Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 461–468

Discussion
Statistical Disclosure Limitation

Donald B. Rubin[1]

Issues of honoring confidentiality constraints when releasing microdata have always been important but appear to be becoming even more critical. Although there exists interesting work on masking data to preserve confidentiality, the valid analysis of specially masked microdata generally requires not only knowledge of which masking techniques were used, but also special-purpose statistical software tuned to those masking techniques. The proposal offered here is to release no actual microdata but only synthetic microdata, constructed using multiple imputation so that they can be validly analyzed using standard statistical software. The possible benefits and feasibility of this rather radical, but possibly stimulating, proposal are explored.

1. Context of Comments

Recently there has been an increased interest in issues concerning confidentiality of data, reflected, for example, by the Panel on Confidentiality, formed under the Committee on National Statistics of the National Academy of Sciences, and the recent special issue of *Statistica Neerlandica* dedicated to R.J. Mokken. The reasons given for this increase include both (1) an explosion in the demand for microdata from users, especially for data paid for with public funds whose analyses could have major social or medical implications, and (2) an explosion in the presentation of legal and ethical positions on the consequences of the improper release of confidential information. The major focus of the Panel on Confidentiality's efforts has been on the dominant legal, ethical, and logistical issues about which I have no professional wisdom to offer; the major focus of the special issue of *Statistica Neerlandica* was on methods for avoiding and measuring risks of inappropriate disclosure when releasing microdata.

Although much of this previous work is intriguing, my comments might be read as accepting as inevitably and hopelessly complex the legal, ethical and logistical issues surrounding the general release of any actual microdata in the future. My focus here is solely on statistical issues arising from the release of synthetic microdata sets for public-use, that is, data sets consisting of records of individual synthetic units rather than actual units.

[1] Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, MA 02138, U.S.A.
Acknowledgements: I would like to thank Fritz Scheuren, IRS, and Herman Haberman, OMB, for their encouragement to pursue these ideas, and Robert Groves for interesting commentary when this perspective was presented at a Washington meeting of the American Statistical Association in memory of William G. Madow. Furthermore, I wish to thank the National Science Foundation and the U.S. Census Bureau for providing partial support for work on these ideas.

In some cases, published summary tables or summary statistics can be adequate and obviate the need for microdata, and then the concern is with the ability of an intruder to violate confidentiality guidelines from these summaries. However, because it is difficult to contemplate even a small portion of the legitimate analyses at the data-release stage, releasing only some summary statistics (such as provided by tables of counts, means, variances, correlations, means of logged variables, etc.) is generally an insufficient response to demands for microdata. Also, in some cases actual microdata can be released "privately" to selected users, with appropriate sanctions for violations of confidentiality concerns, but this does not fully satisfy the demands for publicly available microdata.

My proposal is to initiate a research effort, based on the theory and applied successes of multiple imputation methodology (Rubin 1987), to develop a technology for releasing, for general public-use, only synthetic microdata sets, not one unit of which is an actual unit. This proposal is related to, but more radical than, suggestions to use missing data techniques to analyze masked data (e.g., Little 1993).

This research project is intended to supplement related work on the release of summary data and on the formation of special groups of users to receive actual microdata, but it is also intended to supplant current work on the release of specially masked microdata. Such masked data sets require special software to analyze, especially if the masking rules themselves are unknown or partially unknown to the user. Since the vast majority of users of public-use data will rely on standard statistical software, I believe it is imperative that public-use microdata not only preserve confidentiality, but moreover, preserve the user's ability to obtain valid statistical inferences using standard statistical methods; such concerns have received only limited attention from the statistical community. In principle, my proposal can satisfy both concerns, whereas the use of masked data cannot because of the need for special attendant errors-in-variables software for each analysis (e.g., Fuller 1993). Releasing complex coarsened-data versions of the raw data must inevitably lead to the need for special complex software for proper analysis. Despite the high technical quality of this work and related work on measuring disclosure risks (e.g., Skinner 1992), I think it is a false hope that special software will be developed for each combination of *analysis* × *masking method* × *database type*. Moreover, the users of microdata have their own science to worry about, and they cannot be expected to become experts at using and understanding complicated "demasking" programs operating on a data set whose entries do not look like any individual's microdata.

2. Implications of Proposal

My proposal has strong implications for the four groups most affected by the release of public-use microdata: survey units themselves, whose confidentiality is at stake; data producers, who are being squeezed between the demands of data users and concerns of maintaining confidentiality; legitimate data users, who want to obtain valid inferences from public-use files in as straightforward a manner as possible; and data snoopers, who want to discover information at a "too-personal" level. The implications for each of these groups is nontrivial, and these are discussed before describing the structure of the resultant synthetic multiply-imputed data bases.

