AVOIDING DISCLOSURE IN THE RELEASE OF MICRODATA

Paul Zeisset, U.S. Bureau of the Census

For the purpose of this discussion, we will use the term microdata to refer to files in which each record provides data about an individual person, household, establishment, or other unit. Microdata thus include an agency's own confidential files of questionnaires or basic records from a survey or other data collection. Normally we think of these data as being summarized or aggregated to produce statistics for the reports and publications discussed in the previous paper. Nonetheless, release of information in microdata form to a data user outside the originating agency can serve legitimate and important public purposes--in that the data may be useful for many more tabulations or other analyses than the originating agency is prepared to provide. Further, certain statistical applications (for example, similation models) require the user to have input in microdata form.

Release of records about individuals raises the issue of disclosure. Some files are by law not confidential, for example, from the Census of Governments, where detailed data are released identified to the specific governmental unit. On the other hand, most statistical data bases are covered by statutes which prohibit the release of data from which information may be gained about particular individual entities, be they persons, households, establishments, corporations, or other reporting units. In the latter situation, microdata are releasable only if the information is not specific enough to allow identification of the individual. Invariably names and addresses, social security numbers, and other positive identifiers must be removed. Further, certain other information, such as residential location, is generally abbreviated or withheld.

Federal Agency Examples of Microdata Release

For those of you not familiar with what types of microdata files are being released by Federal agencies, let me give you a few examples.

Probably the best known of all Federal microdata bases are the public use samples of basic records from the 1960 and 1970 censuses of population and housing. From the first release in 1963, these samples have provided nearly the full richness of detail about households derivable from the decennial censuses: age, education, income, occupation, etc., of each family member along with characteristics of the family's housing. The sample originally released in 1963 had little geographic information and the sampling fraction was only 0.1 percent of all American households. But, based on the public acceptance and demonstrated utility of that microdata product, public use samples from the 1970 census were created with a larger sampling fraction (one-percent) and more specific geographic information (that is, areas as small as 250,000 population were identified).

The Census Bureau also releases survey data files on a similar basis, with certain added qualifications regarding the smallest areas that can be identified. Microdata are available from the Current Population Survey, the Annual Housing Survey, and the National Travel Survey, to name just a few. Other agencies frequently contract with the Census Bureau to conduct surveys for them, and these surveys also result in microdata files released by either Census or the sponsoring agency: for example, the National Crime Survey sponsored by LEAA, the Consumer Expenditures Survey sponsored by BLS, and the Survey of Income and Education by HEW. In general, all of these files become available for unrestricted public use after identifiers, detailed geography, and some subject information are removed.

Several agencies also release microdata based on administrative records. The Social Security Administration makes several files available from its Continuous Work History Sample derived from payroll tax records and from records of each applicant for a social security number. The Longitudinal Employee-Employer Data (LEED) file is a one-percent sample of employees covered by Social Security. For every individual in the file there is age, race, sex and a record for each place of employment since 1957, indicating the industry, State, county, taxable wages, and estimated total wages for each year. In view of the disclosure potential of the county and industry identification, purchasers must enter into a written agreement with SSA specifying the purpose for which the file may be used, prohibiting further dissemination without SSA authorization, and specifically precluding any attempt to identify specific individuals or establishments or to match individual records with information in other files on specific individuals.

The National Center for Health Statistics also releases a number of microdata files. In this context the most interesting of these is the file on natality which provides a 50-percent sample of records in its birth registration system. No other federal microdata file allows so large a sampling fraction. Records include the age, race, and education of the father and mother, the State and county of residence of the mother, the birth date, legitimacy (if recorded) and several characteristics of the mother's previous childbearing history. Purchasers of a NCHS microdata file must sign a statement that the microdata file will be used solely for statistical research purposes.

Factors Bearing on the Likelihood of Disclosure

While we are confining our consideration to microdata files with no positive identifiers, it should be recognized that a combination of data elements, such as geographic location, age, race, and occupation, if sufficiently detailed, could be used to identify an individual if the investigator knew those characteristics of his subject in advance. Other information on a microdata

record so identified would then be disclosed about the individual, for instance, his income, marital history, educational attainment, and so forth.

Let me discuss three factors bearing on the likelihood that such disclosure might occur: (1) sample size, (2) geographic and subject detail--or the degree to which records in the file are unique, and (3) recognizability of the sample record.

