VIEWS ON THE PRODUCTION AND USE OF CONFIDENTIAL MICRODATA

William E. Winkler
Bureau of the Census

ABSTRACT

To be of use to researchers, public-use microdata should be analytically valid and interesting. Public-use microdata are analytically valid if they yield results and conclusions that correspond closely to the results from the original, confidential microdata. Microdata are analytically interesting if files contain a sufficient number of variables, say five demographic and six quantitative, to produce more detailed inferences than could be produced using a large number of summary statistics. Many of the existing public-use files have been created by data providers who are unaware of modern record linkage techniques that allow some of their records to be reidentified and associated with individuals. Existing record linkage methods are so powerful that a small percentage of reidentifications is possible in some public-use files by relatively naive persons using commercially available software. Newly developing record linkage methods will initially allow the association of sets of individual records from sets of files for many important economic and demographic analyses that serve the public good and significantly reduce costs. These powerful new record linkage methods, when fully realized, will allow reidentifications in many of the public-use data files even though the data were produced by conscientious individuals who believed they were using effective techniques for assuring confidentiality.

INTRODUCTION

A public-use microdata file should be analytically valid. By this, I mean that, for a very small number of uses, the microdata should produce analytic results that are approximately the same as the original, confidential file that is not distributed. If the public-use microdata do not produce analytic results that are approximately the same, then a user could publish results and make policy decisions that are in error. In such a situation, it is likely that the statistical agency that produced the public-use data would need to show that published results are in error. If the public-use microdata is to meet a moderate number of specific analytic needs (say six), then it is quite possible that the data are no longer confidential. The lack of confidentiality is due to the fact that the microdata will have to meet a number of marginal restraints, particularly in subdomains defined using six different sets of variables.

In producing confidential, public-use data files, statistical agencies should first assure that the files are analytically valid. If the agencies cannot produce analytically valid files (whether or not they are confidential), then the agencies should not give out public-use files. When agencies first produce the public-use microdata, they should check it for analytic validity. Some methods of producing the public-use data may have more of a tendency to yield analytically valid files and make checking for analytic validity easier. Additive noise (Kim 1986, 1990) and the generalizations (Sullivan and Fuller 1990, 1991; Fuller 1993, Little 1993) have a strong tendency to produce analytically valid files. The files that are produced via noise addition have generally not been confidential because a (possibly) small percentage of records can be reidentified with a moderately high probability.

Both coarsening methods (of key variables via global recoding and local suppressions as in Statistics Netherlands' ARGUS) and data swapping (which has been used by the Census Bureau and others) are known to reduce the analytic validity of files (Little 1993). Coarsening is likely to be superior to swapping in many situations and can still make reidentification more difficult (e.g., De Waal and Willenborg 1996, De Waal, Hundepool, & Willenborg 1996).

By adding additional variables to records and increasing the number of records in files, data providers both make the files more useful and at greater risk of reidentification. The main reason for adding more information to the public-use file is to allow several analytically valid uses of the data. Assuring that a public-use file satisfies several analytic constraints on margins, correlations, and subdomains requires a moderate amount of modeling and software development by highly skilled individuals and is not a cost-effective use of the data provider's resources. Whether it is possible to write computationally tractable software and do the modeling to produce artificial data, meeting a number of analytical needs on subdomains with a moderate number of variables (say, twelve), is an open research problem.

REIDENTIFICATION

Whereas the various methods of making reidentification more difficult (additive noise, coarsening, swapping) and the checking of processed files for analytic validity are straightforward, reidentification experiments are much more difficult. In seminal work, Paas (1988) showed that reidentification can be straightforward when there are a large number of multivariate normal variables. Blien, Wirth, and Müller (among others) have criticized Paas' assertion that confidentiality will be difficult to achieve in many situations where there are many variables for matching. Their criticism is based on the fact the individual (intruder) performing reidentification is unlikely to have specific knowledge of whether records in the intruder's file are also present in the public-use file and based on the use of crude methods of reidentification such as those used by Paas or even simpler methods.

I assert that reidentification with public-use files is far easier than some individuals believe who are not familiar with powerful record linkage methods. My assertion is based on three sets of results that I will describe below. The ideas will be illustrated by explaining concepts in terms of the large public-use file produced by Kim and Winkler (1995).

