CHAPTER V – Methods for Public-Use Microdata Files

One method of publishing the information collected in a census or survey is to release a **public-use** microdata file (see Section 2.D). A microdata file consists of records at the respondent level where each record on the file represents one respondent. Each record consists of values of characteristic variables for that respondent. Typical variables for a demographic microdata file are age, race, and sex of the responding person. Typical variables for an establishment microdata file are Standard Industrial Classification (SIC) code, employment size, and value of shipments of the responding business or industry. Most public-use microdata files contain only demographic microdata. The disclosure risk for most kinds of establishment microdata is much higher than for demographic microdata. The reasons for this are explained in Section C.4 of this chapter.

This chapter concerns microdata files that are publicly available, that are **public-use** microdata files. In addition to or instead of public-use files, some agencies offer **restricted-use** microdata files. Access to these files is restricted to certain users at certain locations and is governed by a restricted use agreement.

To protect the confidentiality of microdata, agencies remove all obvious identifiers of respondents, such as name and address, from microdata files. However, there is still a concern that the release of microdata files could lead to a disclosure. Some people and some businesses and industries in the country have characteristics or combinations of characteristics that would make them stand out from other respondents on a microdata file. Public use microdata files contain some measure of risk of disclosing confidential information. A statistical agency releasing a microdata file containing confidential data must do its best to minimize the risk that an outside data user can correctly link a respondent to a record on the file. Aside from not releasing any microdata, there is no way of removing all disclosure risk from a file; however, agencies must make reasonable efforts to minimize this risk and still release as much useful information as possible.

Several Federal agencies including the Census Bureau, National Center for Education Statistics, National Center for Health Statistics, Centers for Medicare and Medicaid Services, Energy Information Administration, Social Security Administration, Bureau of Transportation Statistics, and Internal Revenue Service release microdata files. This chapter describes the disclosure risk associated with microdata files, mathematical frameworks for addressing the problem, and necessary and stringent methods of limiting disclosure risk.

A. Disclosure Risk of Microdata

Statistical agencies are concerned with a specific type of disclosure of personal information that relates to a respondent, and there are several factors that play a role in the disclosure risk of a microdata file. A record is at risk of being identified if a respondent is unique in the database with respect to a set of identifying variables and if the intruder knows that the respondent is on

the file. Data providers that are subject to the privacy rule under the Health Insurance Portability and Protection Act (HIPAA) and/or the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) must take affirmative steps to protect the confidentiality of the reported values before the database is released as a public-use file.

A.1. Disclosure Risk and Intruders

Most national statistical agencies collect data under a pledge of confidentiality. Any violation of this pledge is a disclosure. An outside user who attempts to link a respondent to a microdata record is called an **intruder**. The disclosure risk of a microdata file greatly depends on the motive of the intruder. If the intruder is hunting for the records of specific individuals or firms, chances are that those individuals or firms are not even represented on the file that possesses information about a small sample of the population. In this case, the disclosure risk of the file is very small. The risk is much greater, on the other hand, if the intruder is attempting to match *any* respondent with their record to an external file. We can measure disclosure risk only against a specific compromising technique that we assume the intruder to be using (Keller-McNulty, McNulty, and Unger, 1989).

A.2. Factors Contributing to Risk

There are two main sources of the disclosure risk of a microdata file. One source of risk is the existence of high-risk records. Some records on the file may represent respondents with unique characteristics such as very unusual jobs (e.g. movie star, Federal judge) or very large incomes (e.g. over one million dollars). An agency must decrease the visibility of such records. Another type of high-risk records includes those cases where multiple records in a data file are known to belong to the same cluster (for example, household or school). In this case, there is a greater risk that either one may be identified (even if no information about the cluster per se is provided). A third type of high-risk records can occur when one dimension of the data are released in too fine a level of detail. In this case, if data are released for small areas, such as school districts, variables that would not create disclosure problem at a higher level of aggregation, such as a state or region, may result in an increased risk of disclosure. An example might be, teacher's income by race/ethnicity and age.

The second source of disclosure risk is the possibility of matching the microdata file with external files. There may be individuals or firms in the population that possess a unique combination of the characteristic variables on the microdata file. If some of those individuals or firms happen to be chosen in the sample of the population represented on that file, there is a disclosure risk. Intruders potentially could use external files that possess the same characteristic variables and identifiers to link these unique respondents to their records on the microdata file.

Knowledge of which individuals participated in a survey, or even which areas were in the sample, can greatly help an intruder to identify individuals on a microdata file from that survey. Advising respondents to use discretion when telling others about their past participation in surveys is appropriate but may make respondents wary of participating in the survey. The

disclosure risk of a microdata file is greatly increased if it contains administrative data or any other type of data from an outside source linked to survey data. Those providing the administrative data could use those data to link respondents to their records on the file. This is not to imply that providers of administrative data would attempt to link files, however, it is a possibility and precautions should be taken. In addition, in some cases, the administrative data may be already released as a public use file, so any intruder could use the information to try to identify an individual. The potential for linking files (and thus the disclosure risk) increases as the number of variables common to both files increases, as the accuracy or resolution of the data increases, and as the number and availability of external files increases, not all of which may be known to the agency releasing the microdata file.

