give remarkably good estimates. A good topic for future research will be the quantitative relationship between the skewness of the distribution and the ability of R_0 to give an accurate prediction of the appropriate number of percentage of the total number of observations in a swapping interval, P(a).

Recall that K_0 is the average percentage change induced by the rank swap. Table 4 shows that the theory provides good predictions for the relationship between K_0 and P(a). Again, the distribution of each swapping interval is assumed to be approximately uniform. For relatively small swapping intervals, the K_0-estimator is an extremely good predictor. It even compensates for a "moderately" skewed distribution such as INCOME (Target $K_0 = 10$, Observed $K_0 = 11.43$). However, for very extremely skewed distributions, such as MAINTAIN, even the K_0 estimator fails miserably (Target $K_0 = 10.00$, Observed $K_0 = 4.72$.). As the theory predicts, P(a) and the observed value of K_0 are directly proportional. Compare values (within each field)

of the "PCT OF RECORDS IN SWAPPING INTERVAL" column with the "AVERAGE ABSOLUTE PCT CHANGE/ OBS."column from Table 4. The columns are almost directly linearly correlated. When P(a) doubles, the observed value of K_0 approximately doubles.

Also use Table 4 to confirm the inverse relationship between the observed correlation, R(a, a') and the corresponding observed value of K_0. This is a logical consequence of the validity of Theorems 3 and 4. For pre-defined values of K_0 which correspond to values of R_0 near 1.00 (e.g., $R_0 \geq 0.95$), one can construct a sampling interval for which the observed correlation is near the target. For less skewed distributions, values for this corresponding R_0 can extend to 0.900 and lower. The theory holds. Is implementing such a swap feasible?

Feasibility of Implementation. The procedure is not difficult to program. For testing purposes, the author wrote and executed a SAS modular program by which the procedure was tested. All information in Tables 1 through 4 was generated by this program. The modules and the SAS procedures used are listed below.

Module 1: Determine the bottom- and top-codes for CONTINUOUS VARIABLE #1. (Use PROC MEANS.)

Module 2: Create a data set by stripping off the IDNUM and the value for CONT_VAR_#1. Exclude records where the value for CONT_VAR_#1 is missing, bottom-, or top-coded (Use DATA STEP with KEEP = option).

Module 3: Calculate the record count, mean, and standard deviation of the new set. (Use PROC MEANS).

Module 4: Sort the data set by CONT_VAR_#1. (Use PROC SORT.)

Module 5: Swap the data. (Use a DATA step.) This is the only module which requires some involved programming. It also requires the most Central Processing Unit (CPU) time to execute. In this program, the user sets the following parameters:

(1) Record Count of the set,
(2) Target Factors, R_0 and K_0,
(3) Mean of CONT_VAR_#1,
(4) Standard Deviation of CONT_VAR_#1,
(5) Bottom-code of CONT_VAR_#1, and
(6) Top-code of CONT_VAR_#1.

The program

(1) calculates P(a), by use of a programmed formula,
(2) chooses an appropriate random number,
(3) randomly swaps the data in the prescribed manner, and
(4) sets the swap flags.

Module 6: Sort the output set of the previous module by IDNUM. (Use PROC SORT.)

Repeat Modules 1 through 6 for CONT_VAR_#2, CONT_VAR_#3, … .
Module 7: Combine all information from the output sets of Module 6. (Use DATA step with a MERGE BY IDNUM.)

Module 8: Analyze the correlation coefficients of the continuous variables after the swap. (Use PROC CORR on the set produced in Module 7.)

Module 9: Produce the correlation coefficients for the original set (Use PROC CORR on the original set.) Manually compare the two. (For files with a large number of continuous variables, use PROC COMPARE.)

If the user is satisfied with the results of the swap, invoke Module 10.

Module 10: Update the values of continuous variables. (Use a DATA step with UPDATE statement

Even though the code for Module 5 is the most difficult to write, it is extremely easy to modify. The user can easily change any combination of the following.

(1) To execute for a different value of the Diminishing Factor, R_0 (or K_0), reset the value of the parameter R_0 and re-execute the program.

(2) To execute for a different continuous variable, change the parmeters: set count, mean, standard deviation, bottom-, and top-code. Re-execute.

(3) To change the method by which P(a) is calculated, modify the line of code with the formula, then re-execute. Use this when some users desire P(a) as a function of K_0, others as a function of R_0.

Module 5 is easy to modify. Even a novice SAS programmer can maintain and execute this procedure. Before implementation, there is still work to be done. At present the parameters must be hard-coded from the output of Module 3 to Module 5. An experienced SAS programmer can code the routine so that this update is automatically executed. The user must also restart the program for each continuous variable. Through the clever use of macro-variables and macro-programs, a user should be able to pre-specify all continuous variables, then start the execution. The program would not stop until the completion of Module 9. After verifying the acceptability of the swap, the user could then invoke Module 10. It is also probable that more efficient code could be written in another environment (Unix, C, etc.).

