MASKING MICRODATA FILES

Jay J. Kim* and William E. Winkler*, Bureau of the Census
Jay J. Kim, Room 3134-4, Washington, D.C. 20233-9100

This paper describes an application of a general methodology for producing public-use data files that preserves confidentiality and allows many analytical uses. The methodology masks quantitative data using an additive-noise approach due to the first author and then, when necessary, employs a reidentification/swapping methodology to assure confidentiality.

KEY WORDS: Confidentiality, Noise Inoculation, Reidentification, Swapping

1. INTRODUCTION

While many types of data are collected by government agencies, use of the microdata files is often limited to sworn agents working on secure computer systems at those agencies. The confidentiality restrictions can severely affect public policy decisions made at one agency that has access to nonconfidential summary statistics but not to the microdata that are collected at two or more other agencies. The application of this paper is in producing a public-use data base that contains much data from the March Supplement to the Current Population Survey (CPS) and income data from the Internal Revenue Service (IRS) 1040 Form. The data are for use by the Department of Health and Human Services (HHS) in setting policy regarding earned income credit and other benefits. The microdata is masked in such a manner that both Bureau of the Census and IRS confidentiality restrictions are met. No masked IRS quantitative data can alone be used in reidentifications.

The main methodology is an additive-noise approach (Kim 1986) for masking multivariate normal data that preserves confidentiality and can preserve many essential characteristics of the data such as means, variances, and correlations. The CPS and IRS data of the application are known to be approximately multivariate normal. While the methodology has been extended to general data distributions (Sullivan and Fuller 1989, 1990; also Little 1993), the extension involves transforming general data to multivariate normal, masking, and then transforming the masked data back to the original scale. As we begin with multivariate normal data, we need not be concerned with the two additional transformation steps of the more general Sullivan-Fuller methods. We do note that the set of general software that we developed for arbitrary multivariate normal data could be extended to the general data by inclusion of software performing the two Sullivan-Fuller transforms.

The secondary methodology of this paper is a sophisticated reidentification/swapping technology that is based on existing record linkage concepts (Winkler 1994, 1995a). The matching software uses the masked CPS and IRS quantitative data along with other variables such as age, race, sex, and State to produce reidentifications with the original merged file of unmasked CPS and IRS data. Since we know true matching status, we can minimize the number of pairs of records in which quantitative data is swapped. While swapping can help preserve confidentiality, it can reduce the analytic usefulness of the file (Little 1993). By minimizing swapping, we assure the analytical usefulness of the final file as we show later.

The outline of this paper is as follows. In the second section of this paper we describe the data files, the additive-noise masking methodology, and the reidentification/swapping methodology. The third section provides results. In the fourth section, we discuss some of the limitations of the masking methodology, provide an overview of the general software we developed, and describe an additional methodology called controlled distortion. The final section consists of summary and conclusions.

2. DATA AND METHODS

This section describes the data, the masking methodology, and the reidentification/swapping methodology.

2.1. Data to be Masked

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

i) Total income;
ii) Adjusted gross income;
iii) Wage and salary income;
iv) Taxable interest income;
v) Dividend income;
vi) Rental income;
vii) Nontaxable interest income;
viii) Social security income;
ix) Return type;
x) Number of child exemptions;

xi) Number of total exemptions;
xii) Aged exemption flag;
xiii) Schedule D flag;
xiv) Schedule E flag;
xv) Schedule C flag; and
xvi) Schedule F flag.

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

2.2. Masking Methodology

Masking is via an additive noise approach (Kim 1986, see also Sullivan and Fuller 1989, Sullivan and Fuller 1990, and Little 1993). Adding random noise with the same correlation structure as the original unmasked data is currently the only method (Little 1993) that preserves correlations. Theoretical details are in the appendix of a longer research report that is available from the authors. Masking is done according to the following steps:

i) Calculate the variance/covariance for income types iii) through viii) in section II. This results in a 6×6 variance/covariance matrix.
ii) Take c×100 percent of the above variance-/covariance and generate random numbers using subroutine *RNMVN* in International Mathematical and Statistical Library (IMSL). Note that RNMVN requires the users to provide the variance/covariance which the generated random numbers should have. This process produces 59,315×6 matrix of random numbers. The expected value of the generated random numbers for each of the 6 arrays is 0.
iii) Add the random numbers generated in ii) to the income fields in section 2.1. Note that both the raw income data in section 2.1 [income types iii) through viii)] and the noise in step ii) of this section are of matrix 59,315×6. Thus the addition is done by element by element of the matrices.
iv) Sum up incomes for each individual for income types iii) through viii) in section 2.1 and calculate the difference between the sum and the total income, and the difference between the sum and the adjusted gross income.
v) Sum up noise inoculated incomes of types iii) through viii) for each individual. Add to the sum of the perturbed incomes the difference between the sum of raw incomes and the total income calculated in step iv) above. This gives the masked total income. Masked adjusted gross income is produced similarly.

