



IN THIS ISSUE

Brave New World of Synthetic,
Organic, and Sonic-Optic
Data 1
News from ICPSR 2
Researcher's Notes 2
Census 2010 and Paper-
Based Technology 3
Crossroads Corner 4

.....

DISC News contains articles about local, national and international data issues. It is published twice a semester by the library staff and sent to faculty, graduate students, and librarians, with a focus on the Social Science departments.

Editor: Joanne Juhnke, Special Librarian
Staff contributors: Cindy Severt, Senior Special Librarian; Janet Eisenhauer Smith, Data Analyst/Archivist;

.....

University of Wisconsin-Madison
1180 University Drive
3308 Sewell Social Sciences Building
Madison, Wisconsin 53706 USA

Phone: (608) 262-0750
E-mail: disc@mailplus.wisc.edu
Web: <http://www.disc.wisc.edu/>
Hours: Monday-Friday 8:30-4:30

.....

This issue of DISC News is online at:
<http://www.disc.wisc.edu/pubs/Newsletters/apr08news.html>

Please Note

DISC will be closed:
Mon. May 26 for Memorial Day
Fri. July 4 for Independence Day
Mon. September 1 for Labor Day

**BRAVE NEW WORLD OF SYNTHETIC, ORGANIC,
AND SONIC-OPTIC DATA**

by Cindy Severt

Mention quantitative analysis and it will take most readers of this Newsletter little stretch of the imagination to envision a numeric data file. Most social science microdata follows the classic pattern: individual responses to survey questions, coded for statistical analysis. For decades social scientists have reaped the social benefits of analyzing public use microdata files for insights into social behavior. Over the years, however, the boundaries of traditional social science data have stretched in various ways.

What if the variables in the datasets weren't limited to the survey instrument?

With rapid advances in technology and online access to public data, there has been an increase in the potential for respondent confidentiality to be compromised. As February's *DISC News* highlighted, restricting access to data has been an important approach in protecting confidentiality. Another solution to this problem has been to modify the data itself. Microdata files for public use are routinely modified in order to mask individual respondents' identities, often with unfortunate implications for the statistical quality of the data. Synthetic data is a new approach that generates multiple subsamples from the original survey data, protecting confidentiality while simultaneously preserving critical statistical properties of the data: mean, variance, and covariance. The data is inference-valid, but impossible to link back to an individual. The SIPP Synthetic Beta file (http://www.sipp.census.gov/sipp/synth_data.html) is the first of its kind to be released by the Census Bureau, created by integrating data from the Survey of Program Participation (SIPP), Social Security Administration (SSA), and Internal Revenue Service (IRS). Birth date, death date, marital history, and immigrant status are among the synthesized variables in this collaboratively-produced dataset, designed to reproduce the characteristics of the underlying confidential microdata.

At another point on the confidentiality spectrum one finds biosocial data, population-based sample surveys that combine demographic, social, and behavioral data with biological indicators. Most biosocial data refers to markers long associated with health surveys such as grip strength, pulmonary functioning, blood pressure, heart rate variability, weight and height, perceived age, clinical measurements of various substances in the blood, saliva, or urine, and various other measures of risk factors, exposures, and health outcomes which social scientists can use to better estimate environmental and behavioral effects on health. One example of such data collection

Continued on p. 3

NEWS FROM ICPSR

by Joanne Juhnke

Summer Program Stipend

By the time you read this, the application deadline (April 28) for the ICPSR Summer Program may be going or gone! However, DISC still does have a travel stipend available for defraying the cost of one individual traveling to Ann Arbor for the ICPSR Summer Program—approximately the cost of round-trip airfare from Madison to Detroit. The stipend is limited to UW-Madison, and preference will be given to students. To be considered for the travel stipend, contact your ICPSR Official Representative at DISC, Cindy Severt: 262-0750 or cdsevert@wisc.edu.

TIGER/Line Files@ICPSR

1990-2006 versions of TIGER/Line Files are now available for download. The new Web site, TIGER/Line Files@ICPSR (<http://www.icpsr.umich.edu/TIGER/>) offers data no longer available online from the US Census Bureau. TIGER stands for Topologically Integrated Geographic Encoding and Referencing and provides users with the appropriate data to prepare maps through geographic information system (GIS) software packages.

