The Synthetic Longitudinal Business Database

Saki Kinney, NISS; Jerry Reiter, Duke University
Ron Jarmin, Javier Miranda, Arnold Reznek, U.S. Census Bureau
John Abowd, Cornell University

July 31, 2009
Census-NSF-IRS Synthetic Data Workshop

This work 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.

Overview

- LBD background
- Synthetic data generation
- Analytic validity
- Confidentiality protection
- Future plans

Longitudinal Business Database(LBD)

- Developed as a research dataset by the U.S. Census Bureau Center for Economic Studies
  - Constructed by linking annual snapshot of the Census Bureau's Business Register
  - CES constructed longitudinal linkages, re-timed multi-unit births and dealt with missing data

The ("Real") LBD

- Economic census covering nearly all private non-farm business establishments with paid employees
  - Contains: Annual payroll and Mar 12 employment (1976-2005), SIC/NAICS, Geography (down to county), Entry year, Exit year, Firm structure
- Used for looking at business dynamics, job flows, market volatility, international comparisons…

Why public release?

- Provide multi-mode access to the LBD
  - Public use tabulations – *Business Dynamics Statistics*
  - "Gold Standard" confidential microdata available through the Research Data Center Network
    - Most used dataset in the RDCs
  - Synthetic public use micro data

Why public release?

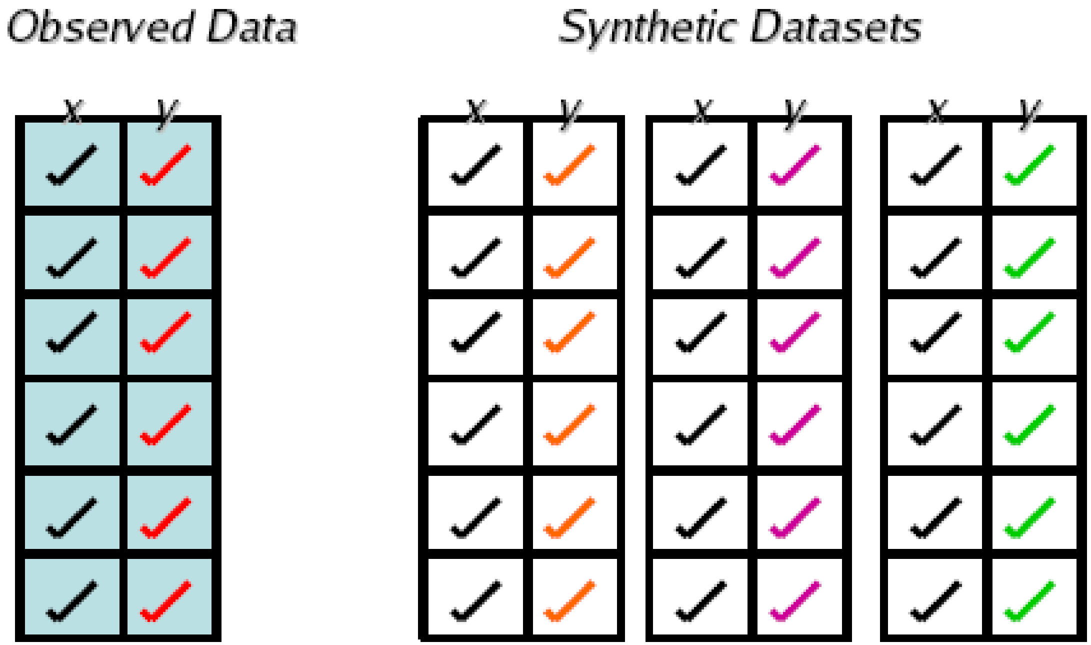
- Provide users with disclosure proofed microdata that permits valid inferences for a subset of uses
  - No need to utilize the RDC Network
  - Reduce the number of requests for special tabulations
  - Aid users requiring RDC access
- Experiment in public use business microdata

Why synthetic data?

