

New Tools for Accessing Data While Still Protecting Confidentiality¹

ANCSDAAP Population Census Conference
Hong Kong, March 25-27, 2009

Daniel H. Weinberg, Ph.D.
Assistant Director for Decennial Census Programs

Nancy M. Gordon, Ph.D.
Associate Director for Strategic Planning and Innovation
U.S. Census Bureau

Contact information: <daniel.h.weinberg@census.gov>

March 2, 2009

Abstract

Traditionally, national statistical offices (NSOs) have released tabulations, and some have released public-use microdata files that have been protected against disclosing confidential information by using techniques like top-coding variables and limiting geographic detail. More recent methods of disclosure avoidance include creating synthetic data and noise addition.

Because not all research and information needs can be met by these tabulations and data files, NSOs have developed three key alternatives for providing non-government researchers with access to confidential microdata to improve statistical modeling. The first, licensing, allows qualified researchers access to confidential microdata at their own facilities, provided certain security requirements are met. The second, statistical data enclaves, offers qualified researchers restricted access to confidential economic and demographic microdata at specific agency-controlled locations. Thirdly, statistical agencies can offer remote access, through a computer interface, to the confidential data under automated or manual controls so that only disclosable results of the analysis are returned to the researcher, who is unable to view or copy any of the confidential data.

New forms of disseminating information to non-technical users are also being developed. One U.S. Census Bureau approach, called *HotReports*, provides statistical information relevant to a particular topic using charts and maps that can be easily understood. The underlying (public-use) data appear and are available for downloading by clicking on the display.

¹ This paper was prepared for the Association of National Census and Statistics Directors of America, Asia, and the Pacific (ANCSDAAP) 24th Population Census Conference, 25-27 March 2009 in Hong Kong, China. It is based on Weinberg et al. (2007). It includes the results of research and analysis undertaken by Census Bureau staff and has undergone a more limited review than official Census Bureau publications. The authors wish to thank Laura Zayatz for her comments and suggestions.

New Tools for Accessing Data While Still Protecting Confidentiality

The United States has conducted a population census every ten years since 1790 and has provided reports from those censuses to legislators, government executives, and the general public as soon thereafter as practicable. For most of U.S. history, those reports have been printed tabulations. It is only since the 1960 Census that the U.S. Census Bureau, hereafter the Census Bureau, has used any other dissemination methods. In the U.S., government agencies collect such data under strict statutory guidelines that require confidentiality -- the protection of a respondent's identity from public disclosure. The necessity for protection leads any agency in the direction of less openness. Yet the public good, and the reason that an agency receives public funds, push the agency in the other direction: release as much information as possible. This conflict must be resolved, at least within the U.S., within the bounds of the agency's enabling legislation.

There are five basic approaches to providing information to the public. First is the provision of tabulations from the data that are collected, always accompanied by statistical notes but sometimes also accompanied by explanatory text. This approach is so common as to not require additional description or discussion. But it is worth noting that for the 2000 census, the Census Bureau started using the Internet to provide these tabulations to the public using the American FactFinder (AFF). The AFF is the "data access and dissemination system" for online access to all tabulations from the 1990, 2000, and (upcoming) 2010 U.S. Population and Housing Censuses, including tabulations from the American Community Survey (ACS), the survey begun in 2005 as a replacement for the "long form" sample census. (It also provides access to results from our economic censuses, to tabulations from a number of business surveys, and to post-census population estimates. The AFF, found at <factfinder.census.gov>, is described further in Appendix A.)

The second method is the provision of "public-use microdata samples" (PUMS) from which independent researchers can produce their own analyses. The Census Bureau pioneered the use of PUMS in the 1960s and now produces a wide variety of such data files for its household surveys. Among the techniques used to protect a respondent's identity on the PUMS are variable suppression, top- and bottom-coding, re-categorization, adding noise, swapping, and geographic aggregation. Yet the computer revolution of the past half-century, especially the Internet, has made it increasingly possible to decode the information on such files, and agencies have responded by reducing the amount of information they release on such files. Furthermore, microdata files from business surveys have only rarely been made public, so independent research using representative business data has been difficult or impossible. Again, though this method is less venerable than tabulations, it has been in use for approaching a half-century and will not be further discussed in the body of paper.

However, a fairly recent development -- the creation of *synthetic data* that mirrors the properties of the collected data yet fully protects the confidential data provided by respondents -- is

discussed in Attachment B. Yet another advance in disclosure avoidance techniques – *adding noise* – is described in Attachment C. As these techniques are further developed, it will be possible to meet more, but not all, needs of data users through publicly released products.

