DATA CONFIDENTIALITY: A RESEARCHER'S PERSPECTIVE

James P. Smith, RAND
1700 Main Street, Santa Monica, CA 90407

Key Words: Data confidentiality

My assignment in this session is to discuss data confidentiality from the perspective of researchers--the potential users of the data. In recent years, data confidentiality has become a critical issue in determining whether some surveys are even able to get off the ground. The most important social science data surveys have vastly more information on respondents than was true even a decade earlier. Not only are far more questions asked in more detail in each round, surveys such as the Panel Survey of Income Dynamics (PSID) and the National Longitudinal Surveys (NLS) are longitudinal in design. Knowing a respondent's state of residence in one year is unlikely to offer much of a clue to his identity. Knowing the state where he lived for each of the last 25 years is quite another thing. In addition, many of these surveys and some new ones on the horizon are now requesting linkages with administrative records such as Medicare and social security earnings records. This explosion in information available on respondents has sharply elevated concerns on data confidentiality.

Advocates on both sides of the issue must concede that neither has a monopoly on virtue. Insuring the confidentiality of respondents is extremely important both on ethical and practical grounds. If respondents believe that the information they provide will be misused, increasing numbers will simply refuse to participate in scientific surveys. On the other side, current scientific and public research increasingly require very precise and detailed information. This essential research would be impossible if we took confidentiality to the limit. The only way of guaranteeing zero probability of disclosure is to not have surveys at all. We are talking then about a tradeoff between two critical values--confidentiality and research. There are no absolutes in this debate--the practical questions are where to draw the line and where the real dangers lie. A good deal more progress will be made on this thorny problem when we eliminate corner solutions and reduce the posturing.

I will organize my paper around the four most frequently proposed solutions to data confidentiality--remote access, Census Special Sworn Employees (SSE), statistical models, and licensing of researchers. In my evaluation, I am asking the question of whether each of these models is viable on a large scale--that is, making survey data available in a useful way to significant numbers in the research community. For special circumstances in limited applications, each model clearly will play a role. For wide wholesale applications, the Census Bureau recently has been pushing the virtues of the first three models while discounting the licensing approach. I will argue here that this ranking has it backwards. The first three models--remote access, SSE, and statistical models--are seriously flawed and should not be promoted as viable solutions to data confidentiality problems. The best long-term solution is to license researchers and research institutions and to simultaneously impose large penalties and strong sanctions for misuse.

REMOTE ACCESS

A frequently-recommended alternative is to offer researchers remote access to confidential data over BITNET or some other form of electronic file transfer. The model here, and often cited as a successful one, is the Luxembourg Income Study (LIS). To oversimplify, LIS is a collection of *Current Population Survey*-styled surveys from the United States and largely, but not exclusively, Western European countries. In order to gain access to the data, a researcher submits a SPSS job by electronic file transfer, which will be run locally by LIS staff after they check the computer programs for potential violations of confidentiality.

Despite the growing list of fans for this model, I do not believe that it is a viable solution for mainstream social science data surveys and analyses. Turn-around time is not the problem, for it is widely reported usually to be a matter of hours and not days or weeks. One reason that it is not a useful model is that LIS data sets are relatively simple, lending themselves to mostly cross-tabular or simple regression analysis. In contrast, many of the new social science surveys that raise confidentiality questions are far more complex, with longitudinal designs and multiple

Reprinted from the 1991 Proceedings of the Social Statistics Section of the American Statistical Association

linkages with detailed administrative data.

Mostly, the LIS model fails as a prototype because it rests on some very naive assumptions about how the research process actually works in practice. This model asserts that research largely involves using well established statistical models on a ready-to-run analytical data file. This description falls well off the mark. For example, most (90 percent or more) of my analysis of micro-data takes place before I submit an SPSS or SAS program. It involves elaborate and painstaking data searching and cleaning. This interactive dialogue with the data not only produces the operational definitions of variables, it also shapes the very research questions that arise.

In addition, a good deal of modern statistical analysis, especially that on the frontier of the sciences, is custom written; for example, involving maximum-likelihood procedures. Computer programs can and often do run into pages and pages. I shudder to think of the poor Census Bureau employee whose job it will be to read through all these programs. If done with any seriousness, turn-around time would certainly be measured in months if not years. Those of you who write your own programs know that it is a myth to think you really could read through another's code and divine whether the intent of each line of code is benign or malign. This is especially true if we try to replicate the LIS model on more complicated data files such as SIPP, PSID, or the new Health and Retirement Survey.

SPECIAL SWORN EMPLOYEES

A related approach that has also been frequently put forth is that researchers become special sworn employees of the Census Bureau (SSE) under a joint statistical agreement or a fellowship. Since they are technically Census Bureau employees, they have the same rights of access to confidential data as regular Census employees. Under this model, the researcher, and his programmer if necessary, comes on site to the Census Bureau headquarters in Suiteland, Maryland for a periods of a few months to over a year. During his or her stay, the researcher can use the confidential data. The drawback, of course, is that the original data or any derivative analytical files cannot be taken when the researcher leaves.

