Should We Continue To Release Public-Use Microdata Files? Yes, Yes, Yes!

Stephen E. Fienberg, Carnegie Mellon University

Public-use microdata are fundamental to much social science research and public policy. The need to maintain the confidentiality of respondents in such data bases has raised the issue of whether we can continue to operate using the current approaches for the release of microdata. We argue that statistical agencies have a responsibility to provide access to such data bases and that advances in statistical methods for disclosure limitation should aid them in this effort.

■ Introduction

Martin David's far-ranging introduction to this topic offers much food for thought and, as always, I find myself in close agreement with him on many aspects of the issues he discusses. I take issue with some of his assertions and assumptions and thus come down in quite a different place on the issues of data release and access. In particular, I remain an unabashed advocate of unrestricted access to microdata files and believe that facilitating access to such data in simple forms via the World Wide Web represents the future of public data access (see Fienberg [1], [2]).

Finally, I believe that recent advances in disclosure limitation offer tools to combat the attacks of intruders and others who wish to compromise the integrity of the public statistical enterprise for private gain. For a good overview of some of these advances, see the special 1998 issue of *The Journal of Official Statistics* on this topic, as well as an earlier issue from about five years ago in the same journal, and the forthcoming proceedings of the conference on *Statistical Disclosure Protection* held in Lisbon, in March 1998. I will make reference here to some of these papers.

Points of Agreement

Let us now turn now to some key points of agreement between David and myself, which underpin both of our positions.

□ As David argues, heightened public concern about supplying data for both administrative and statistical purposes has once more raised concern over unauthorized access to identifiable information. The recent flap over the possible introduction of a new medical insurance number that would allow the linkage of all medical information on an individual is only one such example.

□ The public has little understanding of the statistical uses of administrative data or of the firewalls and other forms of protection used by statistical agencies. The dangers of large commercial data bases amassed for marketing and credit purposes are substantial, and we as a profession need to step forward and make clear the distinction between those private enterprises and statistical ones carried out under Government auspices.

□ Eliminating or further limiting access to Government statistical data, through more restricted release, will have little impact on public perceptions of the safety of data collected under Federal statistical mandates, and will only serve to undercut the value of the data collected.

The onus of access to microdata cannot be placed solely on the shoulders of statistical agencies. Restricted access mechanisms are not foolproof, and disclosure limitation techniques only reduce but do not eliminate, disclosure risk. Thus, we must have legal mechanisms that would severely penalize those who attempt to use statistical data to identify individuals and thereby or otherwise gain access to confidential information.

In fact, I have gone, and would continue to go, much further than David when it comes to the issue of accessing statistical data. There are two different philosophies that people adopt with regard to the preservation of confidentiality associated with individual-level data: (1) *restricted* or *limited information*, wherein the amount or format of the data released is subject to restrictions, and

(2) *restricted* or *limited access*, wherein the access to the information is itself restricted. I have argued elsewhere (e.g., see Fienberg [2]) that Federal statistical data are a public good and that the Federal statistical agencies have a *responsibility* to provide wide and unrestricted access to data that might be of value to secondary users outside the agencies themselves. In the second section, I present a version of this argument.

The Census Bureau has plans for such a major data dissemination system associated with the 2000 Census (see Steel and Zayatz [3]) and, if they come to pass, the public at large may well learn the value of statistical data and the integrity of the Federal statistical enterprise when it comes to the preservation of confidentiality.