The programming code exists which is easy to use and modify. Does this code execute in a relatively short amount of time? Table 4 shows the CPU time required to execute the swap (Module 5). For example, a $K_0 = 0.10$ swap for INCOME took 2 minutes and 17 seconds. The $K_0 = 0.10$ swap for all variables took 18 minutes and 26 seconds (3:07 + 5:56 + 1:59 + 0:48 + 6:36). A multivariate regression of the ratio CPU time/Total Record Count (CPU/$N(\underline{a_i})$ to $P(\underline{a_i})$ yielded the following formulas.

$$CPU\ for\ a_i(\text{sec.}) = N(a_i) * \left(\frac{P(a_i)}{4} + \frac{1}{3000}\right);$$

$$CPU\ for\ Total\ Swap(\text{sec.}) = \sum_{i=1}^{k} N(a_i) * \left(\frac{P(a_i)}{4} + \frac{1}{3000}\right).$$

The CPU time required is a function of (1) k, the number of continuous variables swapped, (2) $N(a_i)$, the number of values that have to be swapped for each continuous variable, and (3) $P(a_i)$, the percentage of the total number of records in each swapping interval. Users may find this formula very useful. If either k or some of the $N(a_i)$ are extremely large, the process could require a substantial amount of CPU time.

Masking the Data. Testing indicates that an ordinal rank swap masks the microdata file as well as the technique of adding independent randomly-generated noise of Paas (1988) and Kim (1986). The testing technique assumed the following.

(1) An intruder could construct a target file of relatively unique individuals, some of whom he was "almost certain" would be contained in the universe file. This target file would contain 600 to 1,000 individuals.

(2) For each record in the target file, the intruder would have non-sensitive information (e.g., the physical location of the housing unit, the number of bathrooms and bedrooms which it contained, and the year in which it was built), which he believed to be very reliable. The intruder is slightly skeptical about certain values in some of these non-sensitive fields.

(3) For each targeted observation, the intruder was also able to make reasonably accurate guesses (within 10 percent of the true value for five sensitive items (household income, home value, mortgage payment, annual maintenance, and property taxes).

(4) Because the intruder has accurate information on non-sensitive items, he can restrict the universe to less than 20,000 observations.

(5) The intruder has a very sophisticated matching software program. The software will link each record from the target file to the "most likely" match in the restricted universe. The intruder has a pre-defined criteria for which linkages

are definitely re-identifications, which are suspect, and which are not re-identifications. For definite re-identifications, it is important to the intruder that he obtain accurate information for all five of the sensitive values. The intruder may realize that the values in the restricted universe have been perturbed, but he does not know to what extent.

(6) The intruder has sufficient knowledge of the software, and the target, and restricted universe files to accurately set the required matching parameters.

The Restricted Universe. The intruder first constructs the restricted universe by blocking the observations into equivalence classes. Two observations are in the same equivalence class if they contain the same values for BATHS, BEDROOMS, YR_BUILT, and REGION. Restrict the universe to only those classes with less than 100 observations. This would produce a file of 18,557 observations.

The Target File. The intruder's target file contains all observations in the equivalence classes with 2 or less observations. This file contained 771 observations.

Introduction of Some Intruder Skepticism. After the construction of the target and restricted universe files, the intruder becomes skeptical of certain values in the non-sensitive variables. Without reconstructing either file, he decides to collapse certain equivalence classes. For (programming) simplicity, assume the region code was inaccurate. Collapse these files accordingly. Two records are now equivalent, if they contain the same values for BATHS, BEDROOMS, and YR_BLT. The target file now has equivalence class blocks containing 7 or less observations. The restricted universe has blocks containing 213 (as compared to 100) or fewer records.

The Matching Software. The intruder has software which utilizes the matching algorithm developed by Fellegi and Sunter (1969). He also posesses software similar to that of Winkler (1988), he can calculate a reasonably matching and non-matching "weights" for the sensitive variables.

The Expectation-Maximization (EM) software of Winkler (1988) obtains good estimates for a set of weights, which would do the best job re-identifying the true corresponding record in the restricted universe. This software independently compares values of each variable in the target file to the corresponding variable in the universe.

If the typical value of that variable has a large number of possible matches in the universe, the EM algorithm assigns a variable a low positive weight for matching cases. If the typical value in that field has a limited number of possible matches, it assigns a large positive value. Negative weights are assigned for mismatch weights. Large negative numbers indicate that

there are a limited number of possible mismatches for the typical value. Small negative weights, indicate that the typical value has many possible mismatches.

In our testing, we have access to a unique identifier, which is attached to each record in the target and universe files. Thus, we are able to positively determine whether the weights were the optimal for discrimination. In actuality, the intruder would not have this information available to him. He would derive a less optimal set of weights, this would cause more "definite" re-identifications, which were incorrect. Weights will differ between the perturbed versions of the universe. Fortunately, for the intruder, the optimal weights did not differ significantly from the form of perturbation used. Table 5 gives the approximate optimal weights.

The Fellegi-Sunter routine independently matches each field on the target file to records on the restricted universe file. This match is restricted to those records in the universe which have the same values as the key variables (BATHS, BEDROOM, YR_BLT) of the target file. The appropriate positive/negative weight from Table 5 is obtained for each variable. Each record in the universe is assigned a value equal to the sum of the 5 weights (one weight for each sensitive field). The record in the universe with the largest aggregate weight is linked to the target.

Table 5 shows that HOME VALUE is a good discriminator of true re-identification. For the typical value of HOME VALUE, there are a limited number of cases which are possible matches and non-matches. INCOME, on the other hand, is a less adequate discriminator. For the typical INCOME value, there are many possible matches and mismatches.

Criteria for Definite, Suspect, and Erroneous Re-identifications. The intruder is interested in accurately identifying all five sensitive values. If one of the five is suspect, he questions the re-identification. In addition, if two or more of the values are suspect, he dismisses the linkage as a case for which no suitable match is found.

If the sum is greater than +10.0, then all five fields match within 10 percent of the intruder's expectation. In this case, he assumes that he has a positive re-identification.