Six income variables are masked directly and the remaining two are masked in a manner that preserves sums. If top-coding is required for the incomes at 100,000 (or -100,000), it can be done after the above five steps. In some situations, data providers censor outliers prior to masking because outliers (even when masked) are particularly easy to reidentify. In our approach, we specifically assume that data are not censored because censoring reduces analytic usefulness of the masked file. It is straightforward make modifications to deal with censored data.

As the users might want to tabulate the counts of individuals depending on the recipiency status of various IRS income and the noise inoculation completely changed the zeros and non-zeros both alike, we are going to add flags indicating whether each amount of unmasked income was zero or not. This will allow them to analyze the data for recipient group and nonrecipient group, separately.

Even after masking, it may be possible to reidentify a certain proportion of records in the masked file with the original, corresponding records in the unmasked file. While the 1/1600 sample assures that most mid-to-low income individuals can not be reidentified in the entire population using information from the public-use file, some individuals with high incomes or unusual combinations of age, sex, race and income characteristics might be reidentified. Specifically, if we can reidentify a mid-income record across masked and unmasked sample files and there are 2000 individuals in the population with essentially the same characteristics as those that were used in the reidentification, then there is only a 1 in 2000 chance of a reidentification. In other words, it is not possible to reidentify such a mid-income individual in the entire population via information in the public-use file. However, it may still be possible to reidentify individuals with high incomes or with unusual characteristics. To minimize the chance of reidentification, we need to employ additional procedures in a manner that does not eliminate the analytical usefulness of the public-use file. Such minimization may be possible because we are the data

providers and have knowledge of the exact truth of reidentifications between unmasked and masked sample files.

2.3. Reidentification/Swapping Methodology

To determine how much reidentification is possible, we proceed in two stages. First, we match the merged raw data file against the masked file using record linkage software (Winkler 1994, 1995a). Based on the reidentification rate, we next swap quantitative data according to a proportion that minimizes the chance of reidentification.

During the first stage, we use blocking variables such as age, race, sex, and State code and matching variables such as the IRS income and CPS quantitative variables. Blocking is a record linkage term that means that we only consider pairs that agree exactly on the blocking variables. The quantitative matching variables need not agree exactly. The matching decision rule is based on an information-theoretic extension of the likelihood ratio test (Fellegi and Sunter 1969) that assigns scores to each pair based on a function of their associated likelihood ratios. Likely reidentifications, called matches, are given higher scores, and other pairs, called nonmatches, are given lower scores. To best separate the pairs into matches and nonmatches, we use a version of the EM algorithm for latent classes (Winkler 1994, 1995a) that determines the best set of matching parameters under certain model assumptions which are not seriously violated in this particular situation.

During the second stage, we first collapse cells (age × race × sex) to assure that there are sufficient candidates for swapping. The collapsing strategy is similar to those used in sampling and nonresponse imputation. Within collapsed cells we randomly swap quantitative data according to a proportion that we specify. Since we know true matching status, we can minimize the swapping proportion because we know exactly which pairs are reidentifications. We note that the specific set of reidentifications varies with each different seed value used at the masking stage. Swapping preserves means and correlations in the subdomains on which it was done and in unions of those subdomains. On arbitrary subdomains, however, collapsing and the amount of swapping can adversely affect the analytic validity of the files. If swapping is done such that each record that is swapped is only swapped with another record in that subdomain, then we say that we have *controlled* that subdomain. Means and correlations among swapped variables within controlled subdomains are necessarily the same. We cannot hope for confidentiality while providing analytic validity in arbitrary subdomains above a certain size. If we were to provide such analytic validity in subdomains above a certain size, then we would necessarily be able to reidentify every record in the file.

3. RESULTS

3.1 Utilities of the Full Sample Data

Since the model building requires mean and variance/covariance or correlation of the variables involved, statistics were calculated for six variables in the raw and masked data. The means of the raw and masked data are almost identical (Table 1).

Table 1. Means of Raw and Masked Data

Type	Raw	Masked
Wage	23,799	23,784
Tax Int	1,825	1,823
Div	587	587
Rent	1,190	1,187
Ntax Int	342	342
Soc Sec	947	948

Table 2. Correlation for Raw and Masked Data

	Raw	Masked
Wage vs Dividend	.18	.18
Wage vs Tax Int	.12	.12
Dividend vs SS	.12	.12
Tax Int vs Rent	.08	.08
Dividend vs Rent	.04	.04
Ntax Int vs SS	.04	.04

Table 2 shows that all correlations are the same to two decimal places.