Points of Disagreement

There are, of course, some specialized statistical data bases that pose difficult problems for disclosure limitation (e.g., those involving enterprises), and it is here that proposals for restricted access have recently come to the fore. David [4] has spoken positively about mechanisms of restricted access. I believe that such approaches represent a copout on the part of statistical agencies and a government unwilling to invest in the research necessary to provide broader unrestricted access. Restricted access through centers such as those David describes is less than satisfactory to the thoughtful and innovative statistical analyst. Just imagine the agency response to a statistician's request for full residual plots at low levels of geography. If every such request needs to be screened before approval, which is essentially the current model, research progress will grind to a halt. Thus, restricted access as a strategy can only be useful as a short-term measure while we continue to struggle to understand how to successfully limit disclosure while at the same time provide access in an essentially unrestricted form to a perturbed form of the original statistical data base. Here, I differ markedly with David. We could devote a full session to the problems associated with the mechanisms for restricted access which David advocates. But restricting access can be separated from putting some burden, legal or otherwise, on the users of statistical data. Such a burden represents a cost that is balanced by the benefit of access. Thus, I believe that David's proposal is worth pursuing.

Concern about the vulnerability of statistical data to the attack of an intruder has led many statisticians and non-statisticians to overstate the dangers of broad-based microdata release. Despite the growth in the sophistication of software and the exponential growth of memory and storage, algorithms for matching do not work as accurately as people claim unless the intruder has more considerable knowledge to bring to bear than is almost always the case.

Outline

The remainder of the paper is as follows. The second section discusses public-use data as a public good, and then, in the third section, I describe briefly one broad strategy of providing expanded public-use data, the use of perturbations, and even simulation. In the fourth section, I briefly describe a project which a number of us hope will advance the general enterprise of expanded access.

There is a simple bottom line to my response to the question posed to the members of our panel at this session. "Should we continue to release public-use microdata?" My answer is "Yes, yes, yes! And in new and expanded forms."

■ Statistical Data as a Public Good

As I have already mentioned, my position with regard to Government statistical data is that they are a *public good* and that restricted access should only be justified in extreme situations where the confidentiality of data in the possession of an agency cannot be protected through some form of restriction on the information released.

Government statistical data such as those gathered as part of censuses and major sample surveys meet two key tests that are usually applied to quantities labeled as public goods: jointness of consumption (consumption by one person does not diminish their availability to others), and benefit to society as a whole (statistical data are used to inform public policy and as the basis for democratic representation). The only issue, then, is whether or not there is nonexclusivity, i.e., whether or not it makes sense to provide these statistical data to some citizens and not to others. But if we have means for providing access to all or virtually all in society, e.g., via the Internet and the World Wide Web, then the costs of providing the data to all are often less than the costs of restricting access, although there are other costs that result from expanded use to those who produce the data.

For a general discussion of the costs and benefits of data sharing, see Fienberg, Martin, and Straf [5], and for more focussed discussion relevant to the present context, see Duncan, Jabine, and de Wolf [6], pp. 29-33, and Duncan [7].

Thus, for me, the question is not if we should continue to supply public-use microdata, but how.

■ Perturbation and Methods and "Simulated" Public-Use Data

There is a general class of methods for disclosure limitation that were labeled *matrix masking* by Duncan and Pearson [8]. Some matrix masking methods alter the data in systematic ways, e.g., through aggregation or through cell suppression, and others do it through random perturbations, often subject to constraints for aggregates. Examples of perturbation methods are *controlled random rounding, data swapping,* and the recently proposed *post-randomization method* or PRAM of Gouweleeuw, et al. [9] and generalized by Duncan and Fienberg [10]. One way to think about random perturbation methods is as a restricted simulation tool, and, thus, we can link them to other types of simulation approaches that have recently been proposed.

Fienberg, Makov, and Steele [11] pursue this simulation strategy and present a general approach to "simulating" from a constrained version of the cumulative empirical distribution function of the data. In the case when all the variables are categorical, the cumulative distribution function is essentially the same as the counts in the resulting cross-classification or contingency table. Thus, this general simulation approach is equivalent to simulating from a constrained contingency table, e.g., given a specific set of marginal totals. Feinberg, Makov, and Steele [11] thus suggest replacing the original data by a randomly generated one drawn from the "exact" distribution of the contingency table under a log-linear model that includes "confidentiality-preserving" margins among its minimal sufficient statistics. They then propose to retain the simulated table only if it is consistent with some more complex log-linear model. This approach offers the prospect of simultaneously smoothing the original counts *and* providing disclosure limitation protection.