Table 5. Optimal Matching/Non-Matching Weights For the EM Procedure For the Rank-Based Proximity Swap Version of the 1993 AHS Microdata File

Field	If values agree within 10 Percent	If values disagree by more than 10 percent
INCOME	+1.8	-0.6
HOME VALUE	+3.0	-3.6
MORTGAGE	+1.2	-5.4

MAINTAIN	+ 1.9	-3.2
TAXES	+ 3.0	-4.0

If the weight is less than 0.0, then at least 2 of the latter 4 variables (INCOME is excluded, since it is a poor discriminator) do not match within 10 percent of the intruder's expectation. He therefore assumes the linkage to be incorrect.

Any case with a weight between 0.0 and 9.9 is assumed to be a "Questionable Linkage."

It is possible for the algorithm to assign a weight greater than 10.0 to an incorrect linkage. Here, there exists a record in the restricted universe on which all five of the values lie within 10 percent of the intruder's expectations for the target. When this occurs, the intruder will incorrectly believe that he has re-identified his target. Refer to these as "Incorrect Re-Identifications."

Table 6 below displays the masking effect of each of the rank swap (i.e., $K_0 = 0.10$, $R_0 = 0.975$, $R_0 = 0.950$, and $R_0 = 0.900$). Since the $K_0 = 0.10$ generally gives smaller swapping intervals than $R_0 = 0.975$, the value of $R_0 = 0.990$ is used as an approximation for the corresponding correlation coefficient.

Analysis of Masking Results. Refer to Table 6. The rank swap appears to effectively mask the microdata file. When $R_0 = 0.975$, only 100 (or 13 percent) of the 771 target records were able to be correctly re-identified. This compares favorably with the results from the Paas (1988) simulation, in which he correctly re-identified 20 percent of his target file. The intruder "re-identifies" 152 records, but 52 (or about 1 out of every 3) are incorrect. Only marginal gains are achieved by reducing the factor below 0.950. The decrease is slight (142 "re-identifications", 89 correct, 53 incorrect), as the target level of R_0 is decreased to 0.900.

To the layman analyst, 13 percent re-identification may seem unsatisfactorily high. However, keep in mind the special circumstances under which the simulation is conducted.

(1) The intruder is certain that each target respondent exists on the microdata file.

Table 6. Ability of an Intruder to Re-identify Microdata Masked by a Rank-Based Proximity Swapping Routine 10 Percent Difference Allowed Between Values on Target and Masked Files

Swapping Method Target R_0	~ 0.990 ($K_0 = 0.10$)	0.975	0.950	0.900

Correct Re-Ids	124	100	90	89
Incorrect Re-Ids	54	52	51	53
Questionable Linkages	378	266	236	204
False Linkages	215	353	394	405

(2) The target file is an offspring of the file used to create the perturbed universe. Values were not generated as the result of two independent surveys. As a result, respondents do not give inconsistent answers (e.g., respondent gives gross income figure for one survey and take-home income for another), there are no keying errors, and no user interpretation errors (e.g., one survey gives (current) home value, the other gives the price at which the house was purchased several years ago).

(3) The software used is as sophisticated as any possessed by the intruder.

(4) The set of weights calculated are optimal for discrimination. If an intruder had to "guess" weights, his matching accuracy would diminish.

Muller, et al (1995) discuss Points (1) and (2) in detail. For these reasons, they are not surprised by the high re-identification rate which Paas achieved. Point (3) can be challenged. Winkler's software is available on the internet. Also, the methodology, from which the program was constructed, exists in the professional literature. Some intruders may have better software. Note that Winkler's software is extremely thorough and effective. It required several years of intense research, combining the results of many noted matching professionals. The validity of Point (4) is hard to evaluate. For this simulation, the weights were not very sensitive to the values in the perturbed file. If an intruder uses Winkler's program, he will do no better than the results given in Table 6.

Comparison With the Addition of Random Noise. Kim (1986) suggested a method of adding random noise to sensitive data. Like the rank swap, it has the ability to predict the amount by which bivariate covariances are diminished. Kim's theorem is as follows.

Suppose ($\underline{a}$, $\underline{b}$) is multivariate normal distribution with standard deviations STD($\underline{a}$) and STD($\underline{b}$). Suppose also that the correlation coefficiecnt of $\underline{a}$ to $\underline{b}$ is R($\underline{a}$, $\underline{b}$). Suppose we have a multivariate noise distribution ($\underline{na}$, $\underline{nb}$) where the means of $\underline{na}$ and $\underline{nb}$ are 0, STD($\underline{na}$) = c * STD($\underline{a}$), STD($\underline{nb}$) = c * STD($\underline{b}$), and R($\underline{na}$, $\underline{nb}$) = R($\underline{a}$, $\underline{b}$). For each element (a_i, b_i) in ($\underline{a}$, $\underline{b}$) randomly choose a value (na_i, nb_i) in ($\underline{na}$, $\underline{nb}$). Let $a'_i = a_i +$

na_i and $b'_i = b_i + nb_i$ in ($\underline{a}'$, $\underline{b}'$), then

$$VAR(\underline{a}') = VAR(\underline{a}) * (1 + c^2),$$
$$VAR(\underline{b}') = VAR(\underline{b}) * (1 + c^2), \text{ and}$$
$$COV(\underline{a}', \underline{b}') = COV(\underline{a}, \underline{b}) * (1 + c^2).$$

Kim is using noise to expand the variance-covariance structure. The regression coefficients are unaltered by the noise. The approach of this paper preserves the variance structure, but contracts the covariance structure by a factor of $1 / (1 + c^2)$. One is able to calculate variance-covariance structures for the five variables in question. By using the SUN-UNIX subroutine RNMVN, he can generate values for random noise, (n_{1i}, n_{2i}, …, n_{5i}), from a distribution with the desired variance-covariance structure. Add these values to the original values to produce a distribution with COV ($\underline{a}'$, $\underline{b}'$) = S_0 * COV ($\underline{a}$, $\underline{b}$).

