# ACCESS METHODS FOR UNITED STATES MICRODATA

Daniel Weinberg, US Census Bureau
John Abowd, US Census Bureau and Cornell U
Sandra Rowland, US Census Bureau (retired)
Philip Steel, US Census Bureau
Laura Zayatz, US Census Bureau

August 20, 2007

This presentation is intended to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on methodological issues are those of the presenter and not necessarily those of the U.S. Census Bureau.

# Traditional Methods of Data Access

1. Tabulations

- Confidentiality of respondents protected by limiting the number of cells relative to the number of observations
- May use complementary suppression

# Traditional Methods of Data Access

## 2. Public-Use Microdata Samples

- Confidentiality of respondents protected by omitting some information and modifying some of the remaining information
- Methods include  top- and bottom-coding, re-categorization, noise infusion, swapping, and geographic aggregation

# Four recent data access approaches

- Licensing – providing restricted data directly to individuals or organizations under a confidentiality protection agreement.
- Research data centers – statistical enclaves for research purposes.
- Remote access – submission of analysis requests (typically computer programs)
- Synthetic data – data that mirrors the properties of the collected data yet fully protects the confidential data provided by respondents.

# Licensing

Aspects of the license document :

- Defines the information subject to the agreement;
- Specifies the individuals who may have access to subject data;
- Describes limitations of disclosure and clearance procedures;
- Lists administrative requirements;
- Requires that copies of publications based on the data be sent to the sponsoring agency;

# Licensing, continued

- 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; and
- Reviews the security requirements for the maintenance of, and access to, subject data, and describes penalties for violations.

# Licensing, continued

Examples of U.S. uses:

- National Center for Education Statistics
- Division of Science Resource Statistics
- Bureau of Labor Statistics
- Panel Study of Income Dynamics

# Research Data Centers (Data Enclaves)

- The nine Census Bureau facilities are partnerships with academic and non-profit organizations (one at HQ).
- Staffed by a Census Bureau employee.
- Meet all physical and computer security requirements.
- Host Census Bureau and other federal agency data.
- Researchers become “Special Sworn Status” employees of the Census Bureau.

# Research Data Centers, continued

Goals:

- Increase the utility and quality of Census Bureau data products;
- Encourages knowledgeable researchers to become familiar with an agency's data products and data collection methods;
- Research can address important policy questions without the need for additional data collection;

# Research Data Centers, continued

Goals, continued:

- Improves the quality of data collection and processing practices by subjecting current methods to testing through additional uses;
- Allows for data linking not possible with aggregates that leverage the value of existing data;

# Census RDC Projects Must Meet Five Standards

- Benefit to Census Bureau programs (13 criteria used)
- Scientific merit
- Requires non-public data
- Must be feasible
- No risk of disclosure

Proposals must also pass a review by the Census Bureau policy office.

# Other Research Data Centers

Examples:

- Bureau of Labor Statistics HQ
- National Center for Health Statistics HQ
- National Institute of Child Health and Human Development (3 locations)
- National Opinion Research Center (2 locations)
- Canada, UK, Germany

# Remote Access

- Data files usually edited in advance to reduce the possibility of disclosure.
- Employ automated and manual filters that block certain kinds of queries and results.
- Must be monitored automatically and/or manually for disclosure avoidance. A difficult issue is complementary disclosure review.

# Remote Access, continued

Methodologies

- "Remote job execution systems" – an email interface that allows users to send programs; processing is usually done in batch mode.
- Web interface with custom–built or custom–tailored software.

# Remote Access, continued

Examples

- Luxembourg Income Study
- Canada, Denmark, the Netherlands, Sweden, Australia
- In the U.S.: National Centers for Education and Health Statistics, Census Bureau

# Synthetic Microdata

- Has a relatively long history (e.g., U.S. 1990 decennial census, 1989 U.S. Survey of Consumer Finances)
- "Fully synthetic data" – posterior predictive distribution from a model-based data analysis [Rubin, Fienberg]
- "Partially synthetic data" – synthesizing either the sensitive values or the identifiers of sensitive cases [Little]

# Synthetic Microdata, continued

Standards to meet:

- Protect confidentiality at least as well as other methods.
- Provide inferences that are consistent with the inferences an analyst would have made from the original data (analytical validity).

# Synthetic Microdata, continued

Major Census Bureau uses (1):

- Survey of Income and Program Participation linked to longitudinal social security benefit histories and longitudinal employee–employer earnings reports.
  - Cleared for release
  - More tests of analytical validity of value

# Synthetic Microdata, continued

Major Census Bureau uses (2):

- Longitudinally integrated employer–employee data from the Longitudinal Employer–Household Dynamics (LEHD) program in “OnTheMap” Internet application tabulating origin–destination data.

# Synthetic Microdata, continued

Major Census Bureau uses (3,4):

- Longitudinal Business Database
  - “Beta” version released for testing
- American Community Survey
  - Not yet released

# Concluding Comment

- Threats to public microdata release are increasing.
- Statistical agencies must respond to this threat in order to meet the needs of their users.
- Statistical agencies have responded – through licensing, research enclaves, remote access, and synthetic data.