(1) Sample size or fraction of the universe

If an investigator were searching for a particular individual in a microdata file his probability of success can be no greater than the chances that the individual's record is present in the file. In a one-percent sample the chances are 99 to 1 against a particular individual having a record in the file, assuming one has no external way of knowing that the individual was included in the sample. A larger sample size would create greater disclosure potential; a smaller sampling fraction would yield less.

(2) Uniqueness

I use the term uniqueness to refer to whether an individual can be distinguished from all other members in a population in terms of information available on the microdata record. That uniqueness is determined by the size of the population and the degree to which it is segmented by geographic information, and the number and detail of characteristics provided for the sample unit.

The smaller the population, the more easily an individual can be unique; the larger the population, the more likely that his or her set of characteristics is duplicated by somebody else's. Size of the population, or of the smallest segment that can be readily identified, can be varied quite directly by varying the amount of geographic information supplied on a microdata file.

It can also be said that the greater the number and detail of characteristics reported about an individual the more likely it is that the individual's representation on the file would be different from that of any other individual in the population. Just 10 characteristics with four categories each create over a million possibilities (4^{10}), and when one considers that some data items may have 100 or more potential categories (e.g., age, occupation, industry, income, place of birth) the number of possibilities becomes astronomical in a file with a large number of characteristics. Many characteristics are, however, likely to be correlated with one another, thus reducing the degree to which an additional item creates additional unique records.

Assuming that we need to control the degree of differentiation available, it might then seem reasonable to designate a minimum category population, for instance, to collapse country-of-birth categories with less than 50 cases in the file. The technique appears inadequate, however, since for instance, while there may be many Russian-born persons sampled, only one may be black, or only one may live in a particular identified area. More important, uniqueness in the sample is not the critical factor, for there may be a hundred such individuals in the population with no possibility of discriminating among them. Uniqueness in the population is the real question, and this can not be determined without a census or administrative file exhausting the population or at least an identifiable subset thereof (such as a file of all doctors). Precluding uniqueness in the sample would be a very conservative approach to avoiding disclosure.

Some public-use microdata files provide characteristics for all or at least multiple members of a household. The association of the characteristics of household members greatly increases the potential for unique combinations (for example, a 66-year-old judge married to a 23-year-old actress would be a rather unusual combination.)

(3) Recognizability

Suppose we determine that a given record is unique. The next question is whether that record can be linked to a specific person, without which disclosure does not occur. I will refer to this property as a record's recognizability, and I'll discuss three factors affecting it: (a) the existence of a population register, (b) inaccuracy or "noise" in the microdata file, and (c) time lag or the degree to which the microdata information becomes out-of-date for an individual.

(a) Population Registers

Suppose there were a list of everyone in the population, including each person's age, place of birth, and a few other items which were also on a public-use microdata file. Such a list, or population register, could make it not too difficult to find the identity of any one with a unique set of those characteristics.

In some countries, Sweden to name one, such registers are publicly available. In this country the best lists would be in the hands of the Internal Revenue and the Social Security Administration, but these are not available to the public. But neither nationwide coverage nor coverage of all segments of the population is required. Reasonable coverage of a defined subpopulation, along with a number of reliable matching characteristics may suffice. A register of some groups like black architects, American Indians, high public officials, or birth records is not improbable. The existence of rather extensive registers of business establishments, in the hands of government agencies, trade associations or firms like Dun and Bradstreet, has virtually ruled out the possibility of releasing microdata files about businesses for statistical purposes.

One needn't associate the idea of a population register with the dossiers of an investigative agency. If Who's Who in America or the Congressional Directory were in computerized form they could be quite useable for the restricted populations they cover. Welfare agencies and credit bureaus might have information useable for matching in computerized form although access to these files is assumed to be restricted. Those lists which are public--city directories, voter registration lists, or the records of motor vehicle agencies, tax assessors or real estate agencies--probably don't contain a broad enough set of characteristics for matching, at least with the microdata files we have examined. There should be no doubt, however, that any new file considered for availability in microdata form should be reviewed for its correspondence to various existing population registers.

(b) "Noise" in the Data

Another factor which affects recognizability is inaccuracy or "noise" (random error) in the microdata. Usually we think of noise in data as undesirable--respondent mistakes, intentional misrepresentation, coding or processing errors--but that noise also reduces disclosure potential in that unreliability in the microrecord degrades its matchability to a referent in the population. The effect is more severe to attempted identification through matching than it is to the more appropriate statistical uses because there is no chance for compensating errors to average out or to appear small in perspective.

If unintended error or unreliability helps reduce disclosure potential, then intentional noise added to a microdata file could be still more effective, particularly in touching all records rather than just some. Doing so without damaging the usefulness of the file for statistical purposes is the problem.