Note that S_0 (for Kim's method) like R_0 (for the method presented here) is the expected value of the ratio of the largest covariance to the smallest. Whereas Kim's method expands the perturbed covariances, this method diminishes them. Never the less, this appears to be a logical measure by which the masking power of the two methods can be compared.

Let $R_0 = S_0^{-1}$. To compare the protection afforded by this masking technique to that afforded by the rank swap technique, a random noise approach is used to construct masked versions of the universe for values of R_0 = 0.990, 0.975, 0.950, and 0.900. These sets were then subjected to the EM and Fellegi-Sunter routines. Table 7 shows the masking ability of random noise.

At all levels of R_0, the rank swap and the addition of random noise provide about the same amount of protection. Look at the R_0 = 0.975 level. Both swaps re-identify about 100 observations correctly and about 50 incorrectly. When the rank swap is used, there are less questionable linkages (266 to 328) and more false (353 to 286) linkages. Similar comparisons can be made at the other levels of R_0. Like the rank swap, little extra protection is gained by using values of R_0 below 0.950. These results re-enforce the conjecture, "For this microdata file,a value of R_0 near 0.950 is the optimal level of distortion."

Table 7. Ability of an Intruder to Re-identify Microdata Masked by Kim's Random Noise Procedure 10 Percent Difference Allowed Between Values on Target and Masked Files

Target R_0	**0.990**	**0.975**	**0.950**	**0.900**

Correct Re-Ids	126	103	96	92
Incorrect Re-Ids	52	54	54	55
Questionable Linkages	411	328	280	230
False Linkages	191	286	342	394

XII. Future Research Topics

Below are listed a few areas which the author feels may be fruitful for future research.

1. Development of a Better Estimator for P(a), the Percentage of Observations in the Swapping Interval. In Appendix B, we estimated the relationship between P(a) and R(a, a'), the target correlation of each value in field a after swapping to its pre-swap value. Assuming a to be uniformly distributed, we achieved a relatively accurate estimate for P(a). Can we do better? Is it possible to derive a more exact estimate by using the higher moments (e.g., the third moment which measures the skewness)?

2. Examination of P(a) for Non-Monotonic Decreasing Distributions. All five continuous distributions in this rank test were monotonic decreasing. Does this method of estimation give more exact results with uni-modal distributions that are less skewed (e.g., normally distributed fields)? How do we handle multi-modal distributions?

3. Examination of Alternative Constructions of the Swapping Interval. The interval used in this research was constructed by using the results of Corollary B.2. in Appendix B. It assumes that if R(a, a') is the target correlation and Var(a) is the variance of a, and if e_i is the difference between the i-th swapped value and R(a, a') times the original value, then $Var(e_i) = (1 - R^2(a, a')) * Var(a)$.

   The variance of the e_i are estimated by assuming the a_i are uniformly distributed. Are there other methods of calculating or estimating the variance of the e_i? Should we use an iterative process which calculates the $Var(e_i)$ directly? Should we stratify and swap within each stratum? Should we stratify and permute within each stratum?

4. Development of a Rank Swap for Categorical Variables. The rank swap works well for continuous variables because they can be sorted in ascending or descending order. Is there some analogous method to rank categorical variables? Should we attempt to develop a metric from R^n to R^1; use this metric on the n sensitive continuous variables to stratify the microdata file; then swap categorical variables within stratum? Should we

first block all categorical variables, then swap entire blocks?

5. Analysis of a Rank Swap on Statistics of Sub-Domains. This research shows that there is a "predictable" relationship between R_0 and K_0. In Section VIII, it was hypothesized that bias of the mean of a subset was directly proportional to K_0. No simulations were done to prove or disprove this conjecture. The author's "gut" feeling is that the conjecture is true for random subsets. It probably does not hold for non-randomly constructed subsets (e.g., houses with more than 8 bathrooms, or units valued over $250,000). Can we quantify the bias as a function of K_0 and the frequency with which each observation (in the subset in question) appears in the universe?

6. Development of Intruder Simulation Software. Paas (1985), Winkler (1988, 1995), and Fienberg, et al. (1995) have developed matching software. All used the software to re-identify individual respondents in "masked" data sets. We used Winkler's software to simulate an intruder in this project. Is our model reasonable? Should we set up some guidelines and develop the software to simulate a sophisticated intruder? What constitutes a reasonable target set? What constitutes a re-identification? What percentage of re-identifications are tolerable for a public use file?

XIII. Conclusions

The technological revolution of the 1980s has allowed data-users to handle larger and more detailed data sets than ever before. This has been accompanied by the public's easy access to very sophisticated matching software. As a result, government agencies have been compelled to release only microdata files with very limited detail. Severe restrictions have been placed on the sampling fraction of the universe, the amount of geographic detail, and the range of values for continuous data released. Users have found this extremely irritating and unacceptable.

Data swapping appears to be a feasible alternative to severe top- and bottom-codes. This technique was first suggested by Reiss in the early 1980s, and has evolved significantly. In the past, the Bureau has shied away from its use. It has been argued that it not only masks the file, but also diminishes its multivariate analytical utility. This paper illustrates that an effective version of the data swap exists which controls the amount of distortion induced.