- Concerns about confidentiality protection for census of establishments
  - LBD is a test case
- Criteria given for public release:
  - No actual values of confidential values could be released
  - Should provide valid inferences while protecting confidentiality

Partially Synthetic Data

- Y = variable(s) to be synthesized
- X = variable(s) not synthesized



Synthetic LBD

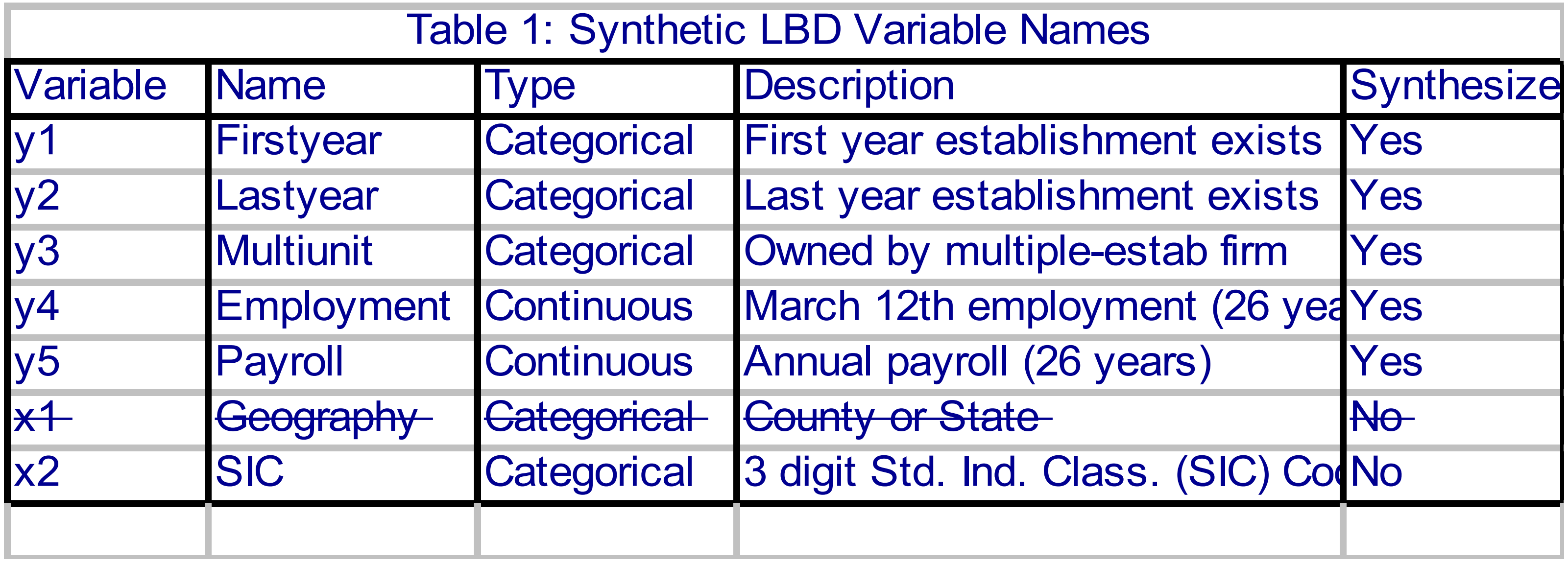


Table 1: Synthetic LBD Variable Names

Variable	Name	Type	Description	Synthesize
y1	Firstyear	Categorical	First year establishment exists	Yes
y2	Lastyear	Categorical	Last year establishment exists	Yes
y3	Multiunit	Categorical	Owned by multiple-estab firm	Yes
y4	Employment	Continuous	March 12th employment (26 yea	Yes
y5	Payroll	Continuous	Annual payroll (26 years)	Yes
~~x1~~	~~Geography~~	~~Categorical~~	~~County or State~~	~~No~~
x2	SIC	Categorical	3 digit Std. Ind. Class. (SIC) Co	No

Notes:

- There is also a randomly generated estab ID number, LBDnum
- Released Synth LBD contains one implicate, excludes geography

Synthesis: General Approach

- Generate joint distribution of Y|X by sampling from conditionals
  - $f(y1,y2,y3|X) = f(y1|X) \cdot f(y2|y1,X) \cdot f(y3|y1,y2,X)$
- Use SIC as “by” group

Synthesis of Synthetic LBD

- Step1: Impute Firstyear | SIC, County
- Step 2: Impute Last year | First Year, State, SIC
- Step 3: Impute Multiunit | Last Year, First Year, SIC, County)
- Step 4: Impute Emp(t)|Multiunit,Last Year, First Year, SIC, Emp(t-1)
- Step 5: Impute Pay(t)|Emp(t),Multiunit, Last Year, First Year, SIC, Pay(t-1)

General approach to synthesis

- Drawing from $f(y_k|X,y_1,\ldots,y_{k-1})$
  - Fit model using observed data
  - Draw new values of parameters from posterior distributions
  - Use new parameters to predict y_k from X and synthetic values of $y_1,\ldots,y_{k-1}$