(c) Time Lag

Time lag is a third factor affecting recognizability. There is inevitably some lag between the date of data collection or reference date and the date the microdata become available, usually at least several months and sometimes several years. As the data become less current they become less useful for many statistical purposes, but they may also become less potentially dangerous to confidentiality.

First, the user will have greater difficulty in reconstructing a given individual's characteristics as of the reference date. Secondly, whatever possible gain the user might expect from the match will presumably be less. Welfare agencies and credit bureaus might have the best files for matching purposes, but the fact that the linked microdata may be one or more years out of date should reduce the utility of the match substantially. A microdata file could be withheld from public use for a number of months or years to reduce its disclosure potential, or "old" files could be released with less stringent protection than contemporary files.

Hypothesized Relationships Among the Various Factors

Now, in examining the relative impact on disclosure potential of the various factors we have discussed, it is useful to hypothesize how an investigator might go about identifying microdata records. There appear to be two different broad types of potential disclosure situations, and they are affected by the various factors in differing degrees. The first scenario is where the investigator searches the file for a specific individual, using certain characteristics of which he is already aware. The second is where the investigator is just "fishing" for a set of characteristics he recognizes.

The first type is quite volatile. If a public-use microdata file were to be useful for investigatory purposes, the breach of confidentiality would be extremely serious. The most obvious factor working against misuse of this type is the sample size. Even considering the largest of the existing public-use microdata files, the six 1970 census one-percent public use samples, and under hypothetically perfect matching conditions, the investigator would have a 94-percent probability of failure with regard to a particular individual.

Only where there is an extremely large number of subjects for whom excellent matching data are available, and under conditions where success in only a few cases will suffice, could the file seem to be of any use. The existence of some sort of population register would be almost a necessity for investigatory use. It is also true that any substantial noise or inaccuracy in the data would preclude an exact match rather effectively.

By contrast, in the second type of disclosure situation the investigator is not searching for a particular individual, but is just "fishing" for a set of characteristics he or she recognizes. Such an occurrence does not immediately seem to be very serious, since there is presumably no profitable purpose to be served by such an investigation. Such an effort might, however, be undertaken in an attempt to discredit the issuing agency or the practice of releasing microdata.

Since one is not starting with a specific set of target individuals, the low probability that any one individual is in the sample is not a problem to the investigator. The investigator selects certain unusual and highly noticeable characteristics, then extracts corresponding records from the sample. The task then is to recognize well-known households or individuals among the extracted records. A population register would be useful but not mandatory here. In the absence of a population register, geographic information on a file is very important since it may be the most specific matching characteristic known to the investigator. Number of characteristics reported is important since the matching will depend on some sort of pattern recognition. Minor aberrations introduced into the data may not inhibit the match if they do not disturb the general pattern, quite unlike the situation with a population register where a minor discrepancy might defeat the match. Compared to searching for a specific individual, the technical requirements for a fishing expedition are relatively modest.

Techniques for Avoiding Disclosure

(1) General Tradeoffs

From the foregoing it should be apparent that a number of factors impact on disclosure potential, and also that no one of them alone can be so restricted as to prevent disclosure by itself. A file which exhausts a universe, or comes close, presents considerable disclosure potential if it contains any unique records. Geographic information must be restricted beyond the point where an individual user could be familiar with a significant proportion of the universe, but whether that point comes at 25,000, 250,000 or 1 million will depend on the detail in the file and other restrictions imposed. The Census Bureau has imposed a 250,000 minimum population criterion across the board, but that is in the context that the Bureau normally provides data files with highly detailed subject matter (for instance, single years of age, detailed occupation). No formula has been worked out adequately representing the tradeoffs between level of geographic identification, detail of individual subject items, and sample size.

(2) Elimination of Categories Identifying Small Salient Groups

Another technique is to avoid categories so detailed that they define a small and easily identifiable group. Providing income groupings so that persons with very high incomes cannot be separately identified is a common technique and may be seen as a more generalized approach to insuring that corporate executives and other highly recognizable individuals not be so easily identified from the rest of the population. A common upper limit for detailed income categories is $50,000 per year, although inflation may soon make a somewhat higher cutoff appropriate.

(3) Allowing No Unique Cases

It has also been proposed (Fellegi, 1972) that microdata files can be made disclosure free by making sure that there are no unique records in the file, which is to say that every set of characteristics is replicated at least once. There is little doubt that this standard would prevent disclosure since any match attempt would never result in only a single qualifying individual. This is, however, an unrealistic standard for a file with many data items, since the number of possible combinations would be astronomically high when in fact relatively few of those data items would be involved in any conceivable match attempt.