First consider the implications for the survey units themselves. Under my propo-

sal, no actual unit's confidential data would ever be released. Rather, all actual data would be used to create a multiply-imputed synthetic microdata set of artificial units, and this artificial data set would contain essentially all of the legitimate (i.e., nonconfidential) information available in the actual microdata, where the meaning of "essentially all" is discussed in Section 3. With such a plan, data collectors could honestly tell potential respondents that "Your data will only be used to create synthetic data for public-use, and none of your data values will ever be released," and thus, confidentiality constraints would be automatically honored from each unit's perspective. The hoped-for implications are that survey units will be more likely to be respondents and to respond truthfully.

Next consider the implications for the data snooper. The fact that the confidential microdata and released units are entirely synthetic will be part of the documentation of the public-use file, so that all users, including those with sinister intents, will know that it contains no actual microdata. Knowing this fact should destroy intruders' interests in snooping at the individual level because each individual is synthetic and therefore the "individual" has no actual identification to be discovered. Snoopers must have better places to snoop!

From the perspective of a legitimate user of public-use microdata, the primary point to keep in mind is that such a user has no interest in any individual's data, and the only reason for analyzing individuals' data in the form of a public-use microdata set is to allow analyses for population (or subpopulation) estimands (such as population correlations and means within strata) using straightforward statistical tools. The desiderata from the user's end are that released microdata (1) should look like actual individual microdata in the sense that they must be analyzable using the full range of standard complete-data statistical tools, and (2) valid inferences for legitimate estimands should be easily obtainable. A multiply-imputed microdata set satisfies these criteria since multiply-imputed data can be validly analyzed simply using repeated applications of complete-data methods. Data users can be told that "Because the synthetic microdata looks just like replicated actual microdata, you can use standard complete-data software to draw inferences." Users should also be told: "Although valid inferences will be obtained, the standard errors will be larger than those from the actual microdata because there is a reduction in information relative to the actual microdata, and this is reflected by the between-imputation variability"; and "At some point, standard errors for highly detailed estimands will be so large that the survey is no longer useful for these estimands, but the same is true of any survey at some level of detail."

Finally, the implications of my proposal are substantial for the data producer, because the proposal requires a heavy investment in technology at the data collector's end so as to allow the production of a multiply-imputed public-use file. My attitude is that the data producer has the knowledge and can have the resources to build an appropriate file, whereas it is entirely unrealistic to expect the varied community of users to duplicate properly the efforts needed to decode an exotic masked data base. That is, we must allocate the resources where the problem can be solved, not recommend the duplication of resources in a hopeless attempt to address the problem. To see exactly what is involved requires consideration of how to create a multiply-imputed synthetic microdata base.

3. A Multiply-Imputed Synthetic Microdata Base

For simplicity, consider an actual microdata set of size n drawn using design D from a much larger population of N units, where X represents background variables, Z represents outcome variables with no confidentiality concerns, and Y represents outcome variables with some confidentiality concerns; X is, in principle, observed for all N units, whereas Z and Y are only observed for the n sampled units and are missing for the N-n unsampled units. A multiply-imputed *population* consists of the actual X data for all N units, the actual (Z, Y) data for the n units in the survey, and M (the number of multiple imputations, typically between 3 and 10) matrices of (Z, Y) data for the N-n nonsampled individuals. The variability in the imputed values ensures, theoretically, that valid inferences for population estimands can be obtained, at least for population estimands up to some level of detail. A model for predicting (Z, Y) from X is used to multiply-impute (Z, Y) in the population. The choice of model is influenced by the X data, the design D, scientific understanding of how to predict (Z, Y) from X, and judgment about the range of questions users may ask of the data.

Given such a multiply-imputed population and a new survey design D^* for the released microdata (possibly, but not necessarily, the survey design actually used to collect the data), we can draw a sample of n^* units ($n^* \ll N$), which looks structurally just like an actual microdata base of size n^* drawn from the population using design D^*, and we can do this M times to create M replicates of (Z, Y) values. The result is a multiply-imputed synthetic data base. Of critical importance, this data set releases only synthetic confidential (i.e., Y) data because $N \gg n$ and $N \gg n^*$. To make sure that no actual microdata are released, it might be wise to draw the samples from the population excluding the n actual units, and to guard against all M imputations of a Y variable being identical within a unit (e.g., make M large enough so that this event is very rare, and examine such cases before releasing them). The resultant multiply-imputed data base can be used to create M repeated microdata bases, each of which is analyzed using complete-data methods of analysis appropriate to design D^*; these repeated complete-data analyses are combined to create one valid inference for the population using the multiple imputation technology given in Rubin (1987), summarized, for example, in Rubin and Schenker (1991).