The Bureau has used the addition of random noise to mask microdata files. This also distorts the data in a "controlled" manner. Both techniques measure the amount of distortion by comparing the bivariate covariances of the data after the swap to the corresponding values before the swap. For a pre-specified amount of distortion, this research shows that data-swapping protects individuals identities as well as (if not better than) the addition of random noise.

In addition, the data-swapping technique suggested here should be relatively quick to code and easy to modify and enhance. Testing has shown that, under reasonable circumstances, it does

not take a long time to execute. For files with many sensitive continuous fields, or those with a large number of observations, this routine may require too much computer resources. However, a formula has been provided from which a reasonable estimate of the expected number of CPU seconds (on a VAX cluster) can be obtained.

Should the Bureau consider rank swapping a potentially feasible method to mask microdata files, there is a plethora of directions in which the research can proceed. These range from improving our estimates for the appropriate swapping interval to the development of an intruder model.

This method is quick, has the ability to control the distortion, and masks the data well. I strongly urge the Bureau to consider implementation of this technique to limit the risk of disclosure in future public use microdata files.

XIV. References

1. Cochran, W. (1977). Sampling Techniques (3rd edition), New York: John Wiley and Sons.

2. Dalenius, T. and Hodges, J. L. Jr. (1959). Minimum Variance Stratification. Journal of American Statistical Association **54**, 88-101.

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

4. Feinberg, 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 50-YABC-2-66205, Task Order 8, Activity 1.

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

6. Greenberg, B. (1987), Rank Swapping for Masking Ordinal Microdata, US Bureau of the Census (unpublished manuscript).

7. Greenberg, B. and Voshell, L. (1990). The Geographic Component of Disclosure Risk for Microdata. Statistical Research Division Report Series **Census/SRD/RR-90/13**, US Bureau of the Census.

8. Kim, J. J. (1986). A Method For Limiting Disclosure in Microdata Based on Random Noise and Transformation. Proceedings of Survey Research Methods Section, American Statistical Association, 303-308.

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

10. Paas, G. (1988). Disclosure Risk and Disclosure Avoidance for Microdata, Journal of Business and Economic Statistics, **6**, 487-500.

11. Reiss, S. P. (1980). Practical Data-Swapping: The First Steps, IEEE Symposium on Security and Privacy, 38-43.

12. Subcommittee on Statistical Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology (1994). Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22, Office of Manangement and Budget, Washington, DC.

13. Winkler, W. E. (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage, Proceedings of the Survey Research Methods Section, American Statistical Association, 667-671.

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

Appendix A. Bias Introduced On the Correlation Coefficient by Independent Ordinal Swaps

Summary. Assume the value of a_i is swapped with $a_{i'}$ and that of b_i with $b_{i'}$. Suppose the method used to swap $\{a_i\}$ is independent of that used to swap $\{b_i\}$. Moreover, assume the correlation coefficient between the pairs $(a_i, a_{i'})$ is R(a, a') and that between the pairs $(b_i, b_{i'})$ is R(b, b'), if R(a, a') and R(b, b') are approximately 1 ,then

$$E[\ COV\ (a', b')\ ]\ =\ R(a, a')\ *\ R(b, b')\ *\ R(a, b).$$

For $R(X)\ +\ R(Y)\ >\ 1.5$, a good approximation of this is

$$E[\ COV\ (a', b')\ ]\ =\ COV(a, b)\ *\ (\ R(a, a')\ +\ R(b, b')\ -\ 1).$$

Lemma A.1. Let $b_i\ =\ m\ *\ a_i\ +\ c\ +\ e_i$, where m and c are chosen to minimize

$$\sum e_i^2,$$

then

$$(1)\sum e_i*(a_i-\bar{a})=0,$$
$$(2)\sum e_i=0.$$

Proof. Let

$$z=\sum e_i^2=\sum\ (b_i-m*a_i-c)^2.$$

Take the partial derivative of z with respect to m, then take the partial derivative of z with respect to c. Set both partials to 0. Solving simultaneously gives the desired results.

Note : Throughout the remainder of this appendix the line, $b_i\ =\ m\ *\ a_i\ +\ c\ +\ e_i$, refers to the line-of-best-fit (i.e., the one which minimizes the sum of the squares of the e_i).

Theorem A.2. Assume $a_{i'}$ is swapped with a_i.

Let $(a_{i'}\ -\ \bar{a})\ =\ m\ *\ (a_i\ -\ \bar{a})\ +\ c\ +\ e_i$, then

(1) m = R(a, a') the correlation coefficient of a with a', and
(2) c = 0.

Proof. Let

$$z=\sum [(a_{i'}-\bar{a})-m*(a_i-\bar{a})-c-e_i]^2.$$

Take the partial with respect to c and divide by -2.

$$\left(\frac{z_c}{-2}\right)=\sum (a_{i'}-\bar{a}) \\ -m*\sum (a_i-\bar{a}) \\ -c$$

The first two summations on the right-hand side are 0. Since $z_c = 0$, this forces $c = 0$.

Take the partial of z with respect to m, then divide by 2n.

$$\left(\frac{z_m}{2n}\right)=\left(\frac{1}{n}\right)*\sum (a_{i'}-\bar{a})*(a_i-\bar{a}) \\ -m*\left(\frac{1}{n}\right)*\sum (a_i-\bar{a})^2 \\ -\left(\frac{c}{n}\right)*\sum (a_i-\bar{a})$$

The last term in the summation is 0. Since $z_m = 0$ and $c = 0$, the first and second terms on the right-hand side are equal. This implies COV(a, a') = m VAR(a). Divide both of these terms by VAR(a) to get the desired result of R(a, a') = m.