As mentioned before, total and adjusted gross income were masked indirectly by summing up masked components of the total and adjusted gross income except the difference between the sum of the unmasked data and total or adjusted gross income. The means of the total and adjusted gross income from the masked data are virtually identical as those from the unmasked data. They are less than .07 percent off from each other. This can be expected since the noise was added to all components which has zero expected value.

Similarly, the variance of the total and adjusted gross income from the masked data are virtually identical to those from the unmasked data.

3.2 Subdomain Estimation - before Swapping or When Swapping Was Controlled for Subdomain

Data users are very often interested in subgroups in their analysis, thus the subdomain estimation very

often matters. The subdomain estimation formula for the masked data is given in Appendix 3 of the longer research report. In short, subdomain mean is not affected by the masking and in the current case only a minor adjustment is needed to the variance-/covariance according to the formula shown in the appendix because the amount of noise added is low (in terms of the variance/covariance). The adjustment also has almost no effect on the correlation. For those persons whose "return type" is 4, (unmarried head of household return), the means and correlations of the unmasked data for the six incomes are computed.

Generally, the estimates of means from the masked data are excellent. That is, for five items, they are virtually identical with those from the unmasked data. However, the estimate of mean nontaxable interest (61) from the masked data is more than 10 percent off from the mean (70) of the unmasked data. Tables 3 shows correlations between the income variables for the unmasked and masked data, respectively.

Table 3. Correlation for Raw and Masked Data for Return Type = 4

	Raw	Masked
Wage vs Dividend	.027	.029
Wage vs Tax Int	.108	.105
Dividend vs SS	.155	.154
Tax Int vs Rent	.172	.171
Dividend vs Rent	.040	.039
Ntax Int vs SS	.056	.052

The table shows that estimation of correlations for this subdomain based on the masked data is generally good. They are the same as those from the unmasked data down to the second decimal place. The statistics were estimated from the masked data for other subdomains such as return type=1 (single return) and Schedule C=1 (Schedule C was filed in the tax return) and similar findings as before were found.

Thus far we have observed the behavior of subdomain estimates when the subdomain is formed by a variable which is not masked. What happens when the subgroup is formed by a masked variable itself? By adding noise, in effect we expand the range of values the variable can take. Thus if we use the same cutoff to form a subgroup for both the unmasked and masked data, there is no guarantee that the same elements will be in the same group in both data sets. To check on the performance of statistics when the subdomain is formed based on the masked variable, wage and salary, shortened to wage, is chosen to be used as a classification variable. A cutoff of 15,000 was used and summary statistics were calculated for the group having no more than that amount for both the masked and unmasked.

When the subgroup was formed based on the unmasked wage, 28,268 persons were in the group, but 28 more people were found in the group when it was formed based on the masked wage. However, the means from the both data sets are virtually identical. Difference in the correlations between the unmasked and masked data is found at the third decimal place. In fact, all the correlations are virtually identical.

3.3. Subdomain Estimation - When Swapping Was Not Controlled for Subdomain

When the records were swapped, not the full records, but substrings were swapped between records composed of eight IRS income fields mentioned above and three CPS income fields such as wage (it will be called CPS Wage), adjusted gross income (it will be called CPS Agi) and aggregated sum of rent (net rent), dividend and interest (it will be called CPS Prop). Thus when swapping is done not controlling for a certain subdomain, the statistics for the subdomain can be changed to a certain degree. To see the effect of swapping on statistics in the subdomain for which swapping was not controlled, statistics were calculated for a subdomain composed of schedule C users. It was repeated twice for 5-percent and 20-percent swapping.

Table 4. Means before And after Swapping for Schedule C Users, n = 7,819

	Raw	Masked	5% Swap	20% Swap
Wage	24,715	24,677	25,338	26,891
Rent	2,820	2,822	2,779	2,746
Tax Int	2,178	2,174	2,171	2,145
Dividend	783	779	773	755
Ntax Int	393	391	366	346
Soc Sec	790	790	803	822

The table shows that i) we can get an excellent subdomain estimation from the masked data for this subdomain even if masking was not controlled for this subdomain; ii) 5-percent swapping does not affect means that much; and iii) as the rate of swapping is increased to 20-percent, the means become more different from the means of the masked data.

The next table shows some selected correlations.

Table 5. Correlations before and after Swapping for Schedule C Users, n = 7,819

Fields	Raw	Masked	Swap Rate 5%	Swap Rate 20%
Wage, Dividend	.6361	.6352	.6143	.6217
Wage, Tax Int	.1903	.1900	.2425	.2413
Dividend, SS	.1535	.1547	.1528	.1346
Tax Int, Rent	.1984	.1978	.1967	.2167
Dividend, Rent	.1291	.1285	.1265	.1304
Ntax Int, SS	.1057	.1062	.1181	.0957

Swapping has some impact on the correlations, but did not harm them too much. 5-percent swapping produced better means but does not necessarily do so for correlation.