The three other methods for disseminating the results of surveys that must remain confidential are the main subject of this paper.² Section I discusses *licensing* – whereby the statistical agency provides restricted data directly to individuals or organizations under a confidentiality protection agreement. Section II discusses *research data centers* – statistical enclaves where “outsiders” can undertake research at statistical agencies. Section III discusses *remote access*, whereby researchers can submit analysis requests (typically computer programs) to agencies and receive the results of those analyses. The last section of the paper, Section IV, covers a new way to integrate and disseminate publicly available statistical information on a specific topic, drawn from many different sources, to non-technical users.

I. Licensing

If public-use microdata samples cannot provide sufficient information to researchers, agencies will sometimes “license” organizations to analyze “restricted-use” confidential microdata. Typically, the license document defines the information subject to the agreement; specifies the individuals who may have access to subject data; describes limitations of disclosure avoidance and clearance procedures; lists administrative requirements; requires that copies of publications based on the data be sent to the sponsoring agency; requires the organization to contact the sponsoring agency in case of (suspected) breaches of security; requires the organization to agree to unannounced and unscheduled inspections; reviews the security requirements for the maintenance of, and access to, subject data; and describes penalties for violations.

The U.S. National Center for Education Statistics uses this method for a large number of its confidential datasets. As noted on their website <<http://nces.ed.gov/pubsearch/licenses.asp>>, “The goal is to maximize the use of statistical information, while protecting individually identifiable information from disclosure.”

The U.S. Bureau of Labor Statistics (BLS) has established a similar program for access to its National Longitudinal Surveys (NLS) of Youth. As its web site <<http://www.bls.gov/nls/nlsfaqs.htm#anch25>> notes,

BLS has established a licensing system through which legitimate researchers at universities and other research organizations in the United States can use NLS data with geographic information at their own facilities, provided that the research project and physical and electronic security measures described in the NLS geocode application are approved by BLS. ... To protect the confidentiality of respondents, the BLS only grants access to geocode files for researchers in the United States who agree in writing to adhere

² This paper focuses on U.S. practice. Eurostat’s Centre of Excellence for Statistical Disclosure Control (CENEX) *Statistical Disclosure Control Manual* (Hundepool et al. 2007) contains a general discussion of European approaches to statistical disclosure control, research data centers, remote execution, remote access, and licensing, with a specific reference to German official statistics.

to the BLS confidentiality policy and whose projects further the mission of BLS and the NLS program to conduct sound, legitimate research in the social sciences. ... Applicants must provide a clear statement of their research methodology and objectives and explain how the geocode data are necessary to meet those objectives. Researchers who are granted access to NLS geocode files may use them at their own facilities, provided that the facilities meet BLS security requirements.

Other confidential BLS datasets can be accessed only at BLS headquarters, once an application is approved. The older NLS cohorts can only be accessed through the Census Bureau's Research Data Centers (see section III), as the Census Bureau does not use licensing because of the way its authorizing statute is written. The U.S. National Science Foundation Division of Science Resource Statistics also uses restricted data licenses to allow researchers to access some of its confidential data.

Licensing is also used in non-government settings in the U.S. The University of Michigan's Institute for Survey Research (ISR) licenses the use of a confidential (geocoded) version of its Panel Study of Income Dynamics (PSID). Other surveys use this method, such as the Fragile Families and Child Wellbeing Study conducted by Princeton University, which releases geographic identifiers to the public via a restricted-use data agreement. Another example is the National Data Archive on Child Abuse and Neglect at Cornell University and its Longitudinal Studies of Child Abuse and Neglect.

II. Statistical Data Enclaves

As noted earlier, the statutory provisions under which U.S. statistical agencies collect data (such as the Confidential Information Protection and Statistical Efficiency Act of 2002, and, for the Census Bureau, Title 13 of the U.S. Code) prevent the release of the full detail of survey data (such as names, addresses, and other information that would allow respondents to be identified) in order to protect their confidentiality. As administrative data about individuals become more and more available through the Internet, statistical agencies must reduce the detail about individuals available through public-use microdata. The availability of such data through the research enclaves can help ensure that valuable research can continue. Further, since business microdata has only rarely been in the public domain, the enclaves allow microeconomic research on businesses that could not otherwise take place.