That procedure does have some relevance when a particular population register is recognized as threatening the confidentiality of a microdata file, for example, a drivers license file with date of birth, state of birth, sex, and marital status. If a four-dimensional cross tabulation of the microdata within the area to be identified had any cells with only one case, categories could be collapsed or areas redefined until that no longer occurred. If more than one population register existed then the resulting microdata could be subjected to additional cross tabulation. This solution should be recognized as being conservative since it is uniqueness in the population, rather than in the microdata file, which assures matchability. Thus, if possible, the multi-dimensional search for the unique case should be performed on the population register file rather than in the statistical microdata.

(4) Noise Introduction

The introduction of noise into microdata is a fourth alternative. In its simplest form it might involve adding or subtracting small amounts at random to values of continuous or interval variables. There are multiplicative as well as additive models, and a few ideas have also come out of the recent literature on randomized response. Clayton and Poole (1976) did some interesting research on the impact of a couple of error introduction techniques on certain univariate applications. But as yet there is little knowledge of the degree to which error introduction would degrade the more common multivariate analyses. If noise were introduced into data on age, for example, the user's concern is not just that age distributions can be faithfully reproduced, but that the noise does not distort sensitive relationships, such as between age and educational progress where one is attempting to

study the cohorts of students ahead of or behind "normal" progress defined by specific age-grade relationships.

(5) Removal of Well Known Individuals from the File

Finally, if disclosure potential lies primarily with a few people with unusual characteristics it is at least worth considering removing them from the file, rather than eliminating some of the information about all of the population. If more than a handful of such individuals is involved there must be concern about bias resulting from their removal. Of course, the originating agency could prepare summary statistics about the individuals removed. But such a procedure should not be relied on to the exclusion of other techniques since the existence of a large population register would make many people recognizable in a detailed file.

Disclosure Prevention Through Restrictions on Use

In the foregoing I have tried to identify ways in which a file may be made acceptable for unrestricted use. Invariable each bit of information removed from a file to make it disclosure-free reduces that file's usefulness for some research purpose. In fact, we at the Census Bureau are continually met with requests to relax our geographic restrictions on microdata to make this or that worthwhile research possible.

Life certainly would be simpler if we could just trust the data user not to misuse the file. Or, if not naive trust, surely strict contractual arrangements could bind the user of a restricted file to observe procedures which would maintain the confidentiality of the individual data.

Our subcommittee carefully considered what conditions could provide adequate protection, in terms of legal authority needed by the user, penalties for misuse, and a set of conditions agreed to by the receiving organization. The Social Security Administration is now releasing certain files on such a restricted basis--not files with individual identification, but files with too much disclosure potential for unrestricted dissemination.

Certain other agencies are not so ready to embrace the idea of restricted dissemination. The statutes of some agencies don't give them the flexibility SSA has. Furthermore, laws such as the Freedom of Information Act make it not altogether certain that regulations could be upheld if they allow one user access to a file but prohibit access to another.

In 1963 the Census Bureau placed certain restrictions on purchasers of its new 1-in-1000 sample. It wanted to keep careful records on the use of the file--for administrative rather than confidentiality reasons. Unfortunately, those signed agreements were soon forgotten by the purchasers, and the files in question passed freely from one to another. This experience certainly indicated to us that an agency could not successfully restrict use without specific attention to enforcement.

The most important reason, of course, for not relying on restricted-use agreements to enforce confidentiality is that there is a great deal to be gained, by the research community and by society at large, by broad and free access to microdata files such as we have discussed. Restricted release should be considered only where a file's disclosure potential cannot be reduced to an acceptable level while maintaining the usefulness of the file, and then, of course, only where the law allows and the restrictions can be successfully enforced.

Unfortunately, our subcommittee did not come up with a neat formula or simple package of rules to follow to produce microdata of optimum usefulness and confidentiality. Research--of both a theoretical and empirical nature--is needed. Our subcommittee report, then, is of greatest value when used as a study guide by responsible agency officials, simultaneously mindful of the importance of confidentiality and the societal benefits of broad access to public data.

References

Clayton, C. A. and Poole, W. K.

1976 Use of Randomized Response Techniques in Maintaining Confidentiality of Data. Draft Report. Research Triangle Institute.

Fellegi, I. P.

1972 "On the question of statistical confidentiality." Journal of the American Statistical Association, Vol. 67: 7-18.