PK-3 Data Resource Center

ICPSR has announced a new website in cooperation with The Foundation for Child Development. The PK-3 Data Resource Center (<http://www.icpsr.umich.edu/PK3/>) provides access and supporting information for four longitudinal datasets selected for

Continued next column

RESEARCHER'S NOTES

by Ana Collares

I am a PhD candidate in the Sociology Department, and my research focuses on the expansion of Brazilian higher education during recent decades. My goal is to evaluate the consequences of this expansion with respect to inequality and social stratification in Brazil, especially the changes in the likelihood of access to higher education for students of different social backgrounds.

In order to do this research, I needed nationally representative data for the period. One of the best sources of data for this investigation is the National Household Sample Survey, or PNAD, which is coordinated by the Brazilian Institute of Geography and Statistics (IBGE). DISC already had several waves of PNAD data, some of them with codebooks translated from Portuguese to English. I recently had the opportunity to work with DISC personnel to acquire the last five waves of the survey (2001-2006). Their help has been crucial to the development of my research.

The PNAD was first implemented in Brazil in 1967, and has been repeated annually since 1970 (except for census years). This rich publicly-available data is multipurpose with a strong focus on the labor market, including questions about general aspects of the population, education, income and housing, as well as migration, marriage and fertility. Each year the core questions of the survey are kept unchanged, allowing for the investigation of national trends across time. Most waves also contain special supplements focusing in depth on specific social issues such as health, education, and female fertility. PNAD is an invaluable resource for students interested in population, labor and social stratification issues in Latin America, with a focus on Brazil.

NEWS FROM ICPSR

Continued from previous column

their potential to inform policy and practice regarding education from pre-kindergarten through third grade. The datasets are:

- Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K)
- National Head Start/Public School Early Childhood Transition Demonstration Study
- National Longitudinal Survey of Youth (NLSY), Child Surveys
- Panel Study of Income Dynamics (PSID), Child Development Supplement

The Foundation for Child Development has announced a small grants program to be funded through its PK-3 Research and Evaluation Forum. A maximum of four awards of up to \$50,000 each will be awarded to researchers proposing to use data from the PK-3 Data Resource Center. See <http://www.icpsr.umich.edu/PK3/spotlight/rfp.html> for more details.

BRAVE NEW WORLD...

Continued from p. 1

is the National Social Life, Health, and Aging Project (NSHAP), an in-home survey of 3,005 persons aged 57 to 84 that collected biomeasures of health and physiological functioning to better characterize the health of survey participants. Height, weight, saliva sample, and distance vision are some of the biomarkers collected (<http://biomarkers.uchicago.edu/timing.html>). NSHAP data is available through ICPSR.

When DNA *is* collected, genetic variations become variables, inevitably leading to questions such as whether or not genetic indicators should be analyzed in the context of understanding complex social traits. At what point do sociology and behavioral / natural science overlap? Are human behavior and culture a product of natural selection?

In yet another twist on social science data, what if the variables in the datasets weren't limited to the survey instrument? What if researchers could create their own variables from raw video footage? DISC staff members Lu Chou and Cindy Severt recently attended a Research Computing Workshop on TRANSANA, a software product for analyzing digital video or audio data. Developed by a graduate student and maintained at the Wisconsin Center for Education Research (<http://www.transana.org/download/index.htm>), TRANSANA works by coding events in video clips and linking those events to an audio transcript. To illustrate, a researcher studying how children learn in a classroom setting might notice recurring episodes of laughter during a 60 minute video. These moments can be tagged to a transcript and selected, not unlike selecting variables by column location. Laughter, silence, activity, or even facial expression can be "extracted" and exported as a tab-delimited file for analyzing with statistical software. What happens just prior to and after the laughter might be equally important, and TRANSANA offers a means of leveraging the richness of that context.

Sociology is an evolving field, and continues to cross-discipline itself into new sub-fields. Whether questionnaire-derived, artificial, biological, or qualitative: in the words of Vincent van Gogh, "there is nothing in the world as interesting as people, and one can never study them enough."