The Census Bureau now has nine and will shortly have ten data enclaves around the United States termed Census Research Data Centers (RDCs). The RDCs are partnerships with academic and non-profit organizations. They are Census Bureau facilities managed by the Census Bureau's Center for Economic Studies (CES), staffed by a Census Bureau employee, and meeting all physical and computer security requirements for restricted access. RDCs offer qualified researchers restricted access to confidential economic and demographic data collected by the Census Bureau and other federal agencies in their surveys and censuses. (See <<http://www.ces.census.gov/index.php/ces/cms/home>> for more information.)

RDCs are aimed at researchers in academia, in independent research organizations, and in federal, state, and local government agencies. Generally, tabulations of confidential data may not be removed from the Census Bureau's RDCs, and therefore estimation of statistical models is the focus of their activities (though tabulations are permitted in other countries' RDCs, such as in Canada). All researchers are required to acquire Special Sworn Status from the Census Bureau, and are then subject to the penalty provisions of its authorizing legislation, should there be a confidentiality violation (a fine of up to US\$250,000 and/or up to 5 years in prison).

The objective of the Census Bureau's RDCs is to increase the utility and quality of data products while maintaining the security and confidentiality of such data. Access to microdata encourages knowledgeable researchers to become familiar with an agency's data products and data collection methods. More importantly, providing qualified researchers access to confidential microdata enables research projects that would not be possible without access to respondent-level information. This increases the value of data that has already been collected.

Access to the microdata also allows for data linking not possible with aggregates – both cross-survey linkages and longitudinal linkages. These linkages leverage the value of existing data. Creative use of microdata can address important policy questions without the need for additional data collection. All of the actual processing of data for approved projects is conducted on servers located in the Census Bureau's secure central computer facility. Researchers located in the RDCs use thin clients (terminals) to access these servers over the Internet via Virtual Private Networks. Researchers may also bring data into the RDCs and arrange for linkage to Census Bureau datasets.

In addition, the best means by which the Census Bureau can check on the quality of the data it collects, edits, and tabulates is to make its microdata records available in a controlled, secure environment to sophisticated users who, by employing the microdata records in the course of rigorous analysis, will uncover the strengths and weaknesses of those records. Each set of observations is the end result of many decision rules covering definitions, classifications, coding procedures, processing rules, editing rules, disclosure avoidance rules, and so forth. The validity and consequences of all these decision rules only become evident when the Census Bureau's micro databases are tested in the course of analysis. Exposing the conceptual and processing assumptions that are embedded in the Census Bureau's microdata databases to the light of research constitutes a core element in the Census Bureau's commitment to quality.

The opportunities for researchers to carry out unique research come at a price. Research conducted at RDCs takes place under a set of rules and limitations that are considerably more constraining than those prevailing in typical research environments. Research proposals are reviewed on the basis of five major standards: (1) benefit to the Census Bureau's programs, (2) scientific merit, (3) clear need for non-public data, (4) feasibility, and (5) whether the planned output will meet the Census Bureau's disclosure avoidance requirements for public release. The output of RDC projects can be methodological or statistical and includes both scientific papers and benefit statements addressing the Census Bureau's needs. Output undergoes disclosure avoidance review under rules established by the Census Bureau's Disclosure Review Board, which may review particularly difficult situations itself.

While the Census Bureau contributes approximately 55 percent of the full costs of the RDC network, the remaining costs must be recovered from sources outside the Census Bureau (which may include funding from other government agencies). The university and non-profit organizations that operate the non-headquarters RDCs typically contribute the space in which the RDCs operate, and provide “release time” to the professor or individual who serves as the RDC’s Executive Director. But they must also pay the salary of the RDC Administrator (a Census Bureau employee), raising those funds in a variety of ways – as a direct contribution of the partner institution, through membership fees from a funding consortium, by charging fees for access, or a combination of these methods.

One recent development that will increase the utility of the RDC network to researchers is the decision to allow the confidential data of other federal agencies to be available through the RDCs. So far, the U.S. National Center for Health Statistics (NCHS) and the U.S. Agency for Healthcare Research and Quality have agreed to make their confidential data available in that way. NCHS has its own RDC at its headquarters, but no other location.

Other countries have adopted the RDC approach. By far the most advanced (in some ways surpassing the U.S. approach on which it was based) is the Canadian RDC network, but for demographic data only (see <<http://www.statcan.ca/english/rdc/index.htm>>). This is a true network in which the leadership is by a coordinating council of partner institutions, and the central statistical office, Statistics Canada, plays a facilitating rather than a lead role (and hosts a “federal” data center), with primary funding coming from the partner institutions and granting agencies.