Longitudinal and panel surveys create a special case of disclosure risk that may be associated with linked files. In this case, the disclosure risk of a microdata file increases if some records on the file are released on another file with more detailed or overlapping recodes (categorizations) of the same variables. Likewise, risk increases if some records on the file are released on another file containing some of the same variables and some additional variables.

As a corollary, there is greater risk when the statistical agency explicitly links a new microdata file on a set of respondents with published data for those same respondents at an earlier point in time. This occurs in longitudinal surveys, such as the Census Bureau's Survey of Income and Program Participation, where the same respondents are surveyed several times and the NCES high school longitudinal surveys where students are followed for 10 to 12 years through high school, postsecondary education, and into the labor force and/or parenthood. The amount of risk is increased when the data from the different time periods can be linked for each respondent. Changes that an intruder may or may not see in a respondent's record (such as a change in occupation or marital status or a large change in income) over time could lead to the disclosure of the respondent's identity.

In general, the disclosure risk of a file increases as the structure of the data becomes more complex - whether it is through the addition of linked data from an external source, or through the addition of linked data for a set of respondents across time, the effect is the same. More complex variable structure also leads to an increase in the likelihood of unique streams of data responses, and thus an increase in the likelihood of disclosure.

A.3. Factors that Naturally Decrease Risk

Sampling is an important factor in decreasing risk of disclosure in microdata files. As we stated previously, if an intruder possesses such a microdata file and is looking for the record of a specific individual or firm, chances are that that individual or firm is not even represented on the file. Also, records on such a file that are unique compared with all other records on the file may not represent respondents with unique characteristics in the population. There may be several other individuals or firms in the population with those same characteristics that did not get chosen in the sample. This creates a problem for an intruder attempting to link files.

The disclosure risk of the file can be decreased even further if only a subsample of the sampled population is represented on the file. Then, even if an intruder knew that an individual or firm participated in the survey, he or she still would not know if that respondent appeared on the file. Data users, however, generally want the whole sample.

Another naturally occurring factor that decreases the risk of disclosure is the age of the data on microdata files and any potentially matchable external files. When an agency publishes a microdata file, the data on the file are usually at least one to two years old. The characteristics of individuals and firms can change considerably in this length of time. Also, the age of data on potentially matchable files is probably different from the age of the data on the microdata file. One caveat is that the difference in age of the data between files may not complicate the job of linking older files if an intruder has access to an external file that corresponds in time to the data collection.

The naturally occurring noise in the microdata file and in potentially matchable files decreases the ability to link files. All such data files will reflect reporting variability, non-response, and various edit and imputation techniques.

Many potentially matchable files have few variables in common. Even if two files possess the "same" characteristic variables, often the variables are defined slightly differently depending on the purpose for collecting the data. Sometimes the variables on different files are recoded differently. The definitions of any variables that are common to both files should be checked to verify that the definitions are the same, otherwise, the variables may actually be measuring different activity. Differences in variable definitions and recodes can make an intruder's job more difficult.

The final factors that decrease risk are the time, effort, and money needed to link files, although, as computer technology advances, these factors are diminished.

A.4 Disclosure Risks Associated with Regression Models

The question of whether disclosure risks exist in regression-type models has become more important over the past decade as federal agencies expand access to their micro data. The risks associated with public use files have increased due to increased computing power coupled with the development of sophisticated data matching software and the increasing availability of electronic databases on the Internet. At the same time, demand for access to microdata files has increased as the researcher community has recognized the value of the files and increased computing power has made analyzing the files much easier. In response to these developments, agencies have developed several modes of restricted access to data: the U.S. Census Bureau has taken the lead on establishing Research Data Centers (RDCs); NCES has made use of licensing agreements; and NCHS has developed remote access systems for users to access micro data files.

The U.S. National Science Foundation and NCES have jointly funded work by the U.S. National Institutes of Statistical Sciences (NISS) to study issues in developing "model servers," which will

allow researchers to estimate models from databases of confidential microdata without having direct access to the microdata. The NISS researchers have investigated how to release useful results (e.g., regression parameter estimates and model diagnostics) while not compromising confidential information (Gomatam et al, 2005). They have also investigated how to estimate regressions using a combination of confidential data from several sources; e.g., several statistical agencies (Karr et al, 2005).

Disclosure risks may arise from the use of regression models, particularly in the standard linear regression model estimated using Ordinary Least Squares methods as well as in logit and probit models (which use binary (0,1) dependent variables) and other Generalized Linear Models (Reznek 2003, Reznek and Riggs, 2004). The risks in regression models that contain continuous variables on the right-hand side are small if the overall sample is large enough to pass tabular disclosure analysis. However, risks may exist in models that contain dummy variables as independent variables. Coefficients of models that contain only fully-interacted (saturated) sets of dummy variables on the right-hand sides can be used to obtain entries in cross-tabulations of the dependent variable, where the cross-tabulation categories are defined by the dummy variables. The same types of cross-tabulations can also arise from correlation and covariance matrices of the variables, and from variance-covariance matrices of model coefficients, if these matrices include dummy variables. These research outputs present disclosure risks if the cross-tabulations present disclosure risks.