CENSUS 2010 AND PAPER-BASED TECHNOLOGY

by Joanne Jubnke

Plans for door-to-door data collection for the United States' 2010 Census took a decided turn for the non-technical in early April. The U.S. Census Bureau had planned to use wireless handheld computers both for collecting

Continued next column

CENSUS 2010 AND PAPER-BASED TECHNOLOGY

Continued from previous column

answers to the questionnaire, and for verifying residential street addresses. Now, due to a mishandled contract with an outside vendor, Secretary of Commerce Carlos Gutierrez has announced that the Census will have to rely on paper forms for the questionnaire data. The paper forms represent a return to the technology that the Bureau was attempting to leave behind for 2010.

The Florida-based Harris Corporation, which had been awarded the contract, will still work with the Census Bureau to provide handheld computers for address canvassing. For more information from the Census Bureau about the handheld computers, see <http://tinyurl.com/4bebl3>.

In an unrelated move in the direction of traditional data collection, the 2010 Census will not provide an option this year for respondents to answer the questionnaire online. This is a change from the 2000 Census, in which an Internet response option was made available for the first time, though with very little publicity due to confidentiality concerns. Until mid-2006, the Census Bureau had indicated that an Internet response option was intended as part of the 2010 Census, but then reversed that decision as plans moved forward. To read more about the Internet response option issue, see <http://www.itif.org/files/eCensusUnplugged.pdf>.

Crossroads Corner highlights web sites recently added to the searchable Internet Crossroads in Social Science Data, available on the DISC web site at <http://www.disc.wisc.edu/newcrossroads/index.asp>.

European Data Center for Work and Welfare (EDACwowe)

The EDACwowe site, at <http://www.edacwowe.eu/en/>, is a searchable collection of annotated links to websites containing data for European research and policy-making in the areas of work and welfare. The central topics covered by the site are income and benefits, social care, and work and employment. The links also touch on related fields such as demographics, education, taxes, health, migration, politics and elections, and quality of life.

EDACwowe organizes its site, via a left-hand menu bar, around the categories of Comparative Data, National Data, and International Repositories. The Comparative Data category is the most detailed, with subheadings for opinion surveys, socio-economic surveys, indicators and statistics, and policies and institutions. Each survey in the Comparative Data category gets a multi-part description on the EDACwowe site, including survey type, participating countries, topics, and availability and searchability of questionnaires and data. The National Data category, by contrast, gives only links and archive names, and the International Repositories category gives a short descriptive paragraph for each link.

EDACwowe is coordinated and supported by the University of Tilburg (The Netherlands) and by the Danish National Centre for Social Research.

USA Counties

The USA Counties database, at <http://censtats.census.gov/usa/usa.shtml>, has long been a part of the U.S. Census Bureau's

Censtat collection of online tools. The drop-down menu system allows users to create reports for the United States, the 50 states and the District of Columbia, or any of the 3,141 counties and county equivalents. Topics range from agriculture to building permits to elections to poverty. The data comes not only from the U.S. Census but also from other federal agencies such as Bureau of Economic Analysis, the Bureau of Labor Statistics, the Federal Bureau of Investigation, and the Social Security Administration.

This past month, the Census Bureau announced that downloadable data files in Excel format have been added to the USA Counties site. Users can now bypass the drop-down menus and directly download files by topic, each file containing data for all of the counties nationwide.

TrafficSTATS

The Traffic Statistics on Travel Safety site (TrafficSTATS) provides a novel online tool for calculating traffic safety risks on various dimensions. Using data from the Fatality Analysis Reporting System and the National Household Travel Survey, users can compare fatality risks for combinations of modes of transportation (e.g. car, SUV, motorcycle), demographic variables (e.g. age, gender), and other parameters (e.g. day of week, region of US). This can lead to fascinating combinations such as "teenagers driving at night in the summer" or "women over the age of 55 on motorcycles." Risks are stated in deaths per 100 million passenger miles, deaths per 100 million trips, and deaths per 100 million minutes of travel. Results can be downloaded in various formats such as XML, CSV, and Excel.

TrafficSTATS is a joint project between Carnegie Mellon University and the AAA Foundation for Traffic Safety, and can be found at <http://www.aaafoundation.org/trafficSTATS/>.