There are also fascinating statistical links here to separate literatures on the existence of maximum likelihood estimates for log-linear models, techniques for generating exact condition distributions using the method of Grobner bases (see Diaconis and Sturmfels [12]), and bounds for contingency tables given a set of marginals as in Buzzigoli and Giusti [13], Fienberg [14], and Roehrig et al. [15].

An extremely important feature of the simulation methodology used here is that information on the variability which it introduces into the data is directly accessible to the user, since anyone can begin with the reported table and information about the margins that are held fixed, and then run the Diaconis-Sturmfels-Markov chain algorithm to regenerate the full distribution of all possible tables with those margins. This, then, allows the user to make inference about the added variability in a formal modeling context in a form similar to the approach to inference on Gouweleeuw et al. [9]. As a consequence, simulation and perturbation methods represent a major improvement from the perspective of access to data over cell suppression and data swapping.

There remain many practical issues regarding the use and efficacy of such methods for generating disclosure-limited public-use samples. For example, how can they be used when the full cross-classification of interest is very sparse, consisting largely of 0's and 1's? And how can we use models to generate the simulated data when users have a multiplicity of models and even classes of models which they would like to apply to the released data?

■ A Pilot Query System for Public Data Access

The National Institute of Statistical Science (NISS) has recently assembled a team of statistical researchers from multiple universities who plan to work with statisticians in statistical agencies to develop a Web-based query system that allows the use of disclosure limitation methods applied sequentially in response to a series of statistical queries, in which the public knowledge of releases is cumulative.

The query system idea draws in part on a pilot project described in Keller-McNulty and Unger [16], and it will use as tools the various disclosure limitation methods being developed in the literature. The idea is to fully automate the methods through algorithms and explore intruder behavior (c.f., Fienberg, Makov, and Sanil [17]).

To get a sense of how this system *might* use the ideas on simulated data bases, consider a data base consisting of a k-dimensional contingency table, for which the queries are only allowed to come in the form of requests for marginal tables, of dimension £ $k - 1$. What we know from statistical theory is that, as margins are released and cumulated by a user, we have increasing information about the table entries.

In response to a new query, the system now examines it in combination with all those previously released margins and decides if the bounds on the cells of the cross-classification are too tight. Then, it might offer one of three responses: (1) yes-release; (2) no-don't release; or perhaps (3) simulate a new table which is consistent with the previously released margins, and then release the requested margin table from it.

A sequential query system need not be restricted to categorical variables nor to queries that come in the form of requests for tables. Further, my favorite simulation techniques for disclosure limitation and my work on bounds for tables will represent only one of many alternative disclosure limitation strategies that need to be explored. Others plan to explore the Argus approach developed by the statisticians at Statistics Netherlands, for example (see [18], [19], and [20]).

The NISS plans are to develop a basic system, test it with one or more public-use microdata files, test intruder behavior in a variety of ways, and elicit agency and other reactions. While there are many theoretical and empirical issues to explore and many exciting questions to address, making such a system function with actual agency data bases offers the real future prospect of improved disclosure limitation *and* increased data access.

■ Summary

In response to the question to our panel, I have attempted to articulate the following position:

□ Statistical data are a public good, and this requires that we develop mechanisms for allowing maximal access to them. The simplest, fairest, and most direct way to do this is through the release of public-access microdata files, available electronically via the World Wide Web.

□ Statistical agencies need to utilize new strategies for disclosure limitation, such as those drawing on extensive forms of perturbation and simulation.

WWW-based query systems offer an exciting new prospect for achieving access to government statistical microdata using modern disclosure limitation controls.

For some, the new world of computation and instantaneous access to data around the world is a curse or at least a threat. For others, it is an opportunity to deliver on the promises of expanding access for research and public purposes.

■ References

[1] Fienberg, S.E. (1997), Confidentiality and disclosure limitation methodology: Challenges for national statistics and statistical research, background paper prepared for the Committee on National Statistics.