B. Mathematical Methods of Addressing the Problem

Although several mathematical measures of risk have been proposed, none has been widely accepted. Techniques that reduce the disclosure risk of microdata include methods that either reduce the amount of information provided to data users or methods that slightly distort the information provided to data users. Several mathematical measures of the usefulness of disclosure-limited data sets have been proposed to evaluate the trade off between protection and usefulness. Again, none has been widely accepted. More research is necessary to identify the best disclosure limitation methodology sufficient for both data users and suppliers of confidential microdata.

Before describing these mathematical methods of addressing the problem of disclosure risk, we must mention several mathematical and computer science problems that in some way relate to this problem. For example, various mathematical methods of matching a microdata file to an outside file can be found in literature concerning record linkage methodology at http://www.fcsm.gov/working-papers/RLT_1997.html. Record Linkage Techniques, 1997 -- Proceedings of An International Record Linkage Workshop and Exposition presents reprints of the major background papers in record linkage as well as discussions of current work.

B.1. Proposed Measures of Risk

Measuring the disclosure risk of a public use microdata file involves measuring the probability that an intruder is able to identify a record. Most research has considered some or all of the following factors:

- the probability that the respondent for whom an intruder is looking is represented on both the microdata file and some matchable file,
- the probability that the matching variables are recorded identically on the microdata file and on the matchable file,
- the probability that the respondent for whom the intruder is looking is unique in the population for the matchable variables, and
- the degree of confidence of the intruder that he or she has correctly identified a unique respondent.

A model for measuring disclosure risk should reflect certain a prior assumptions about the intruder. The level of risk varies depending upon whether the intruder wishes to disclose the reported values of a particular respondent, or the reported values of any respondent, or a group of respondents. (See Steel, 2004). The validity of the measures of risk depend upon the accuracy of the file preparer's designation of the key variable list. This is a set of variables on the microdata file that may be used to identify unique records in the file and that also exist on data that is in the public domain (or could be held privately from some outside commercial source). A frequency count of the records in the microdata file is usually generated using the key variable list. The most common rule applied in preparing public microdata files is the Threshold rule, or sometimes referred to as the k-anonymity rule. This rule requires a minimum number of records, of at least k records, (usually k=3), that are identical with respect to the specified set of key variables. This is also used as a risk measure in mu-ARGUS, a software product developed by Statistics Netherlands and the Computational Aspects of Statistical Confidentiality (CASC) project. (See Websites in Appendix B for further information on CASC).

The percent of records representing respondents who are unique in the population plays a major role in the disclosure risk of a microdata file. These records are often called **population uniques**. The records that represent respondents who are unique compared with everyone else in the sample are called **sample uniques**. Every population unique is a sample unique, however, not every sample unique is a population unique. There may be other persons in the population who were not chosen in the sample and whom have the same characteristics as a person represented by a sample unique. Statistical Policy Working Paper 2 states that "uniqueness in the population is the real question, and this cannot be determined without a census or administrative file exhausting the population." This corollary remains true for each individual record on a sample microdata file. Several methods of estimating the percent of population uniques on a sample microdata file have been developed. These methods are based on subsampling techniques, the equivalence class structure of the sample together with the hypergeometric distribution, and modeling the distribution of equivalence class sizes (Bethlehem, Keller, and Pannekoek, 1990; Steel, 2004; Winkler, 2004).

A measure of relative risk for two versions of the same microdata file has been developed using the classic entropy function on the distribution of equivalence class sizes (Greenberg and Zayatz, 1992). For example, one version of a microdata file may have few variables with a lot of detail on those variables while another version may have many variables with little detail on those variables. Entropy, used as a measure of relative risk, can point out which of the two versions of the file has a higher risk of disclosure.

B.1.a. MASSC.

Another measure of risk used in the Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC) disclosure limitation method (discussed later in Section B.3.d) creates sets of identifying variables, called strata, to find records that may be at risk of disclosure. A unique record in a stratum is a record whose profile is unique for a given set of identifying variables. The record is at risk of disclosing personal information if the record is unique among the set of identifying variables. After categorizing the database into a series of strata represented by different sets of identifying variables, a disclosure risk measure is calculated for each stratum. Unique records falling in a stratum are then assigned a disclosure risk associated with that stratum. MASSC computes four measures of risk to generate an upper bound measure of disclosure risk for a target record, stratum, or file. A measure of disclosure risk is calculated based on whether the target looks like a unique, a non-unique double, a non-unique triple, or a non-unique-four-plus, i.e., a non-unique cluster size of four records or more. An overall measure of the target is generated by taking a weighted average of the four disclosure risk measures where the weights are the relative proportion of each type of record in the adjusted database. By collapsing over the strata, a disclosure risk can be calculated for an entire database as well as an individual record.

B.1.b. R-U Confidentiality Map.