The United Kingdom has also established a Virtual Microdata Laboratory, where academics and government officials can access confidential firm-level (business), controlled-access census, and potentially other microdata files under special license (see <<http://www.ons.gov.uk/about/who-we-are/our-services/vml>>).

The relatively new Research Data Centre of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) – a pilot program from 2004 to 2006 – has stated its goal as “to facilitate access to BA and IAB micro data for non-commercial empirical research using standardized and transparent access rules.” There are currently four locations (see <<http://fdz.iab.de/en/pageTextModulRight.asp?PageID=1>>).

Also, Denmark established an RDC in 1987 for population register-based research; it is now closed (replaced by a remote access system). New Zealand has also had an RDC program since 1997, now in three locations (see <<http://www.stats.govt.nz/products-and-services/microdata-access/data-lab/default.htm>>).

III. Remote Access to Confidential Microdata

Remote access systems make it possible for users to analyze restricted microdata without visiting a statistical enclave. The remote access systems provided by statistical agencies employ

automated and manual filters that block certain kinds of queries and results and must be monitored automatically and/or manually for disclosure avoidance; extracts of microdata and direct access to the records are not permitted. The data files available for analysis are usually partially protected against disclosure using some of the same techniques as those used for public-use files. However, they tend to provide more detail to researchers to carry out their analyses than do public-use files, but less detail than available in an RDC.

The Luxembourg Income Study (LIS, at <http://www.lisproject.org/>) is the oldest of the data suppliers that give users remote access to restricted microdata. LIS began in 1983 to harmonize income data on household surveys from a number of countries. Its data managers developed the LISSY remote access system to allow users from around the world to analyze the household surveys included in its database. That system has (consciously or unconsciously) served as a model for many other remote access systems currently in use and under development.

Canada and Denmark have given users remote access to restricted microdata since 2001; other countries providing remote access systems are Australia, the Netherlands, New Zealand, and Sweden. In the U.S., the National Center for Education Statistics and the National Center for Health Statistics gave users remote access to restricted microdata beginning in 1997 and 1998, respectively. The Census Bureau began disseminating Census 2000 microdata tabulations via remote access in 2003, after pilot tests in 2002 (see Rowland and Zayatz 2001).

In addition to statistical protections, such systems require software and security support. Software run on government systems, particularly those with external interfaces is subject to a variety of regulations – including a detailed and extensive security plan. Platform dependencies must be held to a minimum, several software applications brought together, a large catalog of metadata constructed and fixed, and a detailed user interface maintained. (See Steel 2006 for more information.)

Another important methodological aspect of a remote system is automating complementary disclosure review to avoid disclosure of confidential data that could otherwise result from combining multiple outputs. Although research in this area has been undertaken (see Duncan et al. 2000), no comprehensive mechanism is known to prevent complementary disclosure. This may be due to the difficulty and expense of automating such procedures, frequently resulting in staff of the agency conducting much of the disclosure avoidance review.

One system, the Advanced Query System (AQS) developed by the Census Bureau to provide remote access to data from Census 2000, incorporates completely automated disclosure review. This savings in staff resources is accomplished by having extremely strict limitations on the tables that can be created. As a result, many of the tables available through the AFF (which were reviewed by staff) would not be released if they were created using the AQS. Its usefulness comes not from pushing the boundaries of disclosure avoidance, but by allowing users to choose combinations of variables, or categories for the variables, that were not covered in the AFF.

To avoid complementary disclosures, the AQS will only accept requests that meet certain conditions. Analysts must choose the way each variable is collapsed into categories from the

options permitted by the Census Bureau. The smaller the population in the geographic area for which the tables are requested, the coarser the allowed tabulation choices. Geographic areas must follow the boundaries specified by the Census Bureau, and in addition to selecting geography and a universe, at most three variables may be tabulated at a time. After the tables that meet these conditions are produced, the AQS institutes another automated review using filters that preclude sparse tables and tables with small cell sizes from being returned to the requestor.

IV. Supporting Decision Makers by Providing Easily Accessible Statistical Information

Decision makers and policymakers need to respond to dynamic situations, particularly unexpected major events that require sophisticated access to official data. For example, major natural disasters or significant economic events demand timely access to information that often comes from many different agencies. However, finding them, tabulating them, organizing them, and presenting them in a format that is useful for decision makers may take so much time and so many resources that opportunities for efficient and effective action can be lost. The bottom line is that, when non-technical decision makers need real time access to information from several sources presented in ways they can understand, NSOs need a new way of meeting their needs.

