ACCESS TO CONFIDENTIAL STATISTICAL AGENCY DATA

Stephen H. Cohen
Bureau of Labor Statistics; Office of Survey Methods Research;
2 Massachusetts Ave, NE; Washington, D.C. 20212

Key Words: Confidentiality; Data Access; Federal Statistical Agencies

Introduction

The Federal Statistical Agencies collect a wealth of economic, demographic and social data. These data are collected to meet specific or general requirements in legislation or the code of federal regulations. The agencies publish estimates from that data in various specific tabular forms in paper or on the Internet. Initial publication could be a press release followed by bulletins that present more detailed statistical data and analysis. Usually there still are many additional possible tabulations that characterize or analyze a segment of the population that are not published due to the lack of resources within the agency.

Data rich agency microdata files could allow for very detailed analyses by researchers at their home institutions. Unfortunately, the data richness often means releasing a record potentially representing an individual or establishment that reflects a unique combination of characteristics of that individual or establishment, and the likelihood of some re-identifications (disclosures of confidential information) is increased in comparison to less data rich files. These data are collected under pledges of confidentiality. In some cases, agency employees are subject to severe legal repercussions for releasing identifiable information. The agencies use statistical disclosure control techniques to protect individual identification. These techniques involve data modification or partial suppression to avoid the release of too highly detailed confidential data.

Agencies also release public use data sets for researchers to further analyze on their own. However, since these data are collected under confidentiality laws or pledges, the amount of detail that can be released on public use data sets must be limited. Economic establishment data are never released as public use data sets. Geographic identifiers on demographic and social statistics must be suppressed or aggregated to levels that limit the analysis possible. Certain sensitive economic data elements, such as income in a household survey, must be topcoded. Detailed information such as occupational coding might have to be aggregated up to a higher level to prevent re-identification.

Since the statistical agencies can only produce a limited amount of potential outputs from the full microdata sets, the full potential of these data are not realized. One way of satisfying both concerns, the desire of researchers to have access to such files and the desire to prevent disclosures, is for the agency or research organization to allow researchers access the files under highly controlled conditions.

The availability of new technologies, which offer almost unlimited possibilities of remote and decentralized access, obviously lead to strong user expectations towards more and more flexible dissemination architectures. This article will explore four methods of restricted access procedures that are used to allow researchers to access confidential micro data: research data centers; remote access; licensing agreements; and research fellowships and post doctoral programs.

Research Data Centers

Research Data Centers (RDCs) are secure facilities designed for outside researchers to access confidential statistical agency microdata files. Initially statistical agencies usually only locate RDCs at their headquarters. After an agency has gained sufficient experience with these centers at their headquarters, additional RDCs outside the headquarters are possible. RDCs are both physically and electronically separated from agency data and personnel.

After an agency has decided to create a center by gaining agreement from within and outside, decisions have to be made about which data it holds will be made available for access. These decisions include the survey files, which will be available for analysis and the data elements collected that will be made available. Some files may be considered too sensitive to allow non-agency personnel access such as Internal Revenue Service tax files. Some data files may not be ready for outside use. Permissions may need to be obtained from survey sponsors (some of which may be in other government departments), providers of administrative data underlying the agency's programs, and possibly higher levels within the agency's department (such as departmental legal offices). Files should have adequate documentation

on definitions, data fields, etc. Access to certain sensitive identifiers such as name, address, social security number may not be allowed. Outside researchers might have conditions placed on use that are more restrictive than internal staff. Great care must be taken if an agency can't grant special sworn employee status to researchers, which would subject them to the same penalties as agency employees for confidentiality breaches. Agencies might restrict access for research only or to projects that generate specific benefits to the agency's programs.

In choosing site locations care must be exercised to ensure that selection process is fair. Solicitation announcements should be made in the federal register in addition to distribution to likely candidate organizations. It might be advisable choose the sites with a partner such as the National Science Foundation as the US Census Bureau did. The evaluation process should be fair and objective. As RDCs impose considerable costs on the agency, the agency must decide which options to use to recover the costs associated with RDCs. Costs can be recovered by charging researchers directly or charging the host organization whom can recover their costs by charging laboratory fees. The US Census Bureau (BOC) and the National Center for Health Statistics (NCHS) charge researchers directly at headquarters. BOC charges hosts for remote sites.

The RDCs must be secure facilities not only physically but also procedurally. All materials researchers removed from the facilities must be reviewed for confidentiality. The computer facilities must have no links to or from the outside and the "A" drive and/or other write medians disabled. The site must have a trained on-site employee or contractor who is trained in security and the data sets.