This approach attempts to measure the simultaneous impact on disclosure risk and data utility of applying a specific disclosure limitation technique and can serve as a tool by a data provider for choosing the appropriate parameter value. R is a numerical measure of the statistical disclosure risk in a proposed release of a data file. This could be measured by the percentage of records that can be correctly re-identified using record linkage software. U is a numerical measure of the data utility of the released file. This could be measured by comparing the mean values or the variance-covariance matrix of the original data and the perturbed data. By mapping the values of R and U on the Y and X axis, a confidentiality map is generated which shows the trade offs between, the gains, if any, in reducing disclosure risk by changing the parameters of the disclosure limitation procedure, and the loss in the usefulness of the data by changes in the analytical properties of the file. R-U Confidentiality Map can be constructed for different disclosure limitation techniques and serve as a useful tool in applying a specific disclosure limitation methodology. (Duncan, McNulty, and Stokes, 2001)

B.2. Methods of Reducing Risk by Reducing the Amount of Information Released

Recoding variables into categories is one commonly used way of reducing the disclosure risk of a microdata file (Skinner, 1992). The resulting information in the file is no less accurate, but it is less precise. This reduction in precision reduces the ability of an intruder to correctly link a respondent to a record because it decreases the percent of population uniques on the file. Recoding variables can also reduce the high risk of some records. For example, if occupation is on the file in great detail, a record showing an occupation of United States Senator in combination with a geographic identifier of Delaware points to one of two people. Other variables on the file would probably lead to the identification of that respondent. Occupation could be recoded into fewer, less discriminatory categories to alleviate this problem.

If an agency is particularly worried about an outside, potentially matchable file, the agency may recode the variables common to both files so that there are no unique variable combinations on the microdata file, thus preventing one-to-one matches. For example, rather than release the complete date of birth, an agency might publish only year of birth. Rounding values, such as rounding income to the nearest one thousand dollars, is also a form of recoding.

Another commonly used way of reducing the disclosure risk of a file is through setting top-codes and/or bottom-codes on continuous variables (see Section II.D.2). A **top-code** for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is not published on the microdata file. In its place is some type of flag that tells the user what the top-code is and that this value exceeds it. For example, rather than publishing a record showing an income of $2,000,000, the record may only show that the income is > $150,000. Similarly, a **bottom-code** is a lower limit on all published values for a variable. Top- and bottom-coding reduce the high risk of some records. Examples of top-coded variables might be income and age for demographic microdata files and value of shipments for establishment microdata files. If an agency published these variables on a microdata file with no top-coding, there would probably be a disclosure of confidential information. Examples of bottom-coded variables might be year of birth or year built for some particular structure.

Recoding and top-coding obviously reduce the usefulness of the data. However, agencies could provide means, medians, and variances of the values in each category and of all top-coded values to data users to compensate somewhat for the loss of information. Also, recoding and top-coding can cause problems for users of time series data when top-codes or interval boundaries are changed from one period to the next.

B.3. Methods of Reducing Risk by Disturbing Microdata

Since Statistical Policy Working Paper 2 was published, researchers have proposed and evaluated several methods for disturbing microdata in order to limit disclosure risk. These techniques, described in Chapter II, slightly alter the data in a manner that hinders an intruder who is trying to match files.

Probably the most basic form of disturbing continuous variables is the addition of, or multiplication by, random numbers with a given distribution This **noise** may be added to the data records in their original form or to some transformation of the data depending on the intended use of the file. Probability distributions can be used to add error to a small percent of categorical values. An agency must decide whether or not to publish the distribution(s) used to add noise to the data. Publishing the distribution(s) could aid data users in their statistical analyses of the data but might also increase disclosure risk of the data. Another proposed method of disturbing microdata is to randomly choose a small percent of records and blank out a few of the values on the records (see Section II.D.5). Imputation techniques are then used to impute for the values that were blanked.

B.3.a. Data Swapping

Swapping (or **switching**) and **rank swapping** are two proposed methods of disturbing microdata. The purpose of any swapping methodology is to introduce uncertainty so that the data user doesn't know whether real data values correspond to certain records. Records with a high risk of disclosure are usually selected for swapping. In the swapping procedure, a small percent of records are matched with other records in the same file, perhaps in different geographic regions, on a set of predetermined variables that are used as swapping attributes. The values of variables used as swapping attributes in the file are then swapped between the two records. In the rank swapping procedure, values of continuous variables are sorted and values that are close in rank are then swapped between pairs of records. As the percentage of swapped records increases, the greater the losses in data utility of the microdata file. Although swapping does not change the marginal distribution of any variable in a file, it does distort joint distributions involving both swapped and unswapped variables.

B.3.b. Data Shuffling

Data Shuffling is another data masking procedure that has been successfully applied to numerical data. The procedure involves two steps: first the values of the confidential variables are modified and second, a data shuffling procedure is applied to the confidential variables on the file. This method preserves the rank order correlation between the confidential and non-confidential attributes, thereby maintaining monotonic relationships between attributes.