The Census Bureau has taken a three-part approach to respond to these challenges.

- *HotReports* organize data into usable information for decision makers who are not statistical professionals.
- *DataFerrett* provides powerful statistical tools for sophisticated analysts to use for their own work and to create the displays in *HotReports*.
- *TheDataWeb* provides a fast and reliable infrastructure to gain access directly to data stored remotely to support *DataFerrett* and *HotReports*.³

This paper focuses on *HotReports*, which are interactive reports, that use Internet Web 2.0 interactive presentation technology to change data into information. They use the tools of *DataFerrett*, first, to draw data from *TheDataWeb* about specific variables relevant to an issue and a local area and, second, to display that information in graphs, maps, simple tables, and interactive text boxes.

This guided use of statistical data is possible because analysts, who are knowledgeable about both the issues being examined and the data themselves, design the reports. The policy maker is

³ To normalize data streaming from different databases, vendors, statistical packages, and data structures, technical specialists transform the underlying structural rules of each dataset (the metadata) from their original formats to that of *TheDataWeb*. This transformation enables the tools in *DataFerrett* to operate on, and integrate, information from many databases in a consistent manner. The initial investment in creating consistent documentation for many datasets pays off repeatedly, by enabling this software framework to provide data searching capacity and rules for proper statistical usage and data integration that are invisible to users, who need to learn only one set of tools to be able to access and manipulate data from a variety of sources.

given a package of appropriate information that has been selected from the myriad different and potentially confusing statistical measures that exist. In other words, a HotReport is organized to show analytically useful patterns in the data for a particular geography, or over time. The decision maker reviews information rather than being faced with tabulations from many different, complex, and possibly inconsistent datasets.

The ease of use and dynamic quality of HotReports are particularly helpful in the U.S. to analyze issues such as regional economic development, emergency preparedness and response, public health planning, grant eligibility, and community performance indicators. Data relevant to these topics come from many different programs in many different agencies, and the files released to the public have different data structures.

One HotReport, titled “Community Economic Development,” is intended to give policy makers, economic development experts, and other users interested in local area development a general overview of some of the statistics available about their community. A prototype is available at <ced.census.gov>; the data in it are currently being updated to show the latest available information. This site provides indicators covering economic, demographic, housing, transportation, and community statistics, based on the suggestions of a group of local economic development experts. Although it is not exhaustive of the wealth of data available from the U.S. federal government, this HotReport provides a selection of basic and detailed measures about local areas drawn from multiple government agencies’ surveys and programs based on administrative records.⁴

Another HotReport, designed to help with evacuation planning before Hurricane Wilma struck the coast of Florida in 2005, is available by going to www.census.gov and, in the *Special Topics* section, choosing *Census Bureau Data and Emergency Response*. Then, in the left column under *Past Emergencies*, choose either *Hurricane Katrina* or *Hurricane Wilma*. Then, under *Maps*, choose the *Demographic Profile for Hurricane Katrina [or Wilma] Affected Counties*. The web addresses are:

http://mongoose.dsd.census.gov/TheDataWeb_HotReport/servlet/HotReportEngineServlet?reportid=ff21413952644d76e21de0b6aa29fd15&emailname=whazard@census.gov&filename=katrina_dem3.html

<http://mongoose.dsd.census.gov/TheDataWeb_HotReport/servlet/HotReportEngineServlet?reportid=cf2ddd1cb2ebb7895713473ce73e25d&emailname=whazard@census.gov&filename=wilma_dem4.html>

V. Concluding Comment

Microdata are the foundation on which our understanding of human and business behavior is based. Yet threats to confidentiality are increasing as computers get more powerful and more and more information about identified individuals is available to the public on the Internet. In the

⁴ Surveys include the American Community Survey and the Census 2000 long form (for the smallest geographic areas that will not have ACS data released until 2010). Administrative data and products based on them include: Local Employment Dynamics, County Business Patterns, Population Estimates, Common Core of (Education) Data, Home Mortgage Disclosure Act reports, Integrated Postsecondary Education Data System, State Occupation Projections, and the Quarterly Census of Employment and Wages.

face of such threats, preserving the extent of current microdata using traditional disclosure avoidance techniques becomes less and less possible.