First, if Paas had been able to use modern record linkage methods, he would have been able to reidentify records using fewer variables and the accuracy would likely have been much greater. My statement is based on the following. In unpublished work (Winkler 1984), I showed that record linkage methods based on the model of Fellegi and Sunter (1969) worked far better than those based on multivariate normal models such as used by Paas. The results were so extreme that in many common situations I was able to identify as much as 40% matches from one file to the next using the Fellegi-Sunter methods and less than 5% with the multivariate normal methods. The accuracy with the Fellegi-Sunter methods was also much greater. I had earlier investigated the crude use of combinations of key variables that performed even more poorly. At the time of the research (1984), I used highly developed discriminant analysis procedures for the multivariate normal data and used the most elementary Fellegi-Sunter ideas in crude code that I wrote. I did not use the advanced string comparators, new multivariate comparison methods, LP algorithms for optimizing over large sets of pairs, EM-Algorithm-based methods for automatically finding optimal parameters, and the powerful parsing and standardization routines that have been developed for record linkage over the last 10 years.

Second, if the files have many variables and must meet a variety of detailed analytic needs, then the data provider should develop methods of producing public-use micro data to assure specific analytic and confidentiality needs are met. Assuring that analytic needs are met can greatly increase the risk of reidentification if special precautions are not taken. The method of Kim and Winkler (1995) produces analytically valid files that are confidential using current record linkage technology and have a combination of key variables (such as geographic area, age, race, and sex) and quantitative variables. Previously existing methods for coarsening and swapping would not have produced analytically valid files. Typical additive-noise methods would not have produced confidential files.

Third, it seems likely the new, non-trivial extensions of current record linkage methods will allow applications with administrative lists and reidentification experiments that were previously thought impossible. These extensions, in particular, will allow linking in of additional data sets and the creation of additional sets of matching variables to improve the matching of two or more files. Because there is significant interest in using microdata from sets of administrative lists, Scheuren and Winkler (1996) have developed methods for merging microdata files and performing statistically valid analyses when the only common information in files is name and address information. Before showing how Scheuren-Winkler ideas extend to allow reidentification in the public-use files, we need a specific example of a public-use file.

PRODUCING AN ANALYTICALLY VALID, CONFIDENTIAL FILE

I illustrate by describing the example of Kim and Winkler (1995) because it gives insight into reidentification issues and how difficult it is to produce analytically valid files when there are many variables that must be distorted.

The Census Bureau needed to produce a public-use file that contained five Internal Revenue Service (IRS) income variables, State, age, sex, race, and several Current Population Survey (CPS) variables that could be used in evaluating low-income and other tax-related policies. The confidentiality restrictions were that no IRS data could alone be used in reidentification, that other data could not be used in obtaining confidential IRS data, and that combinations of the data could not be used in reidentification. In particular, the IRS, which had access to all the IRS income data, age, sex, State, and several other variables, could not obtain specific individual's CPS data and other information that was on the files.

Because the non-public-use file of 59315 records was a 1/1600 sample of the population, many individuals assumed that only minor distortions to the data would be needed. Jay Kim and I first increased additive-noise levels until we could no longer reidentify many records. At the high noise levels, the files were no longer analytically valid. They often had correlations at aggregate levels that varied in the second decimal place, and in a few situations, in the first decimal place. By decreasing noise levels, we were generally able to assure that correlations agreed to three decimals places in important subdomains. In the files with the quantitative data distorted at decreased noise levels, we could reidentify 0.8% of the records using IRS form type, State, age, and the income variables. The 0.8% often consisted of records that had a set of characteristics that made them sufficiently different from other records that they could be reidentified in the original population, not just in the public-use sample. With just State and the five distorted IRS income variables, we could reidentify a moderate number of records. This is a very important point because it applies to matching with economic entities such as businesses and to matching with individuals. Combinations of quantitative data can serve to uniquely identify an individual or a company whether or not name and address information is present. Other combinations of information can also be used in reidentification. Bethlehem, Keller, and Pannekoek (1989) had earlier observed how groups of variables could be used for reidentification when more rudimentary record linkage methods are used.