Before the data are perturbed, the non-confidential variables (**S**) and confidential variables (**X**) on the file are identified. The conditional distribution of $f(\mathbf{X}|\mathbf{S} = \mathbf{s_i})$ between the confidential and non-confidential variables is then derived. For i = 1 to n, generate a vector $\mathbf{y_i}$ from $f(\mathbf{X}|\mathbf{S} = \mathbf{s_i})$. The perturbed values of **Y** are the collection of the values $\mathbf{y_i}$ (i = 1, 2, …n).

The shuffling of data records occurs after the values for the confidential variable have been perturbed and ranked. For each confidential variable let $Y = (y_1, y_2, \ldots, y_n)$ represent the perturbed values of the confidential variable $X = (x_1, x_2, \ldots, x_n)$. Let $X^j = (x^1, x^2, \ldots, x^n)$ represent the rank ordered values of X.

For i = 1 to n: Find the rank of y_i. Let this rank be k. Replace the value of y_i by x^k.
In the example below, notice that the rank of the first perturbed observation is 17. The value of the 17^{th} ordered value of X (x^{17}) = 42.79. Hence, the first perturbed observation is replaced by 42.79. Similarly, the rank of observation y_2 is 16 and is replaced by x^{16} = 41.74, the value of X at the 16^{th} rank of X. The process is repeated for every perturbed observation until the all of the perturbed values are replaced with original values from the confidential variable.

Example Data Set

ID#	S	X	Rank of X	Perturbed Y	Rank of Perturbed Y	Shuffled Y
1	41	54.24	27	43.8024	17	42.79
2	53	52.98	25	43.7608	16	41.74
3	40	33.77	4	31.2382	3	32.54
4	51	43.15	18	41.6440	13	40.41
5	37	48.70	22	36.3746	8	36.94
6	41	41.74	16	43.6570	15	40.77
7	24	36.00	7	46.5293	20	46.80
8	57	48.06	21	51.1033	23	48.76
9	52	57.69	29	54.3518	28	55.21
10	27	34.14	5	42.1101	14	40.72
11	39	32.54	3	40.6861	11	38.79
12	54	55.21	28	48.5196	22	48.70
13	52	40.77	15	53.7893	26	53.19
14	47	48.76	23	41.5140	12	39.50
15	41	27.52	1	44.6543	19	45.35
16	52	50.36	24	40.2965	10	38.68
17	20	42.79	17	34.6577	6	35.43
18	42	39.50	12	40.1456	9	38.05
19	52	53.19	26	51.5981	24	50.36
20	45	40.72	14	32.4994	4	33.77
21	52	38.68	10	47.7596	21	48.06
22	42	46.80	20	32.9835	5	34.14
23	50	59.08	30	44.4699	18	43.15
24	48	32.28	2	51.8446	25	52.98
25	33	36.94	8	35.7985	7	36.00
26	50	38.05	9	54.5523	29	57.69
27	46	40.41	13	25.2914	1	27.52
28	43	38.79	11	54.1997	27	54.24
29	56	45.35	19	54.7677	30	59.08
30	41	35.43	6	29.0405	2	32.28

The marginal distribution of the masked (Shuffled Y) variable is the same as that of the original variable X and the product moment correlation (linear relationships) and rank order correlation (non-linear monotonic relationships) are not disturbed. In the example provided the correlation between (S and X) is 0.4507 and that between (Shuffled Y and S) is 0.4474. The rank order correlation between (S and X) is 0.52 and that between (Shuffled Y and S) is 0.54. These estimates will approach each other as the size of the data set increases.

B.3.c. Data Blurring and Microaggregation

Blurring involves aggregating values across small sets of respondents for selected variables and replacing a reported value (or values) by the aggregate. Different groups of respondents may be formed for different data variables by matching on other variables or by sorting the variable of interest (see Section II.D.6). Records are placed in groups of size k, where k is commonly set between 3 and 10 and the original values associated with sensitive variables are replaced with the aggregate value. Data may be aggregated across a fixed number of records, a randomly chosen number of records, or a number determined by (n, k) or p-percent type rules as used for aggregate data. For a definition of the (n, k) and p-percent rules, see Chapter IV. The aggregate associated with a group may be assigned to all members of the group or to the "middle" member (as in a moving average). Aggregating over groups of 3 records or less may not be sufficient for reducing the risk of disclosure, especially if the blurring is performed on only one or two variables in a file. As the size of the group of records increases, the chance of re-identification is reduced. If the grouping is larger than 10 records, there may be greater distortion introduced into the microdata file which may lead to inaccurate published data. **Microaggregation** is a form of data blurring where records are grouped based on a proximity measure of all variables of interest, and the same groups of records are used in calculating aggregates for those variables. Blurring and microaggregation may be done in a way to preserve variable means. However, single variable data blurring or microaggregation may lead to re-identification and therefore should be combined with other disclosure limitation techniques to provide adequate data protection.