As a result, statistical agencies need to use new and different methods to make their information available to the public, in ways that preserve the ability of social scientists to manipulate the data for research purposes while preserving the confidentiality of respondents. This paper has focused on four ways that statistical agencies have responded: licensing, research enclaves, remote access, and new access methods aimed at non-technical users. New disclosure avoidance techniques – synthetic data and noise addition – that will expand statistical agencies options for releasing more public-use products are summarized in appendices.

Appendix A: The American FactFinder

From the American FactFinder (AFF, at <factfinder.census.gov>), one can obtain data in the form of maps, tables, and reports from a variety of Census Bureau sources. From the *Main Page*, users find links to data in the AFF and other Census Bureau sites.

New users will find data and maps on the most popular topics for their area by clicking on the following buttons from the Main Page:

Fact Sheet - for quick access to basic demographic, social and economic data on your city, town, county, state, or ZIP Code.

People - to find more popular tables on a variety of topics (age, education, income, race, and more) for your area.

Housing - to find popular tables on home values, ownership, housing characteristics, and mortgage amounts.

Business and Government - for data on housing starts, government finances, foreign trade, and more.

About the Data - Censuses and Surveys to learn more about how the Census Bureau organizes and presents data from many sources.

Use the *Help* button at the top of each page to go to specific help for the page you are on, and for more information about the site; read the introduction to the main page of the AFF for basic background on using the site.

Experienced users will want to use this direct link to choose their data set from all the data sets available in the AFF:

Data Sets – to access all data sets, tables, and maps.

Maps – to access reference and thematic Maps with tips on creating, using, downloading, and printing.

Reference Shelf – for resources, related sites, and reports.

Tools – for links to data manipulation tools.

Use the *Search box* to enter a keyword to find data, to enter the name of a geographic area, or to click on 'enter a street address' to find data for your area.

Tips for all AFF users include the following:

Site Map – provides an interactive outline of the AFF that describes how the site is organized and how to find a specific page.

American FactFinder Site Tour – is a tutorial that covers the basics of site navigation and finding data, as well as links to the other tutorials on ‘Search’, ‘Tables’, ‘Maps’, and ‘Working with Economic Data’.

The buttons on the banner at the top of each page provide much functionality:

Main - takes you back to the Main Page.

Search - accesses the Search box.

Feedback - enables users to ask a question, report a problem, or make a comment or suggestion.

FAQs - provides responses to the most frequently asked questions (FAQs) about using the AFF.

Glossary – gives definitions and explanations of words and terms shown on the AFF and in the data tables.

Help - provides help for every page in the AFF, as well as linking to a *Table of Contents* for Help, the Glossary, Census Data Information, and the Tutorials.

Appendix B: Synthetic Data

The newest approach to disseminating detailed information based on confidential microdata to the public is synthetic data. Creating synthetic microdata is a disclosure avoidance technology that protects confidentiality by replacing actual microdata with data that have been simulated. Data files are fully synthetic when all variables have been simulated; they are partially synthetic when some variables retain their original values.

Rubin (1993) and Little (1993) proposed the technique of synthesizing microdata to reproduce the statistical properties of the underlying confidential data while replacing all sensitive items with simulated values. Although the approach appears new, synthetic data were actually used for the first time by the Census Bureau in the preparation of the 1990 Decennial Census Summary Files, where it was called “blank and impute” (see Federal Committee on Statistical Methodology 2005, p. 32). In addition, the Survey of Consumer Finances has used methods that would now be called “synthetic” for confidentiality protection since editing the 1989 survey (see Kennickell 1991 and 1997, and the discussion in the appendix).

There are essentially two standards that synthetic microdata must meet. The first is disclosure avoidance standard. The synthetic microdata must be shown to protect the confidentiality of the underlying data to the standards set by the producer (agency), and to ensure that unlawful disclosures are avoided. Second, the synthetic microdata must be shown to provide inferences that are consistent with the inferences an analyst would have made from the original data, although possibly less precise because of the incremental uncertainty associated with the synthesizing process. This second standard is known as analytical validity or fidelity.

The use of synthetic data techniques began to expand with the advent of large-scale microdata based on longitudinally integrated employer-employee data, like those produced by the Census Bureau’s Longitudinal Employer-Household Dynamics (LEHD) program; see Abowd et al. (2004). These data are particularly difficult to protect using standard microdata disclosure avoidance techniques because the critical new information in the data (jobs) consists of the linkages between members of households and employers. Protecting these linkages is far more difficult than protecting products based only on information about households, or only about employers. In their 2001 and 2004 papers, Abowd and Woodcock reviewed the disclosure limitation methods that might be applied to such linkages and concluded that synthetic data provided a viable technique that could simultaneously provide disclosure avoidance and analytical validity if properly applied.