4. DISCUSSION

The discussion covers how representative the masking procedures are and some of their limitations. The section also provides an overview of our general computer software for masking arbitrary multivariate normal files and a new methodology called controlled distortion.

4.1. Representativeness of Results

The masking/swapping procedures were repeated with two additional seed numbers for the random noise-generation routine. The correspondences of means and correlations between unmasked and masked/swapped files were consistent with those given in this paper. We note the actual set of reidentification/swaps varies with the seed numbers because reidentifications depend on how close individual masked data records are to corresponding unmasked data records. The closeness is dependent on the random noise which varies with the seeds.

4.2. Limitations

As Tables 4 and 5 show for subdomains in which swapping is not controlled, means and correlations in the masked/swapped file may not be consistent with those in the original unmasked file. If we provide two or more copies of masked/swapped files corresponding to different seed numbers, then users can check whether a subdomain analysis is plausible. If the users cannot approximately reproduce an analysis (say a hypothesis test) on one copy that is also performed on another copy, then the users can assume that the masked/swapped file does not support that type of an analysis.

When a masked/swapped continuous variable is used for categorization, the number of observations in categories may not be close to those from the unmasked data. This is because the categorization implicitly corresponds to subdomains in which swapping may not be controlled. The summary statistics for categories between unmasked and masked/swapped data can be consistent if the sizes of the categories are large. If the subdomain of interest is of small size, then we should be careful about using statistics for the subdomain.

4.3. Software

The current version of the computer software can be used for masking and swapping general multivariate normal files. The first program (in SAS) produces an output file consisting of the variance/covariance matrix for the raw data. The second program (in FORTRAN) calls the IMSL routine *RNMVN* to produce random multivariate noise with the same variance/covariance as the raw data. The third combines raw data and noise to produce the masked file. The fourth program (in C) does swapping. All software is portable provided the IMSL routine *RNMVN* is available.

4.4. Controlled Distortion

To provide a means (either additionally or alternately) of preserving analytic validity while further reducing identifiability, we have developed a procedure called controlled distortion. *Controlled distortion* allows a user to distort arbitrarily a single record and necessitates complementary distortions in a small set of additional records so that means and covariances are preserved. The intuitive idea is that we may not wish to swap some records (say a few of those having high incomes) that are easily identifiable because we may adversely affect analyses in some subdomains. Controlled distortion, in many situations, can yield better consistency of means and covariances across arbitrary subdomains. Theoretical details are given in the longer research report. We presently have not written computer software.

5. SUMMARY AND CONCLUSIONS

We demonstrated a methodology for producing a confidential, public-use file that contains eight income fields from the 1990 IRS Tax Return file and the remaining data from the 1991 CPS public-use file. The file was produced in two stages. The first stage consisted of adding random noise with the same correlation structure as the original, unmasked data. The second stage involved reidentifying and swapping records via a record linkage approach.

We investigated the masked file and the masked/swapped file. The masked file provides means and correlations (even in many subdomains) that are very close (3 decimal places) to means and variances in unmasked files. The risk of disclosure for the masked file is somewhat high. As much as 0.8% of the records have a probability of disclosure above 20%; the remaining 99% have a disclosure risk

of less than 0.02%. For the entire domain, means and correlations from the masked/swapped file were typically within 3 decimal places from the corresponding means and correlations in the unmasked file. Deviations in many subdomains were higher; sometimes deviating in the second decimal place. The disclosure risk for all records in the masked/swapped file is below 0.1%.

Swapping can distort the correlations, particularly on subdomains. We suggest releasing two copies (one for each seed used in the random number generator) of the masked/swapped files. If users cannot reproduce a statistical analysis using data from one copy that was done on the other copy, then they can be assured that the public-use file will not support the attempted analysis. In that case, there are two recourses. The first is for the data providers to supply two more copies of the public-use file that have been masked and swapped in a manner that supports the originally attempted analysis. If that is not possible, then the only second recourse is to have the statistical analysis performed on the original, unmasked data.

* The views expressed in this paper are those of the authors and not necessarily those of the U.S. Bureau of the Census. A longer research report with technical appendixes is available from the authors.

REFERENCES

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

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

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

Kim, J. J., and Winkler, W. E. (1995), "General Masking Software," computer software and unpublished documentation, Statistical Research Division, U.S. Bureau of the Census.

Little, R. J. A., (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, **9**, 407-426.

Sullivan, G., and Fuller, W. A. (1989), "The Use of Measurement Error to Avoid Disclosure," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 802-807.

Sullivan, G., and Fuller, W. A. (1990), "Construction of Masking Error for Categorical Variables," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 435-439.

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

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

Winkler, W. E. (1995b), "Reidentification and Swapping Software," computer system and unpublished documentation, Statistical Research Division, U.S. Bureau of the Census.