The NCHS has as RDC only at its headquarters while BOC has remote locations in addition to its Washington, D.C. headquarters. The NCHS RDC is a secure monitored facility where external researchers may be allowed access to internal restricted data files for approved projects. Restricted data files are those which contain information, such as lower levels of geography (e.g., state, county, or lower), but do not contain direct identifiers (e.g., name or social security number). Restricted data files may be used in the RDC by researchers wishing to control for geographic area in their models or they may be used to merge additional data onto the NCHS collected data files for enhanced analyses (e.g. The HCHS contextual data file.) To gain access to the NCHS RDC researchers must follow the strict procedures that govern the use of the RDC:

- researchers must submit a research proposal that describes the work to be done and the output proposed to be released
- no materials may be brought into the RDC
- no materials, printed or electronic may leave the RDC without a disclosure review
- researchers must sign a Researcher Affidavit of Confidentiality
- the RDC is open only when staff are available for supervision
- use of the RDC is subject to space availability, consistency with the NCHS mission and
- the feasibility of the proposed project.

Except for very unusual circumstances, researchers are not allowed access to files with direct geographic identifiers. Should a researcher request an NCHS data file merged with external data, RDC staff will merge the files then remove the geographic identifiers leaving the researcher access to a files that consists of the NCHS data merged with the additional data. Should the researcher need clustering variables to stratify on geography, RDC staff will construct a set of dummy geographic indicators.

RDCs are not special tabulation shops where, for example, local economic development agencies can get detailed geographic/industry breakouts that are not released in agency publications.

Remote Access

For many researchers, working at an RDC is burden because of travel and being away from his/her host institution. Remote access allows researchers to submit analytical programs to run against confidential microdata files from outside the statistical agency. Here too many decisions need to be made. Decisions need to made on the languages that will be supported, medium to be used to submit the programs and review procedures for the output generated. Usually, remote access in not a method that can produce tabulations not previously released.

At NCHS SAS was chosen as the analytic language because it is in wide use and is sufficiently well structured that an automated scanning system could be used. A number of functions available in SAS have been disabled because they are capable of producing unstructured output that present an unreasonable risk of disclosure. Disabled functions include PROC TABULATE, PROC IML, PROC PRINT, LIST, and others.

The current NCHS remote access system operates by e-mail but an internet-based system is under development and testing. The internet-based system offers an user-friendlier interface and is capable of improved turn-around time.

The RDC staff will construct a dummy data file configured exactly like the real data (univariate distributions, variable locations and lengths are the same, and paths are the same) that the researcher can use for developing and debugging programs prior to sending them to the remote access system. The use of the dummy data file results in less iteration on the remote access system thus increasing overall efficiency. The remote access system operates entirely automatically: the system scans the e-mail for arriving computer programs, validates the user, scans the program for non-allowable commands (such as those which could result in a case listing), verifies that it is not trying to access unauthorized data files and, if no problems are found, executes the program against the real data. After execution, the system scans the analytic output generated by the user's program for disclosure problems. Questionable output is routed to an RDC staff person for manual resolution. Users can submit requests to the remote access system 24 hours a day although output is only returned during normal working hours because staff randomly spot check the system to ensure that the system is working properly in all respects. Generally users receive their output within a few hours after submitting their e-mail.

Licensing Agreements

A licensing agreement is a formal agreement that permits confidential microdata to reside on a researcher's personal computer in their home institution. These agreements are formal legal documents between the agency and the host organization that specify the conditions under which the specific data set licensed may be used and the penalties for violation. Licensing agreements are issued for some microdata sets by:

- Social Science Research Organizations such as
  - ✓ Inter-university Consortium for Political and Social Research (ICPSR)
  - ✓ Survey Research Center at the University of Michigan (SRC-UMICH)
- U.S. Government Agencies such as
  - ✓ National Center for Education Statistics (NCES)
  - ✓ National Science Foundation (NSF)
  - ✓ Department of Housing and Urban Development (HUD)
  - ✓ Health Care Financing Administration (HCFA)
  - ✓ Social Security Administration (SSA) and the
  - ✓ Bureau of Labor Statistics (BLS).

There are several common themes that run through the licensing agreements.

The principal researcher must demonstrate that the data is required for research; i.e., public use data, if it exists, is not adequate. The goals of the research that require non-public data must be stated in the application. The licensor must approve the goals of the research before the application process can proceed.

The agreement specifies which people in the licensee's institution must sign the form. For an academic department it is typically a Dean and not the department chairman. The principal researcher (PR) must supply a list of names of people who will be authorized to use the data. The list must be updated as personnel working on the data changes. Those people must be informed of their responsibility not to share the data with people outside the group. The PR must indicate the group's experience, if any, with handling other licensed data sets.

The agreement also includes a statement concerning which law(s) protects the data (e.g., Privacy Act of 1974). A data security program must be developed and implemented. The licensee's institution must allow inspections of the area where the data are used and stored The inspections can be unannounced. Penalties for violations of aspects of the agreement are listed on the form (e.g., denial of use of other data from the licensor, fines, prison terms, etc.). There is a requirement that no attempt will be made to determine the identity of respondents. In general, the licensee is not allowed to link the licensed data to other microdata files.

Articles, reports, and statistical summaries generated from the data must be reviewed by the agency before they are published or otherwise communicated. The results must adhere to the agency's disclosure limitation practices (e.g., all non-zero cells in a publicly released table must represent some minimum number of respondents).