First Year

- Impute Firstyear | SIC, County using variant of Dirichlet-Multinomial
  - “Prior” information is obtained by collapsing categories
  - Synthetic values obtained from sampling from multinomial distribution

Last Year

- Impute Last Year| First Year, State, SIC
- Simple multinomial approach
  - Dirichlet-multinomial with flat prior
  - Sample from multinomial probabilities obtained from matching categories in observed data

Multi-unit Status

- Impute in two stages:
  - Categorical response: Always MU, sometimes MU, never MU
  - Imputed using simple multinomial approach
- Given change in status occurs, impute when change occurred (future)

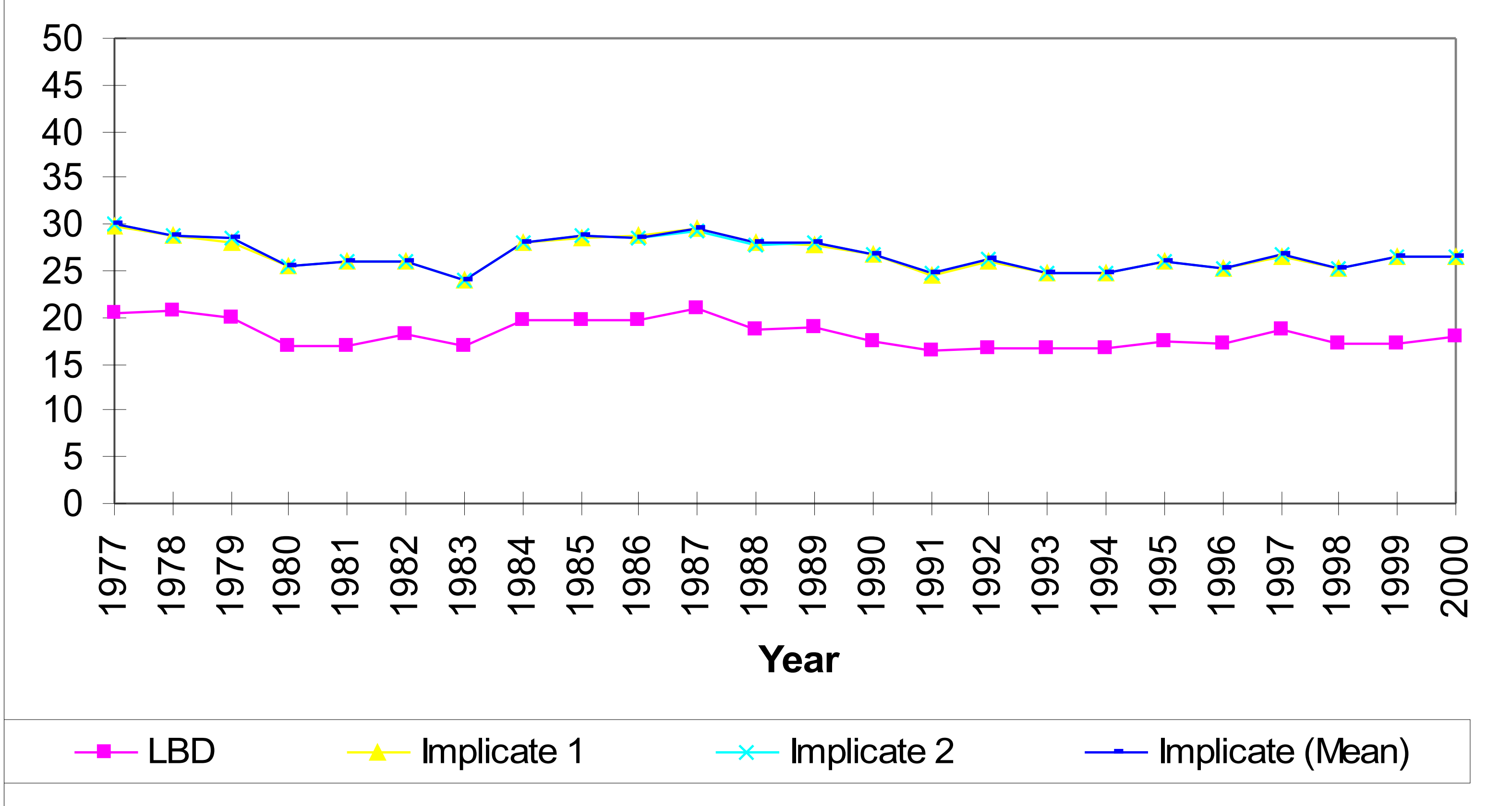
Employment and Payroll

- Highly skewed longitudinal continuous variables
- Imputed using a set of normal linear models with kde transformation of response
- Impute year by year, employment and then payroll