Another proposed disturbance technique involves super and subsampling (Cox and Kim, 1991). The original data are sampled with replacement to create a file larger than the intended microdata file. Differential probabilities of selection are used for the unique records in the original data set, and record weights are adjusted. This larger file is then subsampled to create the final microdata file. This procedure confuses the idea of sample uniqueness. Some unique records are eliminated through non-selection, and some no longer appear to be unique due to duplication. Some non-unique records appear to be unique due to nonselection of their clones (records with the same combination of values). Biases introduced by this method could be computed and perhaps released to users as a file adjunct.

B.3.d. Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC)

Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC) is a disclosure limitation methodology that consists of the following four major steps. The first step, Micro Agglomeration, partitions the records into risk strata in preparation for the level of modification to the data to reduce the risk of disclosure. Some recoding of variables is done during this phase. Individuals in each risk stratum are grouped so that the variance is small with respect to a given key set of identifying variables. In the second step, Substitution, values of sensitive variables are swapped with values from records that are the closest to them in terms of a certain distance measure. In the third step, Subsampling, records are randomly selected for subsampling within each strata. In the fourth step, Calibration, weights are assigned to records using certain key variables to preserve the domain counts from the original dataset. The calibration step reduces bias due to the substitution and it reduces variance due to the subsampling step. In the methodology, every record in the database is subject to modification or swapping, however, when applying this methodology, only a small random portion of the records are actually modified. (Yu, Dunteman, Dai, and Wilson, 2004).

B.4. Methods of Reducing Risk by Using Simulated Microdata

B.4.a. Latin Hypercube Sampling.

Latin Hypercube Sampling (LHS) is another technique that involves creating a replacement file containing replacement values for the sensitive variables in the microdata file. The LHS method ensures that the synthetic data set has nearly the same univariate statistical characteristics of the original data such as mean, standard deviation and coefficient of skew ness. LHS can be used to generate a synthetic data set for a group of uncorrelated variables. In the case where the variables are correlated, a restricted pairing algorithm is first applied to reproduce the rank correlation structure of the real data. Variables are first shuffled on the file and a cumulative distribution function is created for selected variables and used to generate the synthetic values. (Dandekar, Cohen, Kirkendall, 2001). Latin Hypercube Sampling provides one method of using multiple imputation techniques to produce a set of pseudo-data with the same specified statistical properties as the true microdata.

B.4.b. Inference-Valid Synthetic Data.

Another variation in the use of synthetic data for releasing public use data files is by drawing samples from the posterior predictive distribution of the adjusted confidential data. In this approach, the actual confidential variable(s) in the micro data file, Y, are replaced using some controlled data adjustment constraint algorithm. The initial step generates a predicted value for Y and a residual for each Y variable 10 times, called “implicates.” Statistical models using the data can then average the results from the ten implicates to generate standard error estimates. Depending on the variables which need protection and the variables that the researcher is interested in, the values for the confidential variable can be replaced by a posterior predictive distribution for that confidential variable based on a given set or combinations of variable keys.

By customizing the distribution of the predicted Y's plus the residuals for the relevant confidential variable, i.e., the posterior predictive distribution, various micro datasets can be created and the statistical inferences from the synthetic data are valid with the inferences generated by the actual reported values. Multiple public use files can be created from the same underlying data using this method with each public use file customized to different groups of users. The inference valid synthetic data methodology was applied to the Survey of Income and Program Participation (SIPP) data after the SIPP data was linked to earnings data from the Social Security Administration. (Abowd and Lane, 2003).

B.4.c. The FRITZ Algorithm for Disclosure Limitation.

The Federal Reserve Imputation Technique Zeta (FRITZ) system is used for both missing value imputation and disclosure limitation in the Survey of Consumer Finances (SCF). The FRITZ model reviews the data along a sequential predetermined path and imputes values one (sometimes two) at a time. The model is also iterative in that it imputes for the missing values in the data file, and then uses that information as a basis for imputing values in the second step, and continues the process until all values for the missing or sensitive estimates are stabilized and final. The file is reviewed for variable keys that cause excessive disclosure risks and those cases are selected for protection. All dollar values in the SCF are set to missing and the FRITZ algorithm is applied to generate imputed values. The subsequent analysis of this methodology indicates that while the imputations provided the protection to the sensitive individual records, it had only minimal effects on the distributional characteristics of the file (Kennickell, 1998).

B.5. Methods of Analyzing Disturbed Microdata to Determine Usefulness

There are several statistical tests that can be performed to determine the effects of disturbance on the statistical properties of the data. These include the Kolmogorov-Smirnov 2-sample test, Fischer's z-transformation of the Pearson Correlations, and the Chi-Square approximation statistic to the likelihood ratio test for the homogeneity of the covariance matrices.

These procedures are mainly conducted to see if the means and the variance-covariance and correlational structure of the data remain the same after disturbance. Even if these tests come out favorably, disturbance can still have adverse effects on statistical properties such as means and correlational structure of subsets and on time series analyses of longitudinal data. If an agency knows how the file will be used, it can disturb the data in such a way that the statistical properties pertinent to that application are maintained. However, public-use files are available to the entire public, and they are used in many ways. Levels of disturbance needed to protect the data from disclosure may render the final product useless for many applications. For this reason, agencies limit the amount of modification to the data in the microdata file, or attempt to limit disclosure risk by limiting the amount of information in the microdata files. Disturbance may be necessary, however, when potentially linkable files are available to users, and recoding efforts do not eliminate population uniques.