Clearly, the critical effort is the data producer's creation of the multiply-imputed population and subsequent creation of multiply-imputed microdata samples; the use of plural "samples" here is intentional, since there may be good reasons to draw several synthetic samples according to different designs, each tuned to be most informative for its own set of questions (e.g., weighted to represent different subpopulations). Also clear is the fact that simple hot-deck (or jackknife or bootstrap) methods that simply redraw observed units will not work because they would release actual confidential microdata. Some creative mix of implicit and explicit modelling will be required, and the research conducted in search of satisfactory methods should have many beneficial side-effects for handling missing data in complex surveys in general.

The correct way to address the feasibility of this prescription is by an example, even a relatively trivial one, that illustrates that something of use along these lines is possible. Before discussing an illustrative

example in Section 5, it is helpful to clarify the information content of a synthetic multiply-imputed database relative to the real data on which it is based.

4. Information Loss with Synthetic Data

Although it is intuitively clear that there must be a loss of information when comparing an actual microdata set with its associated multiply-imputed synthetic data set, several aspects of this information loss are rather subtle. An easy way to delineate the boundaries of the situation is to state two important facts.

Fact 1. The information in the actual data, consisting of actual (Z, Y) values from the survey coupled with the population values of X, is greater than the information contained in the multiply-imputed population; if the imputation model for predicting (Z, Y) from X is correct, as M increases, the information in the latter is essentially the same as the information in the former.

Fact 2. The information in the microdata consisting of the actual (X, Y, Z) values for the n actual sampling units may be greater than, equal to, or less than the information in the multiply-imputed synthetic microdata consisting of (X, Y, Z) values for the n^* synthetic units; the relationship will depend on the estimand under investigation, the relative sizes of n and n^*, the number of multiple imputations, the survey designs D and D^*, and the ability of X to predict (Z, Y).

At first reading, Facts 1 and 2 may appear paradoxical, but the apparent paradox is resolved by realizing that the actual microdata excludes X values known in the population for units not in the sample, and if X is a good enough predictor of (Z, Y), the (Z, Y) values in the population can be well known from the $X \rightarrow (Z, Y)$ relationship estimated in the microdata and the X values in the population; this is not a new insight, it is simply the basis for standard ratio and regression adjustments. The possibility that multiply-imputed synthetic data can be more precise than the microdata on which the imputations were based is documented in the example of Section 5.

Of course, in common practice, even if $n^* > n$ and design D^* is more efficient for the estimand under study than design D, the available X values known for the entire population may not be very powerful predictors of (Z, Y), and so we should be realistic and expect the synthetic microdata to have less information than the actual microdata for almost all estimands. Several possible approaches for moderating this conclusion are rather obvious.

For example, after having collected the actual microdata and selecting n^* units for the synthetic microdata, one possible modification to current practice is to conduct a supplemental survey of the n^* units to collect nonconfidential Z data that are predictive of Y. That is, the n actual microdata units are surveyed about both nonconfidential Z data predictive of Y and confidential Y, whereas the n^* units to be used as the basis of the synthetic microdata are only surveyed about the nonconfidential Z. Now with Z observed for the synthetic microdata base, if $n^* > n$, it may be that the information in the synthetic multiply-imputed data set is nearly as large as, or even larger than, the information in the actual microdata, at least for the variety of estimands well-captured by the models used for imputation.

Moreover, if users of the synthetic microdata find that the answers being obtained have too much between-imputation vari-

ability to satisfactorily address important legitimate questions, they could have the option of having their questions addressed by the actual microdata in one of two ways. First, they could apply for special status as private users with appropriate sanctions for violating confidentiality constraints. Second, they could submit their questions, possibly in the form of specific requests for software runs, to the data producer who would address their questions using the actual microdata and review the answers for violations of confidentiality constraints before releasing them. If appropriate burdens for making such requests are implemented and it is emphasized that the information increase from synthetic to actual microdata is typically small for legitimate estimands, such requests may be limited to fairly appropriate ones, and the burden of these special requests on the data producer may be outweighed by the straightforward benefits of preserving confidentiality.

