Overview of Statistical Disclosure Methodology for Microdata

Laura Zayatz

Census Bureau

laura.zayatz@census.gov

BTS Confidentiality Seminar Series, April 2003

Microdata

Respondent level data

Each record represents one responding household or person

Usually demographic data

Microdata Disclosure Risk

High visibility records

Linkable external files

Addressing Risk

Can we measure it? (file level – function of ever-changing external database environment)

Getting specific (record level)

Reduce the amount of information

Distort the information

Making the Data Safe

Remove obvious identifiers

Set a geographic threshold

Know external data

Know your data

Look for problems

Look for Problems: External Files

Get to know external files

- State and local government agencies
- Other federal agencies
- Private firms ($)

Perform reidentification experiments

- Hunt and peck
- Data fusion

Look for Problems: Your File

Special uniques

- 2 way cross tabulations of all variables
- Work your way up

Variables that you know exist on external databases (even if you cannot access them)

Be Careful with ……

Geographic detail

Contextual variables/Sampling information

Establishment data

Longitudinal data (life events often generate records on external files)

Administrative data

Data that will also be published in the form of tables at smaller geographic levels

When You Find a Problem…..

Reduce risk by reducing the amount of information

Reduce risk by perturbing the data

Reducing the Amount of Information

Delete a variable

Recode a categorical variable into larger categories (perhaps using thresholds)

Recode a continuous variable into categories

Round continuous variables

Use top and bottom codes (provide means,…)

Use local suppression

Enlarge geographic areas

Perturbing the Data

Noise addition

Swapping

Rank Swapping

Blanking and imputation

Microaggregation

Multiple imputation/modeling to generate synthetic data

Census 2000 Public Use Microdata Samples (PUMS) --- Decrease in Detail

Internal reidentification study looking at all products (microdata and tables) from 1990 census

Increased concern about external data linking capabilities

The Files

17% of households receive the long form

Maximum of 6% of the population appears on PUMS

2 mutually exclusive files

5% state file, PUMAs contain 100,000 people, less variable detail, 2003 release

1% characteristics file, SuperPUMAs contain 400,000 people, same variable detail, 2003 release

Changes for the 5% and the 1% Files

Round all dollar amounts

- $1-7 = $4
- $8-$999 round to nearest $10
- $1,000-$49,000 round to nearest $100
- $50,000+ round to nearest $1,000

Changes for the 5% and the 1% Files

Round departure time

- 2400-0259 in 30-minute intervals
- 0300-0459 in 10-minute intervals
- 0500-1059 in 5 minute intervals
- 1100-2359 in 10-minute intervals

Changes for the 5% and the 1% Files

Noise added to ages of people in households with 10 or more people

Ages must stay within certain groupings

Blank original ages

New ages generated from a given distribution of ages in that grouping

Small Amount of Additional Noise

Certain characteristics of small, unusual subgroups of people and housing units

Vulnerable to disclosure via publicly available datasets

Not disclosing the details

Resulted from reidentification studies

Changes for the 5% File Only

Categories Must Have 10,000 People Nationwide

	1990	2000
Language	305	74
Ancestry	292	143
Birthplace	312	167
Tribe	27	23
Hispanic O.	48	29
Occupation	506	443

Multiple Races

White

Black

American Indian or Alaska Native

Asian: Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian

Native Hawaiian or Pacific Islander: Native Hawaiian, Guamanian or Chamorro, Samoan, Other Pacific Islander

Some Other Race

Race on the PUMS

Yes/No for each major race grouping (6)

Designation of 1 of about 70 single race groups (includes some write-ins)

Designation of multiple race groups

At least 10,000 (5% file) and 8,000 (1% file) people nationwide

Data Swapping

5% and 1% files

Swapping of “special uniques” --- high risk records

Pairs of households that agree on certain demographic characteristics but are in difference geographic areas are swapped across PUMAs and SuperPUMAs

Issues that Have Not Changed

Topcoding (half-percent/three-percent rule)

Property taxes (categorization)

Conclusion

Know your data

Know external data

Proactively look for problems

If possible, if you find a disclosure problem, let users help you choose the best method for reducing risk of disclosure

Thank You for Coming

laura.zayatz@census.gov

301-457-4955