C. Necessary Procedures for Releasing Microdata Files

Before publicly releasing a microdata file, a statistical agency must attempt to preserve the usefulness of the data, reduce the visibility of respondents with unique characteristics, and ensure that the file cannot be linked to any outside files with identifiers. While there is no method of completely eliminating the disclosure risk of a microdata file, agencies should perform the following steps before releasing a microdata file to limit the file's potential for disclosure. Statistical agencies have used most of these methods for many years. They continue to be important.

C.1. Removal of Identifiers

Obviously, an agency must purge a microdata file of all direct personal and institutional identifiers such as name, address, Social Security number, and Employer Identification number. An internal file with the names or other direct identifiers removed may still be at risk of **indirect disclosure**, if sufficient data are left on the file with which to match with information from an external source that *also contains names or other direct identifiers.* In such a case, the identity, as well as all information in the file associated with that person or establishment will be disclosed if the file is released without further modifications.

C.2. Limiting Geographic Detail

The match does not need to be exact. An intruder could link the characteristics of all respondents with the same sample unit with similar information from an external source of data. Other variables on a file may cause an indirect disclosure problem if they could be used to distinguish a small geographic unit on the basis of certain socioeconomic characteristics. Once an individual or establishment’s records are associated with a small geographic area, the possibility of identification is greatly increased. Geographic location is a characteristic that appears on most microdata files. Agencies should give geographic detail special consideration before releasing a microdata file because it is much easier for an intruder to link a respondent to the respondent's record if the intruder knows the respondent's city, for example, rather than if he or she only knows the respondent's state.

Based on these considerations, the Census Bureau does not identify any geographic region with less than 100,000 persons in the sampling frame. A higher cut-off is used for surveys with a presumed higher disclosure risk. Microdata files from the Survey of Income and Program Participation, for example, still have a geographic cut-off of 250,000 persons per identified region. Agencies releasing microdata files should set geographic cut-offs that are simply lower bounds on the size of the sampled population of each geographic region identified on microdata files. This is easier said than done. Decisions of this kind are often based on precedents and judgment calls. More research is needed to provide a scientific basis for such decisions (Zayatz, 1992a).

Some microdata files contain contextual variables. Contextual variables are variables that describe the area in which a respondent or establishment resides but do not identify that area. In general, the areas described are smaller than areas normally identified on microdata files. Care must be taken to ensure that the contextual variables do not identify areas that do not meet the desired geographic cut-off. An example of a contextual variable that could lead to disclosure is average temperature of an area. The Energy Information Administration adds random noise to temperature data (because temperature data are widely available) and provides an equation so the user can calculate approximate heating degree-days and cooling degree-days (important for regression analysis of energy consumption).

C.3. Top-Coding High Risk Variables That Are Continuous.

The variables on microdata files that contribute to the high risk of certain respondents are called **high-risk variables**. Examples of continuous high-risk variables are income and age for demographic microdata files and value of shipments for establishment microdata files. As stated previously, if an agency published these variables on a microdata file with no top-coding, there would probably be a disclosure of confidential information. For example, intruders could probably correctly identify respondents who are over the age of 100 or who have incomes of over one million dollars.

Appropriate top-codes (and/or bottom-codes in some cases) should be set for all of the continuous high-risk variables on a microdata file. Top-coded records should then only show a representative value for the upper tail of the distribution, such as the cut-off value for the tail or the mean or median value for the tail, depending on user preference. Angle (2003) developed a methodology for estimating the distribution of top coded values using a distribution more general than the traditional Pareto, and illustrates it using annual wage and salary income. The model's estimate of the right tail truncated by top-coding has been shown to have many of the dynamics of the right tails of empirical annual wage and salary income distributions. This methodology uses a probability density function model for generating the right tail of an income distribution that has been truncated by top-coding. The model's parameters are estimated in the fit of the model to data below the cutoff for top-coding. The model's right tail is used in the estimation of statistics of the whole distribution. The model is able to generate the distribution of top coded values even after lowering the threshold level for minimum top-codeable annual wage and salary income well below the 99^{th} percentile. (Angle, 2003).

C.4. Precautions for Certain Types of Microdata

There are certain types of microdata that may raise the risk of disclosure when reviewing a file for release.

C.4.a. Establishment Microdata

Most microdata files that are publicly released contain demographic microdata. It is presumed that the disclosure risk for establishment microdata is higher than that for demographic microdata. Establishment data are typically skewed, the size of the establishment universe may be small, and there are many high-risk variables on potential establishment microdata files. Industry publications and trade associations may also exist and function as outside sources of information for a data user. Publicly available administrative databases may also be available for matching to the establishment microdata files and create additional disclosure risks. Also, there are a large number of subject matter experts and many possible motives for attempting to identify respondents on some types of establishment microdata files. For example, there may be financial incentives associated with learning something about the competition. Agencies should take into account all of these factors when considering the release of an establishment microdata file.