In any case, two points are clear. First, the issue of information loss about legitimate estimands when moving from actual to synthetic microdata is complicated and is analogous to many issues of information loss when moving from larger to smaller surveys or from one survey design to another. Second, excessive information loss can be addressed by new supplemental, nonconfidential survey methods designed with the construction of the synthetic microdata set in mind, as well as by limited access to actual microdata by special request.

5. The Census Industry and Occupation Coding Project

This example, although simple relative to the general problems of release of microdata, is important because it demonstrates that the release of microdata with entirely synthetic data on Y variables can result in perfectly acceptable inferences about Y, at least as nearly as we can tell after years of trying to find deficiencies with the microdata. Only a brief summary of the salient points is given here; the basic plan and scope of work were outlined in Rubin (1983), and the multiply-imputed data set has been described and studied in Rubin and Schenker (1986, 1987), Rubin (1987), Treiman, Bielby and Cheng (1989), Weld (1989), Clogg, Rubin, Schenker, Schultz and Weidman (1991), and Schenker, Treiman and Weidman (1993).

The basic situation in this U.S. Census Bureau project is as follows. There exists one research file of approximately 10^5 units with X, Z and actual Y values, and a public-use file of 10^6 units with X, Z, and multiply-imputed Y values, which were created using models estimated from the 10^5 research-file units. In this context, the Y data are 1980 industry and occupation codes, and the X, Z data are extensive background information (e.g., gender, weeks worked per year, region of country), as well as 1970 industry and occupation codes. The 1980 codes were predicted from the background information and the 1970 codes using Bayesian logistic regression techniques. The resultant models were used to create five imputed 1980 codes for each of the million units on the public-use file; hence, here $n = 10^5$, $n^* = 10^6$, $M = 5$.

Experience with these data sets shows that inferences for population estimands involving Y based on the multiply-imputed public-use file appear to be valid, and often, perhaps even typically, *more* precise than the corresponding inferences based on the research file, despite the fact that only entirely synthetic Y values exist in the public-use file and all the real Y values exist in the research file. This result is not that

surprising considering the discussion in Section 4 because the public-use file has much more information on (X, Z) than does the research file, and (X, Z) are good predictors of Y; in particular, 1970 codes are very good predictors of 1980 codes, although there is still substantial residual variability (e.g., some 1970 codes map into more than 60 possible 1980 codes). Of course, if inferences were drawn using all the information from both files, they would be even more precise, but the essential point is that perfectly acceptable inferences for population estimands involving Y *can* be achieved using multiply-imputed data with entirely synthetic Y values – an existence theorem in a simple case.

6. Can This Work in General Practice?

The fact that a synthetic multiply-imputed data set can work in a particular case does not mean it can work in all situations with confidentiality concerns about the release of public-use microdata. Even if the likelihood of general success is small, however, the potential size of the payoff seems so great that, in statistical terms, the expected payoff (probability of success × size of payoff given success) is large, possibly substantially larger than the expected payoffs from competitive approaches with higher probabilities of success. Many difficult and complex modelling issues need to be addressed, but as in all such enterprises, attacking a specific problem or two with an eye on developing general tools seems to be the approach to take.

7. References

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression. Journal of the American Statistical Association, 86, 68–78.

Fuller, W.A. (1993). Masking Procedures for Disclosure Limitation. Journal of Official Statistics, 9, 381–404.

Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 405–424.

Rubin, D.B. (1983). Progress Report on Project for Multiple Imputation of 1980 Codes. Report distributed to the U.S. Bureau of the Census, the National Science Foundation, and the Social Science Research Council.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Rubin, D.B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81, 366–374.

Rubin, D.B. and Schenker, N. (1987). Interval Estimation from Multiply-Imputed Data: A Case Study Using Census Agriculture Industry Codes. Journal of Official Statistics, 3, 375–387.

Rubin, D.B. and Schenker, N. (1991). Multiple Imputation in Health-Care Databases: An Overview and Some Applications. Statistics in Medicine, 10, 585–598.

Schenker, N., Treiman, D.J., and Weidman, L. (1991). Analyses of Public-Use Data with Multiply-Imputed Industry and Occupation Codes. Applied Statistics, 42, 545–556.

Skinner, C.J. (1992). On Identification Disclosure and Prediction Disclosure for Microdata. Statistica Neerlandica, 46, 21–32.

Treiman, D.J., Bielby, W., and Cheng, M. (1989). Evaluating a Multiple-Imputa-

tion Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard. Sociological Methodology, 18, 309–345.

Weld, L.H. (1989). Significance Levels from Public-Use Data with Multiply-Imputed Industry Codes. Ph.D. thesis, Department of Statistics, Harvard University.