Therefore the line-of-best-fit is $(a_{i'} - \bar{a}) = R(a, a') * (a_i - \bar{a}) + e_i$.

Theorem A.3. Suppose $a_{i'}$ is the value swapped for a_i and that each b_i is not swapped, then

$$R(a',b)=R(a,a')*R(a,b)+\left(\frac{1}{n}\right)*\sum e_i*(b_i-\bar{b})$$

Proof.

$$R(a',b) = \left(\frac{1}{n}\right) * \sum (a_{i'} - \bar{a}) * (b_i - \bar{b})$$
$$= \left(\frac{1}{n}\right) * \sum [R(a,a') * (a_i - \bar{a}) + e_i] * (b_i - \bar{b})$$
$$= R(a,a') * R(a,b) + \left(\frac{1}{n}\right) * \sum e_i * (b_i - \bar{b})$$

Theorem A.4. Suppose $a_{i'}$ is the value swapped for a_i and $b_{i'}$ is the value swapped for b_i. Suppose

$$(a_{i'} - \bar{a}) = R(a,a') * (a_i - \bar{a}) + e_i$$
$$(b_{i'} - \bar{b}) = R(b,b') * (b_i - \bar{b}) + f_i$$

Then

$$R(a',b') = R(a,a') * R(b,b') * R(a,b)$$
$$+ \left(\frac{1}{n}\right) * \sum e_i * (b_{i'} - \bar{b}) + \left(\frac{1}{n}\right) * \sum f_i * (a_{i'} - \bar{a}) - \left(\frac{1}{n}\right) * \sum (e_i * f_i)$$

Proof. Use Theorem A.3 twice as follows.

$$R(a',b') = [R(a,a') * R(a,b')]$$
$$+ \left(\frac{1}{n}\right) * \sum e_i * (b_{i'} - \bar{b})$$

$$= [R(a,a') * R(b,b') * R(a,b)$$
$$+ \left(\frac{R(a,a')}{n}\right) * \sum f_i * (a_i - \bar{a})]$$
$$+ \left(\frac{1}{n}\right) * \sum e_i * (b_{i'} - \bar{b}) .$$

Setting $R(a,a') * (a_i - \bar{a}) = (a_{i'} - \bar{a}) - e_i$
gives the desired result.

Note that the swapping of the a_i is done independently of the swapping of the b_i. Thus, $E(e_i) = 0$, independent of the value of $b_{i'}$; and $E(f_i) = 0$, independent of the value of $a_{i'}$. Therefore, the last 3 terms in Theorem A.4 have an expected value of 0 and we get the following result.

Theorem A.5. If the values of a and b are swapped independently, and if R(a, a') is the correlation coefficient of a with a' and R(b, b') that of b with b', then

$$E\,[R(\underline{a'}, \underline{b'})] = R(\underline{a}, \underline{a'}) * R(\underline{b}, \underline{b'}) * R(\underline{a}, \underline{b}).$$

Note that $R(\underline{a}, \underline{a'}) * R(\underline{b}, \underline{b'}) = [\, R(\underline{a}, \underline{a'}) + R(\underline{b}, \underline{b'}) - 1 + \underline{(1 - R(a, a'))*(1 - R(b, b'))}\,]$. The underlined product is small when $R(\underline{a}, \underline{a'}) + R(\underline{b}, \underline{b'}) > 1.5$, hence a reasonable approximation (when one does not have a calculator handy and when both correlation coefficients are assumed to be approximately 1) for $(R(\underline{a}, \underline{a'}) * R(\underline{b}, \underline{b'}))$ is $[\, R(\underline{a}, \underline{a'}) + R(\underline{b}, \underline{b'}) - 1\,]$.

When $R(\underline{a}, \underline{a'})$ and $R(\underline{b}, \underline{b'})$ are approximately 1, the bias on the correlation introduced by two independent ordinal swaps is approximately $[\, R(\underline{a}, \underline{a'}) + R(\underline{b}, \underline{b'}) - 2\,]$ times the original correlation. Note that the bracketed term is always non-positive.

Appendix B. Construction of a Swapping Interval for a Given Value for the Correlation Coefficient Between the Swapped and Unswapped Values of a Field

Summary. This appendix gives a method for constructing an appropriate set of values from which a value for a_i can be swapped to approximately yield $R(\underline{a}, \underline{a}')$, the correlation coefficient between the set of swapped and unswapped values. It assumes the swapping set for a_i is the subset of all values from $\{a_{i+1}, a_{i+2}, \ldots a_{i+j}\}$, which were not used in a previous swap.

Throughout the section,

a_{topc} = the top-coded value;

a_{botc} = the bottom-coded value; and

N_{total} = the number of observations in the range, $a_{botc} < a_i < a_{topc}$.

The Method. Let j be the length of the swapping interval.
Then

$$j = N_{total} * 2^{1/2} *(1 - R^2(\underline{a}, \underline{a}'))^{1/2} * STD(\underline{a}) /(a_{topc} - a_{botc}).$$

Theorem B.1 Assume $a_{i'}$ is the value swapped with a_i. Let $R(\underline{a}, \underline{a}')$ be the correlation coefficient between $\underline{a}$ and $\underline{a}'$. Let

$(a_{i'} - \bar{a}) = m*(a_i - \bar{a}) + c + e_i$ be the line-of-best-fit for the pairs $(a_i, a_{i'})$.