In 2001, a consortium of several U.S. federal agencies undertook a research project to create a public-use file from the Survey of Income and Program Participation linked to longitudinal social security benefit histories and longitudinal employee-employer earnings records. After investigating several feasible disclosure avoidance methods, the group agreed to experiment with partially synthetic data, and those data are now available. In 2005, the Census Bureau’s LEHD program released its first official synthetic microdata product – a public use application called *OnTheMap*, which is based on its longitudinally integrated employer-employee data (see <http://lehdmap2.dsd.census.gov/>). That application relates workers’ residence and workplace

addresses, allowing the user to map the residence locations of all individuals working in a user-specified geographic area, or the work locations for all workers living in a specific geographic area, again specified by the user.⁵ Finally, synthetic microdata from the American Community Survey were released in 2008 as part of the ACS 2006 Public Use Microdata Sample. In this case, the synthetic data techniques are being used to protect the confidentiality of ACS responses from residents of group quarters, such as dormitories, nursing homes, and prisons.

⁵ The synthetic data approach was necessary to meet the Census Bureau's disclosure avoidance requirements in large part because the user specifies geographic areas based on block boundaries, and blocks often contain only a small number of housing units or a small number of employers.

Appendix C: Adding Noise for Tabular Data⁶

The disclosure avoidance technique of adding noise to the underlying data prior to tabulation is an alternative to cell suppression, which has been used for decades for tabular data (Evans et al. 1998; Massell et al. 2006). Each respondent's data are perturbed by a small amount, say 10 percent (the actual percentage is confidential), in either direction. Noise is added in such a way that cell values that would normally be suppressed in tabulations (thus indicating the need for protecting them) are changed by a large amount, while cell values that are not sensitive are changed by a small amount. Because the distribution from which the noise factors are chosen is symmetric, the average value of each variable after adding noise is approximately its initial average value.

Adding noise has several advantages over cell suppression. It enables data to be shown in all cells in all tables. More importantly, it allows for the release of more usable data – in one survey conducted by the Census Bureau, twice as many usable cells were released than when suppression was used.⁷ It is a much less complicated and less time-consuming procedure than cell suppression. In particular, adding noise eliminates the need to coordinate cell suppression patterns between tables (known as complementary suppression). Because noise is added at the microdata level, additivity of the table is guaranteed.

To perturb a respondent's data, by say about 10 percent, multiply its data by a random number that is close to either 1.1 or 0.9. One could use any of several types of distributions from which to choose multipliers, though the actual distributions need to remain confidential within the statistical agency. The overall distribution of the multipliers should be symmetric around 1.0, but exclude a range of values near 1.0. Although the noise procedure does not introduce any bias into the cell values for census or survey data, it causes the variance of an estimate to increase by an amount equal to the square of the difference between the original cell value and the noise-added value. NSOs should incorporate this information into published coefficients of variation.

Applications to Real Data

A problem with the noise technique is that it can add excessive amounts of distortion to cells that would be shown much more precisely under deterministic methods such as cell suppression and controlled tabular adjustment. A natural question to ask then is whether there is a way to modify the method such that it adds less noise to the non-sensitive cells, while retaining the amount of protection provided to the sensitive cells. Massell and Funk (2007a) developed methods to reduce the overall amount of noise added to business data without compromising the level of protection.

The Census Bureau's magnitude data are almost always published in some rounded form, often in integer form representing thousands or millions of dollars. This type of rounding can be done at the record level prior to any tabulations being created, or applied directly to table values that

⁶ This Appendix is based on Zayat (2009).

⁷ By usable we mean that the value of the cell after adding noise was close to the initial value.

have not had rounding applied. Noise is designed to protect individual respondents by changing their response values by small percentages. For very small values, the amount that rounding changes them can be larger than the amount that adding noise changes them. Rounding can therefore systematically remove the effect of noise on small response values. Massell and Funk (2007b) investigated a few types of rounding methods that could work to sustain the protection provided by noise.

