SynLBD: Providing firm characteristics on synthetic establishment data

Saki Kinney, NISS; Jerry Reiter, Duke University
Javier Miranda & Arnie Reznek, U.S. Census Bureau

28th August 2013
WSC/ISI Hong Kong

Excerpt

Phase 1 supported by NSF Grant ITR-0427889. A portion of this work was conducted by Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Census Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau. Results have been screened to ensure that no confidential data are revealed.

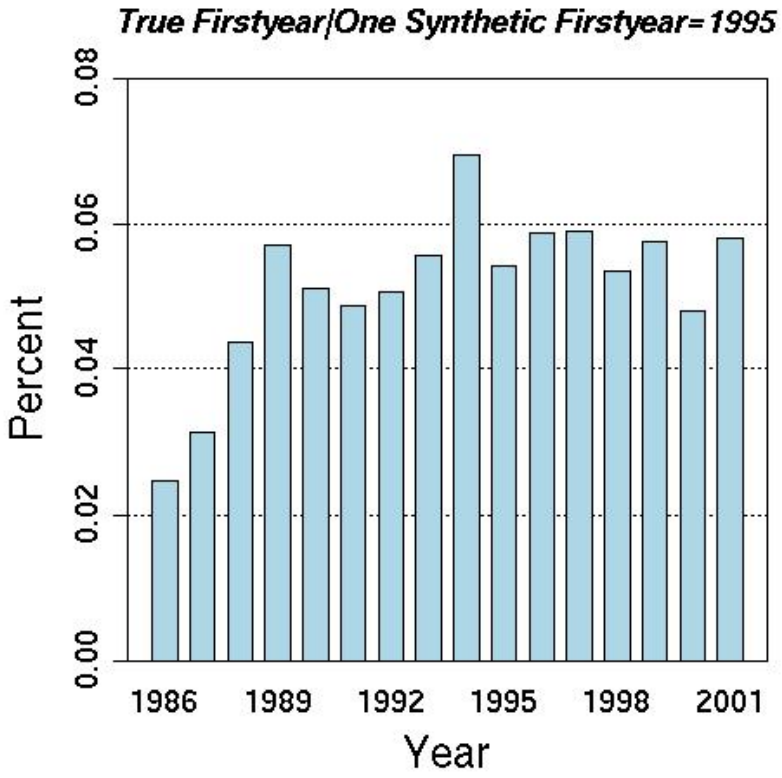
NISS | The Statistics Community Serving the Nation

Confidentiality Protection: Synthesizing Birth and Death Years

- High probability that an individual establishment's synthesized plant birth/death year is different from its actual birth/death year

- Between implicates establishment likely to have different birth, death, lifetime

Example: Synthesizing Synthetic First Year



Confidentiality Protection: Breaking Links Across Implicates

- ID variable in synthetic data generated randomly
- Can’t group (across implicates within year) observations generated from same establishment

Confidentiality Protection: Synthesizing Employment and Payroll*

- Synthesized employment and payroll vary substantially around regression lines
- Correlations between synthetic employment/payroll and observed synthetic payroll are low
  - 'Live' establishments in observed data may be 'dead' in synthetic data

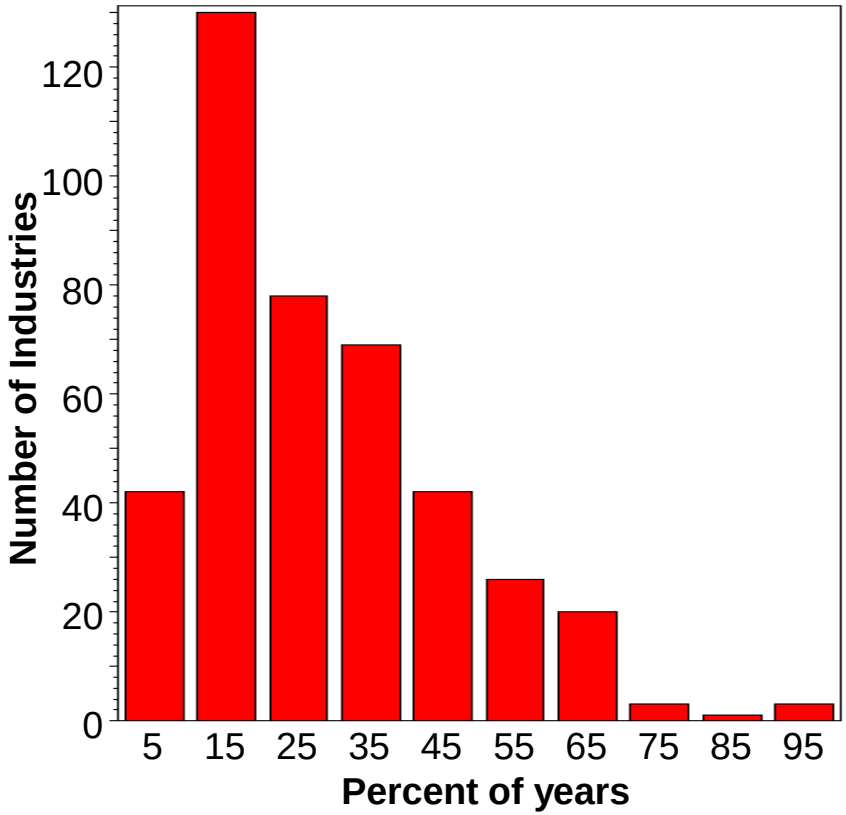
*Results from Phase 1

Confidentiality Protection: Protecting Isolated observations

- We believe the most at-risk establishments are the ones isolated in the distributions of employment and payroll –in the upper tail
- Synthesizer produces observations in upper tail but not necessarily close to real ones

Difference between maximum actual and synthetic employment

Percent of years where distance <5%, by industry



Looking forward

- Phase 2 (underway)
  - Demonstrate confidentiality protection
    - Safely include multiple implicates, geography, firm implicates
    - How will released synthetic longitudinal data be updated with new years of data.