Some examples of datasets released under licensing agreements include: NCES'S s Schools and Staffing Survey and The Early Childhood Longitudinal Study; BLS'S Census of Fatal Occupational Injuries and The National Longitudinal Survey of Youth; and NSF's

Survey of Doctorate Recipients and Survey of Earned Doctorates.

Fellowships and Post Doctoral Programs in Principal Statistical Agencies

Research Fellowships and post-doctoral programs provide the unique opportunities for researchers to address some of the complex methodological problems and analytic issues relevant to agency's programs. Fellows and Post-doctoral candidates will conduct research in residence at agency, use agency data and facilities, and interact with agency staff. They will adhere to the same confidentiality agreements as regular employees.

Research fellowship applicants should have a recognized research record and considerable expertise in their area of proposed research. The American Statistical Association (ASA) administers the Research Fellowship Programs, with some support from the National Science Foundation (NSF) for three Federal statistical agencies: the BOC, the BLS, and the NCES. The ASA also administers a Research Fellowship Program for the NCHS and the Bureau of Economic Analysis (BEA).

Application materials are nearly identical for both programs. Fellowship applicants must provide the following:

- curriculum vitae
- names and addresses of three references
- three (3) copies of a detailed research proposal that includes:
  - ✓ a short descriptive project title
  - ✓ an abstract one-half page or less)
  - ✓ a proposed project term (approximate dates)
  - ✓ background information on research topic, references, etc.
  - ✓ a statement of relevant work already accomplished
  - ✓ proposed research with sufficient detail for evaluation of expected results
  - ✓ a statement regarding significance of expected results
  - ✓ statement citing the advantages of conducting the research at BLS and
  - ✓ a proposed budget.

Proposals are reviewed within the BLS and by an external Program Review Board consisting of representatives of ASA, BLS, the American Economic Association, and the American Association of Public Opinion Research. The proposal will be evaluated on the applicability of the research to BLS programs, the value of the proposed research to science, and the quality of the applicant's research record. Qualified women and members of minority groups are encouraged to apply.

Examples of recent research activities are indirect estimation, imputation, statistical computing, quantitative methods and data, confidentiality, data linkage and cognitive processes.

Research fellows are contract employees and subject to the same confidentiality pledge signed by agency personnel.

The federal government has many varied post-doctoral programs. In statistics they range from long established programs in the National Institutes of Health to recently developed programs at the BLS and the BOC. The goal of these programs at the BLS and at the BOC is to develop an interest in recent Ph.D. graduates on survey methodology research. It is hoped that these researchers might then become attracted to a career with the statistical agencies in survey methodology or goes on teach sample design word in academic institutions.

Post-doctoral researchers at BLS and BOC would become temporary employees; thus, subject to the confidentiality pledges that all employees must sign.

Post-doctoral candidates are restricted to recent college graduates, typically having received their Ph.D. degrees within the last 2 to 3 years. Their research proposal is written in conjunction with a senior methodologist within the agencies.

Summary

In some case, researchers can gain access to data rich detailed economic, demographic and social data for analysis. Various methods involving technology or legal arrangements make this possible. Technology enables researchers to run computer programs against an agency's statistical data from a home institution or to gain access to the data from secure agency facilities both of which do not compromise the confidential microdata. Via legal arrangements, researchers can, in some cases, get confidential data licensed for their use or gain legal status to access data just as agency employees.

The opinions expressed in this paper are those of the author and do not necessarily represent the policies of the U. S. Bureau of Labor Statistics.

References

Massell, Paul, Overview of Data Licensing Agreements at U.S. Government Agencies and Research Organizations, CDAC Paper

Jabine, Thomas B.(1993), "Procedures for Restricted Data Access," *J. Official Statistics*, vol. 9, no. 2, pp. 537-589.

Massell, Paul B.(1999), "Review of Data Licensing Agreements at U.S. Government Agencies and Research Organizations," paper presented at the Workshop on Confidentiality of and Access to Research Data Files, sponsored by the Committee on National Statistics (CNSTAT), Washington, D.C.

Massell, Paul B., Laura Zayatz, (2000), "Data Licensing Agreements at U.S. Government Agencies and Research Organizations," *Proceedings of ICES-II (International Conference on Establishment Surveys).*

George.T. Duncan, Thomas B. Jabine, Virginia A. de Wolf (eds.), *Private Lives and Public Policies*, National Academy Press (1993), in "Chapter 6 : Technical and Administrative Procedures," pp. 141-179.

Jabine, Thomas B., "Procedures for Restricted Use Access." *Journal of Official Statistics*, 9:2, 1993, pp. 537-589.

National Center for Education Statistics, "Restricted Use Data Procedures Manual."

Reznek, Arnold., Joyce. Cooper, and J. Bradford Jensen. "Increasing Access to Longitudinal Survey Microdata: the Census Bureau's Research Data Center Program." *American Statistical Association 1997 Proceedings of the Section on Government Statistics and Section on Social Statistics.* Alexandria, VA, 1997, pp. 243-248.