Current Uses for Census Bureau Public-Use Data Products

One of the initial data products to add noise to the underlying microdata to avoid cell suppression (and consequent sparse published tabulations) was the Census Bureau's program of Quarterly Workforce Indicators, which also uses some synthetic data (for discussion of synthetic data see Appendix B). Other Census Bureau programs that now use noise addition are the Commodity Flow Survey and tabulations for firms that have no employees (whose data are from administrative records). The Census Bureau plans to use the technique to publish tabulations from the 2007 Economic Censuses of the Island Areas (Guam, American Samoa, the Marianas, and the U.S. Virgin Islands) and for the Survey of Business Owners. Promising results suggest that it may also be used for the entire 2012 Economic Census.

References

- Abowd, J.M. and S. Woodcock. 2001. "Disclosure Limitation in Longitudinal Linked Data," in P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (Amsterdam: North Holland), 215-277.
- Abowd, J.M. and S. Woodcock. 2004. "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data," in J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases* (Berlin: Springer-Verlag), pp. 290-297.
- Abowd, J.M., J. Haltiwanger, and J. Lane. 2004. "Integrated Longitudinal Employee-Employer Data for the United States," *American Economic Review Papers and Proceedings*, Vol. 94, No. 2 (May), pp. 224-229.
- Duncan, G., S. Roehrig, and K. Kannan. 2000. Final Report on the American FactFinder Disclosure Audit Project for the U.S. Census Bureau.
- Evans, B. T., L. Zayatz, and J. Slanta. 1998. "Using Noise for Disclosure Limitation for Establishment Tabular Data," *Journal of Official Statistics* Vol. 14 No. 4, pp. 537-552.
- Federal Committee on Statistical Methodology. 1994, revised 2005. *Report on Statistical Disclosure Limitation Methodology*. Statistical Policy Working Paper 22 (Revised). <<http://www.fcsm.gov/working-papers/spwp22.html>>
- Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. S. Nordholt, G. Seri, and P-P. De Wolf. 2007. *Handbook on Statistical Disclosure Control*, version 1.01. A [Eurostat] Centre of Excellence for Statistical Disclosure Control. March. <http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf>
- Kennickell, A.B. 1991. "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," SCR Working Paper, prepared for the Annual Meeting of the American Statistical Association, Atlanta, GA, August <http://www.amstat.org/Sections/Srms/Proceedings/papers/1991_001.pdf>.
- Kennickell, A.B. 1997. "Multiple Imputation and Disclosure Limitation: The Case of the 1995 Survey of Consumer Finances" Chapter 8 in Federal Committee on Statistical Methodology (eds.) *Record Linkage Techniques - 1997*. Proceedings of an International Workshop and Exposition, Arlington, VA, March 20-21. <<http://www.fcsm.gov/working-papers/akennickell.pdf>>
- Little, R.J. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9(2), pp. 407-426.

Massell, P. and J. Funk. 2007a. "Protecting the Confidentiality of Tables by Adding Noise to the Underlying Microdata," *Proceedings of the 2007 Third International Conference on Establishment Surveys (ICES-III)*, Montreal Canada, June 18-21.

Massell, P. and J. Funk. 2007b. "Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection," *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology*, Arlington, Virginia, November 5-7.

Massell, P., L. Zayatz, and J. Funk. 2006. "Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey," Privacy in Statistical Databases, CENEX-SDS Project International Conference, PSD 2006, Proceedings, Lecture Notes in Computer Science (LNCS) 4302, Springer 2006, ISBN 3-540-49330-1.

Rubin, D.B. 1993. "Discussion of Statistical Disclosure Limitation," *Journal of Official Statistics* 9(2), pp. 461-468.

Rowland, S. and L. Zayatz. 2001. "Automating Access with Confidentiality Protection: The American FactFinder." *Proceedings of the Government Statistics Section, American Statistical Association*.

Steel, P. M. 2006. "Design and Development of the Census Bureau's Microdata Analysis System: Work in Progress on a Constrained Regression Server." Presentation at Federal Committee on Statistical Methodology Statistical Policy Seminar. November 28-29.

Weinberg, Daniel H., John M. Abowd, Sandra K. Rowland, Philip M. Steel, Laura Zayatz. 2007. "Access Methods for United States Microdata." U.S. Census Bureau Center for Economic Studies Discussion Paper CES-WP-07-25. August. Found at <http://www.ces.census.gov/index.php/ces/cespapers>.

Zayatz, Laura V. 2009. "New Ways to Provide More and Better Data to the Public While Still Protecting Confidentiality." *Proceedings of the 2008 Joint Statistical Meetings*, Denver, Colorado, August.