Then $VAR(e_i) = (1-R^2(\underline{a}, \underline{a}')) * VAR(\underline{a})$.

Proof. By Theorem A.2., $m = R(\underline{a}, \underline{a}')$ and $c = 0$. Assume the error e_i is independent of a_i. Therefore,

$$VAR(\underline{a}') = VAR [(R(\underline{a}, \underline{a}')*a_i + e_i)] = R^2(\underline{a}, \underline{a}')*VAR(\underline{a}) + VAR(e_i) = VAR(\underline{a}).$$

Solving we find

$$VAR(e_i) = (1-R^2(\underline{a}, \underline{a}')) * VAR(\underline{a}).$$

Corollary B.2. Assume $a_{i'}$ is the value swapped for each a_i, since $E(e_i) = 0$, then

$$E(e_i^2) = VAR(e_i) = (1 - R^2(\underline{a})) * VAR(\underline{a}).$$

Theorem B.3. An Estimate of s (due to Jim Fagan). Assume that the swapping length is N(a) and the typical swapping set has s elements. Then $s \simeq 0.75 * N(a)$.

Proof. Consider the element $a_{N(a)+k}$. It will be swapped with exactly 1 of 2*N(a) elements in the set S below.

$S = S_1 \cup S_2 = \{a_k, a_{k+1}, \ldots, a_{N(a)-1}\} \cup \{a_{N(a)}, \ldots, a_{N(a)+k-1}, a_{N(a)+k+1}, \ldots, a_{2*N(a)+k}\}$.

The set S_1 has N(a) - k elements, and the set S_2 has N(a) + k elements.

Assume after each swap of a_i, the elements are replaced. Then S for the swap of a_{i+1} will always have 2*N(a) elements. Therefore, the probability that $a_{N(a)+k}$ will be swapped from an element in S_2 is

$$p(k) = (N(a) + k)/ (2 * N(a)).$$

Now calculate the average P(k) for k = 1, 2, …, N(a). With replacement this average is

$$\bar{p} = 3/4 + 1/(4 * N(a)).$$

Conclusion. With replacement, about 0.75 of the elements in the set $\{a_{k+1}, \ldots, a_{k+N(a)}\}$ would be swapped with values whose indices are greater than or equal to k.

Therefore, $s_{wr} \simeq 0.75 * N(a)$. We then assume $s_{wor} \simeq s_{wr}$.

Table B.4 shows the results for a Monte-Carlo test on the ratio of s to N(a). From this, one may conclude that $s \simeq 0.72 * N(a)$.

Table B.4 Monte Carlo Testing for the Expected Swapping Set Size, s

N(a)	s	s/N(a)
100	73	0.730
250	180	0.720
500	361	0.722
1000	721	0.721
1500	1082	0.721
2000	1444	0.722
2500	1806	0.722

Theorem B.4 (due to Moore and Fagan). Assume the set $\underline{a} = \{a_i\}$ is uniformly distributed between the range, a_{botc} to a_{topc}. Suppose the desired coefficient between the swapped and unswapped values is $R(\underline{a}, \underline{a}')$. Then a reasonable estimation for a p-percent rank swapping interval is

$$(P(\underline{a}) / 100) = (18/7) * (1-R^2(\underline{a}, \underline{a}'))^{1/2} * STD(\underline{a})/(a_{topc} - a_{botc}).$$

Proof. (1) Let $N(\underline{a})$ = the maximum number of observations in the swapping interval of fixed length, then

$$P(\underline{a})/100 = N(\underline{a})/N_{total}.$$

(2) Uniform distribution implies

$a_{i+k} = a_i + d*k$, where $d = a_{i+1} - a_i$.

(3) The swapping set for a_i is the set of all unswapped values in $\{ a_{i+1}, a_{i+2}, \ldots, a_{i+N(a)}\}$.

(4) Now a_{i+1} may have already been swapped with one of the $N(\underline{a})$-1 members of the set $\{a_{i-N(a)+1}, \ldots, a_{i-2}, a_{i-1}\}$.

Now a_{i+2} may have already been swapped with one of the $N(\underline{a})$-2 members of the set $\{a_{i-N(a)+2}, \ldots, a_{i-2}, a_{i-1}\}$.

Now a_{i+k} may have already been swapped with one of the $N(\underline{a})$-k members of the set $\{x_{i-N(a)+k}, \ldots, a_{i-2}, a_{i-1}\}$.

Let p(k) be the probability that a_{i+k} is still in the swap set. From the proof of Theorem B.3, we see, that

$$p(k) = (N(\underline{a}) + k)/ (2 * N(\underline{a})).$$

Let q(k) = the probability that the value of a_{i+k} is swapped for the value of a_i. Then q(k) can be estimated by the probability a_{i+k} is still in the swap set times the probability that it is selected (i.e., p(k) / E(s)).

From Theorem 3, $E(s) \simeq (3/4) * N(a)$, so

$$q(k) = \frac{2}{3} * \frac{(N(a)+k)}{(N(a))^2}.$$

(5) Recall that the a_i's are uniformly distributed, so that $(a_{i+k} - a_i) = (k * d)$. Therefore,

$$\begin{aligned} VAR(e_i) &= E[(a_{i,j} - a_i)^2] = \sum_{k=1}^{N(a)} q_k * (a_{i,k} - a_i)^2 \\ &= \sum \frac{2}{3} * \frac{(N(a)+k)}{N(a)^2} * (k^2 * d^2) \\ &= \frac{(2*d^2)}{3*N(a)^2} * \sum [N(a)*k^2 + k^3] \\ &= d^2 * [\frac{2}{9} * (N(a)+1) * (2N(a)-1) + \frac{1}{6} * (N(a)+1)^2]. \end{aligned}$$

This shows that under "ideal" conditions (with N(a) large, so $(N(\underline{a}) + 1) \simeq N(\underline{a})$ and $(2*N(a) - 1) \simeq 2*N(a)$), then

$$VAR(e_i) \simeq (7/18) * (N(\underline{a}) * d)^2 .$$

(6) From Theorem B.1, $VAR(e_i) = (1-R^2(\underline{a}, \underline{a}')) * VAR(\underline{a})$.