C.4.b. Longitudinal Microdata

There is greater risk when the microdata on a file are from a longitudinal survey where the same respondents are surveyed several times. Risk is increased when the data from the different time periods can be linked for each respondent because there are much more data for each respondent and because changes that may or may not occur in a respondent's record over time could lead to the disclosure of the respondent's identity. Agencies should take this into account when considering the release of such a file. One piece of advice is to plan ahead. Releasing a first cross-sectional file without giving any thought to future plans for longitudinal files can cause unnecessary problems when it comes to releasing the latter. The entire data collection program should be considered in making judgments on the release of public-use microdata.

C.4.c. Microdata Containing Administrative Data

The disclosure risk of a microdata file is increased if it contains administrative data or any other type of data from an outside source linked to the survey data. Those providing the administrative data could use that data to link respondents to their records. This is not to imply that providers of administrative data would attempt to link files, however, it is a theoretical possibility and precautions should be taken. At the very least, some type of disturbance should be performed on the administrative data or the administrative data should be categorized so there exists no unique combination of administrative variables. This reduces the possibility that an intruder can link the microdata file to the administrative file. There are concerns that agencies should not release such microdata at all or should release it only under a restricted access agreement.

C.4.d. Consideration of Potentially Matchable Files and Population Uniques

Statistical agencies must attempt to identify outside files that are potentially matchable to the microdata file in question. Comparability of all such files with the file in question must be examined. The Census Bureau uses re-identification and record linkage experiments to

determine if their files are matchable to outside files on a certain set of key variables. The National Center for Education Statistics matches microdata files under consideration for release to commercially available school files to identify unique matches. Re-identification of microdata refers to the ability to use public available information to attach names, addresses, and other partially unique identifiers to individual records in a public-use file. An identifier is partially unique if it can be used in conjunction with other variables to re-identify a record even though it may not exactly identify a linkage between two records by itself. Record linkage software has been developed to handle a large variety of both minor and major spelling variations and errors in the variables used in the matching process.

Another measure of the risk of re-identification for a file is the number or proportion of population uniques, where consideration is restricted to those variables thought to be available on external files. Statistical models have been developed that relate the distribution of the sample uniques in a file to the distribution of the population uniques. However, these models only provide an estimate for the percentage of sample uniques that are true population uniques. This estimate tends to have a high variance and estimating the percentage doesn't provide any guide to determining which uniques are artifacts of sampling and which are population uniques. Record linkage experiments also can provide a measure of re-identification risk, but are heavily dependent on acquiring or modeling external data sources (Winkler 2004). A record linkage experiment may identify some population uniques, but should not be considered as an assurance that all risky records have been discovered.

D. Stringent Methods of Limiting Disclosure Risk

There are a few procedures that can be performed on microdata files prior to release that severely limit the disclosure risk of the files such as data swapping and data coarsening. One must keep in mind, however, that the usefulness of the resulting published data will also be extremely limited. The resulting files will contain either much less information or information that is inaccurate to a degree that depends on the file and its contents.

D.1. Do Not Release the Microdata

One obvious way of eliminating the disclosure risk of microdata is to not release the microdata records. The statistical agency could release only the variance-covariance matrix of the data or perhaps a specified set of low-order finite moments of the data. This greatly reduces the usefulness of the data because the user receives much less information and data analyses are restricted.

D.2. Recode Data to Eliminate Uniques

Recoding the data in such a way that no sample uniques remain in the microdata file is generally considered a sufficient method of limiting the disclosure risk of the file. A milder procedure allowing for broader categorization--recoding such that there are no population uniques--would suffice. Recoding the data to eliminate either sample or population uniques would likely result in

very limited published information.

D.3. Disturb Data to Prevent Matching to External Files

Showing that a file containing disturbed microdata cannot be successfully matched to the original data file or to another file with comparable variables is generally considered sufficient evidence of adequate protection. Several proximity measures should be used when attempting to link the two files. An alternative demonstration of adequate protection is that no exact match is correct or that the correct match for each record on a comparable file is not among the K closest matches. Microaggregation or data shuffling could be used to protect data, perhaps using (n, k) or p-percent type rules as used for tables. In this way, no individual data are provided, and intruders would be prevented from matching the data to external files. See Chapter IV for a definition of the (n, k) and p-percent rules. Microaggregation, data blurring, and other methods of disturbance that hinder file matching, however, may cause distortions in published data. Taken to a degree that would absolutely prevent matching, the methods would usually result in greatly distorted published information.

E. Conclusion

Public-use microdata files are used for a variety of purposes. Any disclosure of confidential data on microdata files may constitute a violation of the law or of an agency's policy and could hinder an agency's ability to collect data in the future. Short of releasing no information at all, there is no way to completely eliminate disclosure risk. However, there are techniques that, if performed on the data prior to release, should sufficiently limit the disclosure risk of the microdata file. Research is needed to understand better the effects of those techniques on the disclosure risk and on the usefulness of resulting data files (see Section VI.A.2).