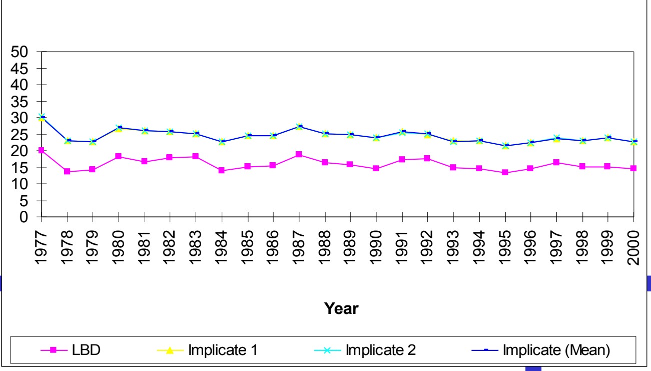
Analytic Validity Tests

- Compare observed data and synthetic data for whole LBD
  - Job creation and destruction
  - Employment volatility
  - Gross employment levels

Job Creation Rates: LBD and Implicates by Year

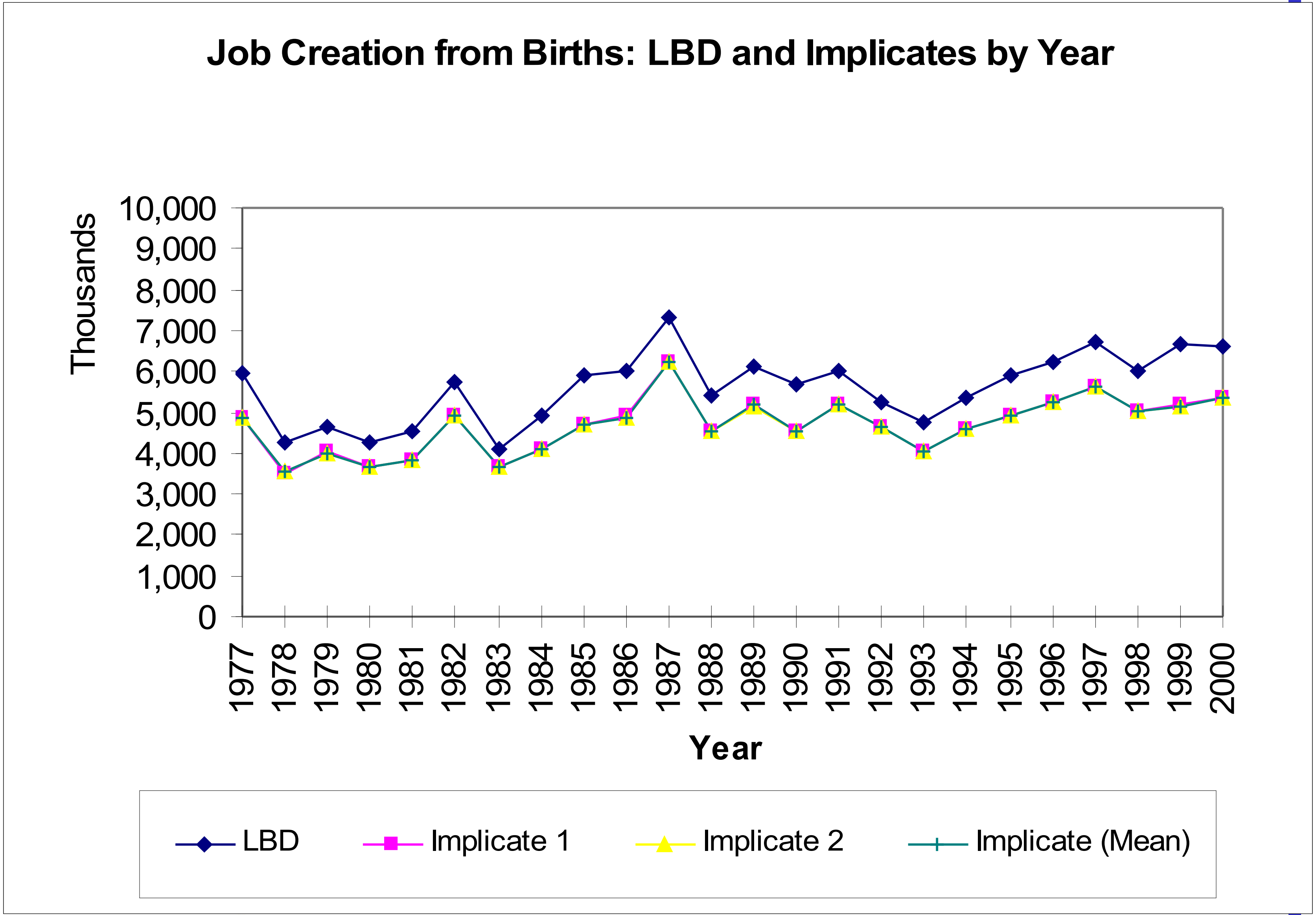


Job Destruction Rates: