Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 455–459

Discussion
Statistical Disclosure Limitation

Brian V. Greenberg[1]

National statistical agencies and offices collect information about a nation's population and institutions and make the information available to the public. Data are often collected under pledges of confidentiality and agencies are usually required by law to respect the privacy of respondents. Statistical agencies have the responsibility to design data release strategies which will not violate pledges of nondisclosure either through intent or neglect. In addition to legal concerns, statistical offices must be mindful that violating pledges of confidentiality may undermine an agency's ability to collect data due to loss of public trust and confidence. A statistical organization seeks to release detailed information to the public to support research and policy analysis; however, it is through fine levels of accurate detail in data items that high risk of disclosure may arise. An important area for statistical research lies in developing methods which allow for public release of as much useful information as possible while abiding by the legal requirements to protect individual privacy and adhering to pledges of data confidentiality given to respondents.

Broadly speaking, the goal of a statistical office is to maximize the level of information provided to the public subject to the requirement that the risk of disclosure be acceptably low. Currently there are no clear ways to quantify "level of information" or to quantify "risk of disclosure," yet these factors must enter into any proposed data release strategy. Development of techniques to quantify and evaluate these concepts and the application of this understanding to the design of data release strategies are important disclosure avoidance research imperatives. Both of these issues have been addressed in the papers by Fuller and Little.

Microdata records are data records at the respondent level, and the risk in the release of a microdata file is that an investigator may be able to determine the identity of a respondent who provided information. In so doing, an investigator will learn about the respondent through information which was provided to the data collection organization and is contained in the microdata record. The respondent is typically a person or household from a demographic survey or census. The respondent can be an economic establishment if the microdata file contains responses from businesses. The objective of a data release strategy for microdata is to provide as much useful data as possible for the user community while not allowing an investigator access to information to link a record to the respondent.

The risk of linking a microdata record to a respondent can arise from two sources.

[1] U.S. Bureau of the Census, FOB 4, Room 2017, Washington, D.C. 20233, U.S.A.

The first can be thought of as recognition due to a combination of highly visibile characteristics which are unique or unusual on a record and which would allow a person who is knowledgeable about the population to recognize a respondent. In such cases, the threat is not a computer match, but rather unique characteristics which may serve to highlight a person well known to those familiar with the universe from which the sample cases belong. The other source of risk arises from a computer matching of the public use microdata file to files held by any other organization, either public or private. Based on the number and complexity of overlapping fields on both files, the risk is that an investigator may match one file against the other and with a high degree of confidence accurately link records. The fields on a microdata file which may be used to link to another data set are sometimes referred to as *key* variables.

To determine the ability to link microdata files one must consider a number of nonstatistical issues. Exactly what data files are available for matching based on characteristics? How many overlapping fields are there on both files? How old are the data on either file? How comparable are the data on the different files? Are the externally held files in easily accessed computer format or are they paper files in a file cabinet? What would be the cost to an investigator and what is the probability of success should an attempt be made to link files; and what are potential benefits? Will an investigator operate in a rational mode: that is, if the costs are high and probability of success and benefits low, can we say an investigator would not undertake such an activity? All these items are very hard, if not impossible, to quantify but they must enter into planning for any data release strategy.

One method for reducing the amount of information identifying respondents is through *grouping* response information into categories. For example, rather than releasing exact date of birth, an agency might release month of birth, or quarter of birth, or perhaps year of birth. We do not release respondent address on a public use microdata file, however, we can release city of residence, or county of residence, or state of residence; or we can release no information at all about the residence of respondents (i.e., we can release a nation file).

Percentages can be recoded into deciles or quintiles. Income can be recoded into intervals of size, for example, $4,000 for incomes up to $100,000 and all incomes equal to or greater than $100,000 can be recoded as "$100,000 or more." That is, one can set a topcode of $100,000 on income. Virtually all quantitative variables released on public use microdata files are topcoded because cases in the tails of a distribution may be highly visible or may render a case vulnerable to identification. Examples of topcoded items include income, monthly rent, value of principal residence, etc. Some variables are also bottomcoded, such as year of birth, year home was built, and so on.

Another method for masking data records to prevent accurate linking of a public use microdata file to an external file is by distorting values prior to release. One method for data distortion in microdata files is through the introduction of an error term sometimes referred to as noise. Under *noise introduction* for continuous variables, actual data values are altered by a small amount in some specified manner. Typically, one thinks of an error term being added to response values or to transformations of response values

$$x + \epsilon \text{ or } T(x) + \delta$$

where ϵ and δ are error terms and $T(x)$ is a transformation of the variable x. The new

masked values

$$x' = x + \epsilon \text{ or } x' = T^{-1}[T(x) + \delta]$$

are released to the public.

When matching an external file against a microdata file having noise, there will be false matches. Due to the introduction of noise the most likely match based on some notion of distance built into the matcher may not be a true match. Some matchers have the ability to attach a likelihood of a true match and then select as matches only those candidates whose likelihood exceeds a specified threshold. This measure is probabilistic and a "likely" match may be a false match. There is an inverse relation between the ability of an investigator to match well to an external file and the level of noise added prior to microdata release.

There is also an inverse relation between the amount of noise added and the usefulness of masked data. The introduced noise to prevent excessive correct matching will also reduce the utility of the data. If a data set is useful for some purposes, it may yield very distorted information when used for others. For these reasons, masking data through noise introduction is not a typical data release strategy for multipurpose files.