I believe the problems with this model are so overwhelming that it is simply not feasible on any reasonable scale. The most obvious difficulty is that very few researchers have the time available to spend from 6 months to a year with their programmer off-site. Six months is actually a very short time to complete a complex research project from start to finish. The upshot is that there are only a very limited number of researchers capable of fitting in with this approach. Since scientists of the first rank are among the busiest, this model will exclude many of the best analysts. But even if the problem of the supply of researchers did not exist, the Census Bureau doesn't have a place to house them. The Census already has difficulty finding room for the small number of visiting scholars who come each year. It is simply not feasible in the foreseeable future that the SSE model could be applied on any reasonable scale.

STATISTICAL SOLUTIONS

In a third approach, a number of statistically-based procedures have been offered to solve the confidentiality dilemma. There are too many variants of these models to discuss them all. Among other variants, they include providing variance-covariance matrices to the researcher--transforming original data by rounding, recoding into intervals or adding random noise to the original data.

If the researcher can only obtain variance-covariance matrices, he will clearly not be able to identify respondents. But at what a cost! Models are essentially predetermined, leaving the researcher little to do except to push the run button. Variables not included in the moment matrix set are unavailable for analysis. Even something as basic as operational definitions of variables are outside the researcher's hands. I don't know how to specify my own moment matrices until I have extensively interacted with the microdata--an option not open to me in this approach. Finally, moment matrices also severely constrain the class of statistical models that can be explored.

The other variants of the statistical approach usually reduce to some type of transformation of the original data. This includes rounding, recoding the data, or more elaborate procedures to add random noise to the original data. On a limited basis, this is standard survey practice and is quite appropriate as the survey people try to obscure unusual values of variables that send strong signals about who the actual respondent

might be. But should it be used for more extensive transformation of the data? Systematic random noise or multiple imputation are good examples of proposals for more extensive use of this approach. I think not. Adding noise or error to the data is by no means innocuous to applied research. At our current knowledge, we have no way of knowing how severe the consequences of such imposed error is, nor do we have procedures to incorporate it into wide classes of models. It is somewhat ironic that the Census Bureau, who devotes so much of its resources to painstakingly enhancing the quality of the data, would in this model be systematically undoing its prior good deeds.

These three proposed solutions fail because they place an intruder between researchers and the data. In these models, someone other than the original researcher ends up making key decisions about how variables will be coded and what class of statistical models can be used. This breeds conformity and lack of originality--the death knell of good science. If the only solution involves full researcher access to the data, how can we insure that researchers will not violate the trust placed in them? That brings me to the solution I will try to champion--licensing of researchers.

LEGAL PENALTIES ON USES

The solution I favor involves licensing or bonding of researchers. This procedure includes statutory penalties on users for improper use of the data, and legal contracts that bind users for certain proscribed analysis and perhaps for a specified period of time. I find this procedure far more promising than the others because it places the fewest limitations on the actual research process. It is also not untested. Ohio State with the National Longitudinal Survey, and Michigan with the Panel Survey of Income Dynamics are releasing some sensitive data in this form, and the early results look promising. The Ohio State model has been in place since 1980, and there are no known instances of violations.

CONCLUSION

I think it would be a great mistake under the guise of confidentiality to rely on solutions that severely restrict extramural research. Some researchers believe that government agencies have other motives that are not as noble as confidentiality to limit extramural access to data. Limited public access to data not only gives intramural researchers a monopoly on the data, it also provides federal agencies an effective mechanism to control areas of sensitivity.

Even if those concerns are considered far-fetched, there is a more serious problem that clearly is not. Researchers outside the government can be far more provocative and daring in the questions they ask, how they ask them, and what their answers are. Internal government agency research tends, and quite appropriately so, to be far more conservative as higher-ups in the system are vigilant in overseeing what the political ripples might be. For example, it is a far different thing for the Census Bureau to argue for the appropriateness of a new poverty line than it would be for an individual researcher to do so on his own. Research that is often seen initially as inflammatory, provocative, and politically incorrect often has the largest impact on redirecting the science. It would be ironic to those of us who have been lecturing statistical bureaus in other countries where data access is quite limited that they should follow the more open U.S. model. Great progress has been made as countries move towards the American system. Hopefully, we reverse roles as our data becomes increasingly limited.

Let me finish by turning to the title of this session: "Who do you trust?" In the end, I believe it is not the scientific research community that the American public is worried about with confidential data. One reason is empirical. There does not appear to be an example of researcher violation of confidentiality with microdata. The other reason is that people know that individual researchers have no incentive to try to identify respondents. Rather, it is the current guardians of the data--the federal bureaucracy--(and especially the IRS) that is not trusted. The public would feel far more comfortable if we could solve the confidentiality issue by allowing access to the research community and prohibiting its use by the government.

Why are the guardians of the data so comfortable with access by intramural researchers? I presume that they would argue an employee of the federal government who revoked his trust would be severely penalized. Put the same penalties and more on the individual researcher where the incentive to violate respondents' confidentiality is much lower.