[2] Fienberg, S.E. (1997), A glimpse at the future of social science statistical data: New forms of data analysis, new types of access, and new issues for data providers, *IASSIST Quarterly,* **21**, (No. 2), pp. 8-11.

[3] Steel, P. and Zayatz, L. (1998), Disclosure limi-

tation for the 2000 census of housing and population, in *Statistical Data Protection (SDP '98) Proceedings,* IOS Press, to appear.

[4] David, M.H. (1998), Killing with Kindness: The Attack on Public-Use Data, presented at this session.

[5] Fienberg, S.E.; Martin, M.E.; and Straf, M.L. (1985), *Sharing Research Data,* Committee on National Statistics, National Academy Press, Washington, DC.

[6] Duncan, G.T.; Jabine, T.B.; and de Wolf, V.A. (eds.) (1993), *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics,* Panel on Confidentiality and Data Access, Committee on National Statistics, National Academy Press, Washington, DC.

[7] Duncan, G.T. (1995), Restricted data versus restricted access: A perspective from *Private Lives and Public Policies,* in *Seminar on New Directions in Statistical Methodology,* Statistical Policy Working Paper No. 23, Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC, Part 1.

[8] Duncan, G.T. and Pearson, R.B. (1991), Enhancing access to microdata while protecting confidentiality: prospects for the future (with discussion). *Statistical Science,* **6**, pp. 219-239.

[9] Gouweleeuw, J.M.; Kooiman, P.; Willenborg, L.C.R.J.; and deWolf, P.-P. (1998), Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics,* **14**, pp. 463-478.

[10] Duncan, G.T. and Fienberg, S.E. (1998), Obtaining information while preserving privacy: A Markov perturbation method for tabular data, in *Statistical Data Protection (SDP '98) Proceedings,* IOS Press, to appear.

[11] Fienberg, S.E.; Makov, U. E.; and Steele, R.J. (1998), Disclosure limitation using perturbation and related methods for categorical data, *Journal of Official Statistics,* **14**, pp. 485-502.

[12] Diaconis, P. and Sturmfels, B. (1998), Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics,* **26**, pp. 363-397.

[13] Buzzigoli, L. and Giusti, A. (1998), An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals, in *Statistical Data Protection (SDP '98) Proceedings,* IOS Press, to appear.

[14] Fienberg, S.E. (1998), Fréchet and Bonferroni bounds for multiway tables of counts with applications to disclosure limitation, in *Statistical Data Protection (SDP '98) Proceedings,* IOS Press, to appear.

[15] Roehrig, S.F.; Padman, S.; Duncan, G.; and Krishnan, R. (1998), Disclosure detection in multiple linked categorical datafiles: A unified network approach, in *Statistical Data Protection (SPD '98) Proceedings,* IOS Press, to appear.

[16] Keller-McNulty, S. and Unger, E.A. (1998), A data system prototype for remote access to information based on confidential data, *Journal of Official Statistics,* **14**, pp. 347-360.

[17] Fienberg, S.E.; Makov, U.E.; and Sanil, A.P. (1997), A Bayesian approach to data disclosure, Optimal intruder behavior for continuous data, *Journal of Official Statistics,* **13**, pp. 75-90.

[18] Hundepool, A.; Willenborg, L.; Wessels, A.; Van Gemerden, L.; Tiourine, S.; and Hurkens, C. (1998), m-*ARGUS user's manual,* Department of Statistical Methods, Statistics Netherlands.

[19] Hundepool, A.; Willenborg, L.; Van Gemerden, L.; Wessels, A.; Fischetti, M.; Salazar, J.-J.; and Caprara, A. (1998), t-*ARGUS user's manual,* Department of Statistical Methods, Statistics Netherlands.

[20] Willenborg, L. and De Waal, T. (1996), *Statistical Disclosure Control in Practice,* Lecture Notes in Statistics, Vol. 111, New York: Springer Verlag.