To reduce drastically the reidentification risk, we developed a swapping method (using modified record linkage software) that swaps large portions of records and preserves covariances on a large set of specified subdomains. We applied the swapping-based methods to the small subportion of records that could be easily reidentified. Both the additive noise and swapping methods preserve means on the specified subdomains. The key to developing a method for producing public-use files was having very good methods of reidentification based on record linkage methods available in commercial or publicly available software. In concluding our paper, Kim and I stated that the reidentification risk for all records in the public-use file is below 0.1%. Later, we decided the risk was below 0.01%, an estimate many individuals would agree with. Recent work, however, indicates that newer, more sophisticated modeling and reidentification methods will be able to reidentify records in the public-use file.

CURRENT EXTENSIONS OF RECORD LINKAGE

Scheuren and I assumed that a small number of records can be accurately matched from two files using name and address information only, and that the first file contained a normal variable that was moderately correlated with a normal variable from the other file. The two quantitative variables (one from each file) are sufficiently different that they can not directly be used in matching. The situation is analogous to the situation where one agency might have a list of companies and the inputs of raw materials that they use in production and another agency would have a list of companies and the types, quantities, and values of the goods they produce. An economist would be interested in modeling using merged microdata but would not effectively be able to merge the files with the name and address information. Other files or a priori information could also be used to model the relationships between the two variables.

For the specific empirical example in Scheuren and Winkler, we considered pairs of quantitative variables from the well-matched subset and regressed one variable on the other to get predictors. Using the model from the subset, we added predictors to all records in one file and used the additional variable to improve matching to the point where we could perform reasonably accurate regression analyses. In extreme situations, R-square values rose from noise levels (below 0.1) when only name and address information was used to values accurately representing truth (ranging between 0.45 and 0.7) when all information (including the new predicted variables) was used.

FUTURE EXTENSIONS

In demonstrating that the Kim-Winkler file is not confidential, university researchers might obtain record linkage software, build a large National file using credit files and other available information, and create a model for the relationships between variables in their file and in the public-use file. If the model-predicted variables are reasonable and there are a moderate number of them, then the university researchers might accurately be able to put names from their file with the correct corresponding records in the public-use file. In Scheuren-Winkler, the small subsample was used to model the relationships between the two quantitative variables. With large independent data sets, the university researchers could develop much more accurate models with more data that would improve linkages much more dramatically than in the Scheuren-Winkler example.

CONCLUSIONS

I expect that the first large scale applications of the new linkage methods will be by economists who wish to wish to combine sets of large files and perform analyses that have not previously been feasible. Quantitative data and other characteristics associated with companies can serve to make linkages reasonably straightforward. Eventually, researchers will also develop projects for linking sets of files of individuals. These methods are particularly appealing in countries such as Canada, the U.S., the U.K., France, and Australia that have not been able to merge most of their economic and population files because unique identifiers are not present. Once the sets of methodological and software tools have been developed for legitimate policy questions that serve the public good, the tools may be used in reidentification experiments.

BIBLIOGRAPHY

Bethlehem, J. A., Keller, W. J., and Pannekoek, J., (1989) "Disclosure Control of Microdata," *Journal of the American Statistical Association,* **85**, 38-45.

Blien, U., Wirth, U., and Muller, M. (1992), “Disclosure Risk for Microdata Stemming from Official Statistics,” *Statistica Neerlandica*, **46**, 69-82.

De Waal, A. G., and Willenborg, L.C.R.J. (1996), "A View of Statistical Disclosure Control for Microdata," *Survey Methodology*, **22**, 95-103.

De Waal, A. G., and Willenborg, L.C.R.J. (1995), "Global Recodings and Local Suppressions in Microdata Sets," Proceedings of Statistics Canada 95, 121-132.

Fuller, W. A. (1993), “Masking Procedures for Microdata Disclosure Limitation,” *Journal of Official Statistics*, **9,** 383-406.

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

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

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

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

Kim, J. J., and Winkler, W. E. (1995), “Masking Microdata Files,”American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 114-119.

Paas, G. (1988), "Disclosure Risk and Disclosure Avoidance for Microdata," *Journal of Business and Economic Statistics*, **6**, 487-500.

Scheuren, F., and Winkler, W.E. (1996), “Recursive Merging and Analysis of Administration Lists,” American Statistical Association, Proceedings of the Section on Survey Research

Methods, 123-128 (presently available on http://www.amstat.org in section on govt statistics).

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

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

Winkler, W. E., (1984), "Record Linkage," Energy Information Administration technical report.