Equating the variance terms in Steps (5) and (6), we get our estimate for N(a) as

$$N(\underline{a}) = \{ (18/7)* (1 - R^2(\underline{a}, \underline{a}')) * VAR(\underline{a}) \}^{1/2} * (1/d).$$

(7) But the N_{total} a_i's are assumed to be uniformly distributed in the interval (a_{botc}, a_{topc}). Therefore,

$$d = (a_{topc} - a_{botc})/N_{total}.$$

Substituting this expression for d in Step (6) and dividing both sides by N_{total}, we get the desired result, namely,

$$(P(\underline{a}) / 100) = \{ (18/7) * (1-R^2(\underline{a}, \underline{a}')\}^{1/2} * STD(\underline{a})/(a_{topc} - a_{botc}).$$

Conjecture B.5 (For a Slightly Skewed Distribution). Theorem B.3 was proven for uniform distributions on the interval (a_{botc}, a_{topc}). For non-uniform distributions, the value of P(a) should be smaller. Empirical testing seems to indicate that rounding 18/7 down to 2 gives reasonable results. Namely,

$$(P(\underline{a}) / 100) = 2^{1/2} * (1-R^2(\underline{a}, \underline{a}'))^{1/2} * STD(\underline{a})/(a_{topc} - a_{botc}).$$

Appendix C. Determination of an Appropriate Fixed Interval Length To Give a "K" Percent Average Absolute Difference

Summary. In Appendix B, we have shown that in order to obtain a given $R(\underline{a}, \underline{a}')$, the correlation coefficient between the swapped and unswapped values for $\underline{a}$, a reasonable estimate for the interval length is

$$N(\underline{a}, R(\underline{a}, \underline{a}')) = N * 2^{1/2} * (1 - R^2(\underline{a}, \underline{a}'))^{1/2} * STD(\underline{a}) / (a_{topc} - a_{botc}).$$

Here

N is the number of non-missing, non-top-coded, or non-bottom-coded observations for the field $\underline{a}$;

$STD(\underline{a})$ is the standard deviation for the set of these N observations; and

a_{topc} and a_{botc} are the top- and bottom-codes, respectively.

In this appendix, we will show that if, for some fixed value of K_0, you desire the condition

$$E(|a_i - a_{i'}| / a_i) = K_0,$$

then a reasonable approximation for the interval length is

$$N(\underline{a}, K_0) = N * (8/3)^{1/2} * K * \bar{a} / (a_{topc} - a_{botc}),$$

where $\bar{a}$ is the mean of the N observations.

Appendix B Results. Let $d_i = (a_i - a_{i'})$, where the value a_i is switched with that of a_i'. Theorem B.1 shows that

$$E(d_i^2) = (1 - R^2) * STD^2(a).$$

Assumptions on the Distribution of d_i. Assume d_i is uniformly distributed on the interval $[-z_i, z_i]$ for each i.

This implies

(1) $E(d_i^2) = z_i^2 / 3$, and

(2) $E (|d_i|) = z_i / 2$; therefore

(3) $E(d_i^2) = (4/3) * E^2(|d_i|)$.

Note that these results hold even though z_i may vary with i.

An Appropriate Expression for E($|d_i|$). Let $d_{ij} = a_i - a_j$, for every possible value of a_j that can be switched with a_i. Then define

$$k_{ij} = |d_{ij}| / a_i.$$

Now take expected values over j with i held constant, to get

$$\begin{aligned} k_i &= E(k_{ij}) \\ &= E(|d_{ij}|) / a_i \\ &= \bar{|d_i|} / a_i. \end{aligned}$$

Here the bar was used to indicate the expected value was taken over a second index "j". We can just as easily write this as
E ($|d_i|$), so

$$E(|d_i|) = k_i * a_i.$$

Ideally, k_i will be independent of a_i, so

$$E(E(|d_i|) = \bar{k} * \bar{a}.$$

Again, in an ideal world, E($|d_i|$) is constant with respect to i, so that

$$E(E(|d_i|) = E(|d_i|) = \bar{K} * a, \text{ with } \bar{K} = k.$$

Conclusion. Putting everything together one realizes,

$$\begin{aligned} N(\underline{a}, R(\underline{a}, \underline{a}')) &= N * 2^{1/2} * (1 - R^2(\underline{a}, \underline{a}'))^{1/2} * STD(\underline{a}) / (a_{topc} - a_{botc}) \\ &= N * 2^{1/2} * E^{1/2}(d_i^2) / (a_{topc} - a_{botc}) \\ &= N * (8/3)^{1/2} * E(|d_i|) / (a_{topc} - a_{botc}) \\ &= N * (8/3)^{1/2} * K * \bar{a} / (a_{topc} - a_{botc}). \end{aligned}$$

Results. When implemented on 1993 Annual Housing Survey data, the above estimate proved amazingly accurate in spite of all the ideal assumptions (uniform distributions, independent variables, and constants instead of expectations). Table 4 shows the results.