For a particular individual user with specified and well-defined uses, a noise introduction scheme can be used to create a public release microdata file. Such a scheme can be tailored to the anticipated user needs. Typically one would expect a fairly sophisticated data user to request such an arrangement. However, for a broad-based, general purpose product which will be employed for a wide range of applications by a variety of potentially unsophisticated users, such a strategy has many problems.

First of all, even though noise might have been added to the data to preserve some distributional properties of the entire data set, subsets of the data may suffer significant distortions. In addition, the levels of noise required to protect highly skewed populations may render the final product quite unacceptable for many uses. When distortions are introduced into the data, one may introduce improbable or inconsistent response combinations on a record. This may be particularly true as it applies to categorical data.

One of the major drawbacks in the use of noise introduction procedures for standard public-use microdata files is that the final product looks like genuine data. The data will convey a specious sense of precision to users – especially unsophisticated users – and the problems described above can be exacerbated greatly. In addition, such data do not convey the information that disclosure avoidance measures have been applied. It is important for the data releasing agency that this fact be clearly recognized by data users and all segments of the public with concern for this issue.

For some of the reasons listed above, releasing data in grouped categories is a primary disclosure avoidance measure for public-use microdata files. It would have been nice to see further discussion on the uses of grouped data and some research into grouping methodologies in these papers. For example, how does one design an optimal group allocation strategy taking into account both data protection and data utility?

Both authors refer to the relation between noise introduction and imputation. In fact, there is a strong relation between grouping and noise introduction through the vehicle of imputation. That is, consider a microdata record of responses, x_i

$$(x_1, ..., x_n)$$

and replace each response by a category Y_i

where $x_i \in Y_i$, to form a record of categories

$$(Y_1, ..., Y_n).$$

The vector $(Y_1, ..., Y_n)$ is a typical record of grouped data.

To create a surrogate record of responses, we can impute for the first variable, v_1, subject to the condition that the remaining variables $v_i \in Y_i$ with $i = 2, ..., n$. Add the further restriction that the value of v_i be in the category Y_1 and denote the value of the first imputed variable by z_1. Impute next for the second variable, v_2, subject to the condition that $v_i \in Y_i$ for $i \neq 2$, and further require that the impute, z_2, be in Y_2. Continue in this way to obtain a record of imputes

$$(z_1, ..., z_n)$$

where $z_i \in Y_i$ recalling that $x_i \in Y_i$. Letting

$$z_i - x_i = \epsilon_i$$

for continuous variables, we can think of

$$z_i = x_i + \epsilon_i$$

as variables with noise introduced. What can we say about such a record and such a procedure?

The level of protection is the same as under the strategy of releasing the grouped data $(Y_1, ..., Y_n)$. The data utility will be greater than for the release of the grouped data only for those attributes controlled for in the imputation process. Thus, for a special purpose product with specified anticipated uses, releasing the file of $(z_1, ..., z_n)$ may have some advantages. But for a general purpose, multiuser product, the problems outlined above pertain, and a file of grouped data $(Y_1, ..., Y_n)$ may serve the user community better.

The paper by Fuller is an excellent extension of his ground-breaking development of measurement error models. By adding noise for data protection as described by this paper, one can view the additional noise as further measurement error and the methods developed earlier by the author can be brought to bear for analysis of the data. The procedure is very sensitive to levels of protection and he adds only enough noise to make the likelihood small of linking a masked record to the original respondent data. In this sense the author does attempt to minimize the amount of noise added subject to the condition that the risk of reidentification (as defined in his model) is acceptably low. The procedures and framework developed by Fuller are very satisfying from both analytical and conceptional perspectives.

The downside of his work is that measurement error models have not been widely adopted by the user public. This point is acknowledged by the author. In the absence of employing such an analysis on the masked data, the benefits he has built into the procedure do not materialize. One is thus left with all the drawbacks of noise introduction without the advantages.

In the Little paper the goal is a statistical analysis of data masking procedures and subsequently produced data with a focus on model-based likelihood methods when applicable. The author considers a large number of masking procedures and initiates this analysis in each. He mentions the need for more analysis of grouped data – a point with which I concur. Little discusses the relation between imputation and the creation of noise introduction for disclosure avoidance. He brings to the analysis many of the considerations he has addressed over the years in his work on imputation methods. In this paper Little is taking initial steps setting the stage for more analysis of disclosure avoidance procedures along the lines of model-based, likelihood methods.

I encourage more work in the area of comparing the various disclosure avoidance methods with an emphasis on realistic applications and scenarios. There is no shortage of procedures for masking data and there is absolutely no reason to believe that any one procedure is uniformly better than all others for all applications. Much more work needs to be done in the area of determining under what circumstances one family of techniques is superior to others with an emphasis on user needs and under what circumstances one procedure is to be preferred to others for concrete applications.

It would be nice if some benchmark *set of statistics* can be identified against which one can test different data protection schemes. What are the attributes of a data set which we would like to see invariant under data transformation for disclosure avoidance? It would be nice if there were some benchmark *data set* on which different disclosure avoidance methods can be applied, examined by others working in this area, and possibly replicated. This would allow a more applied discussion of the effect of proposed methods on data and promote a better shared understanding of findings. By being able to take a critical look at the actual effects of methods for disclosure avoidance, researchers and users in this area will be better able to evaluate their merits.