LBD and Implicates by Year



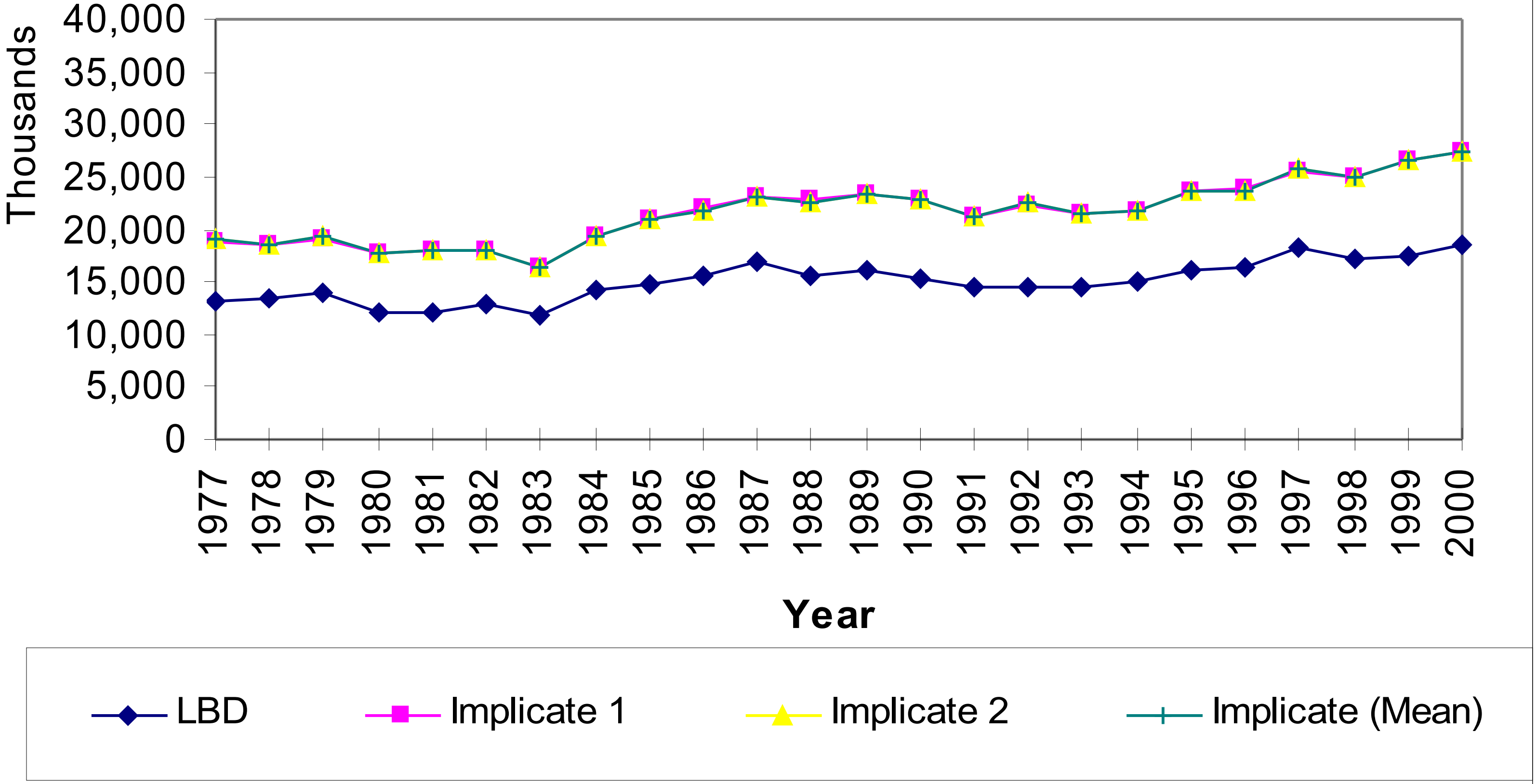
NISS

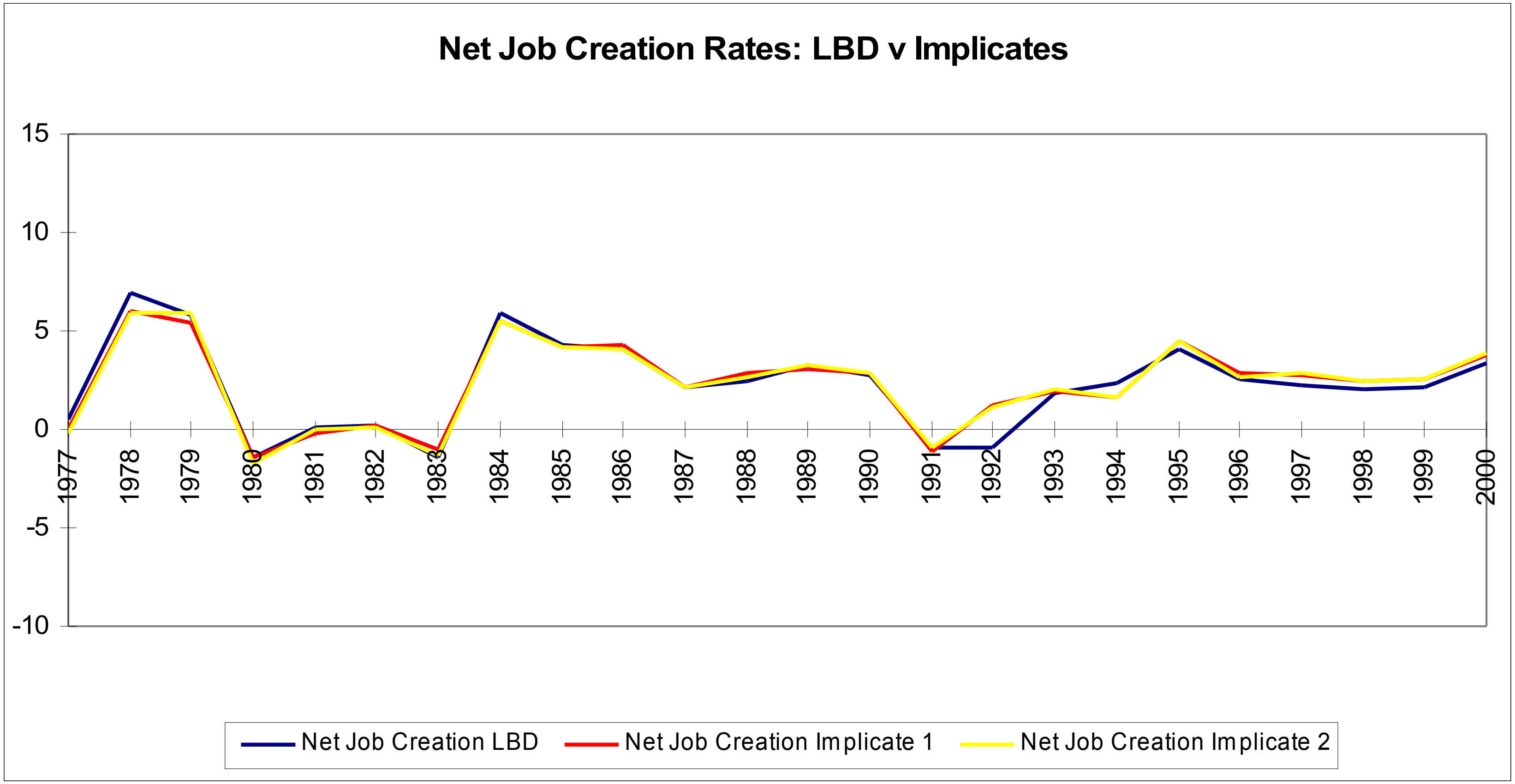
The Statistics Community
Serving the Nation



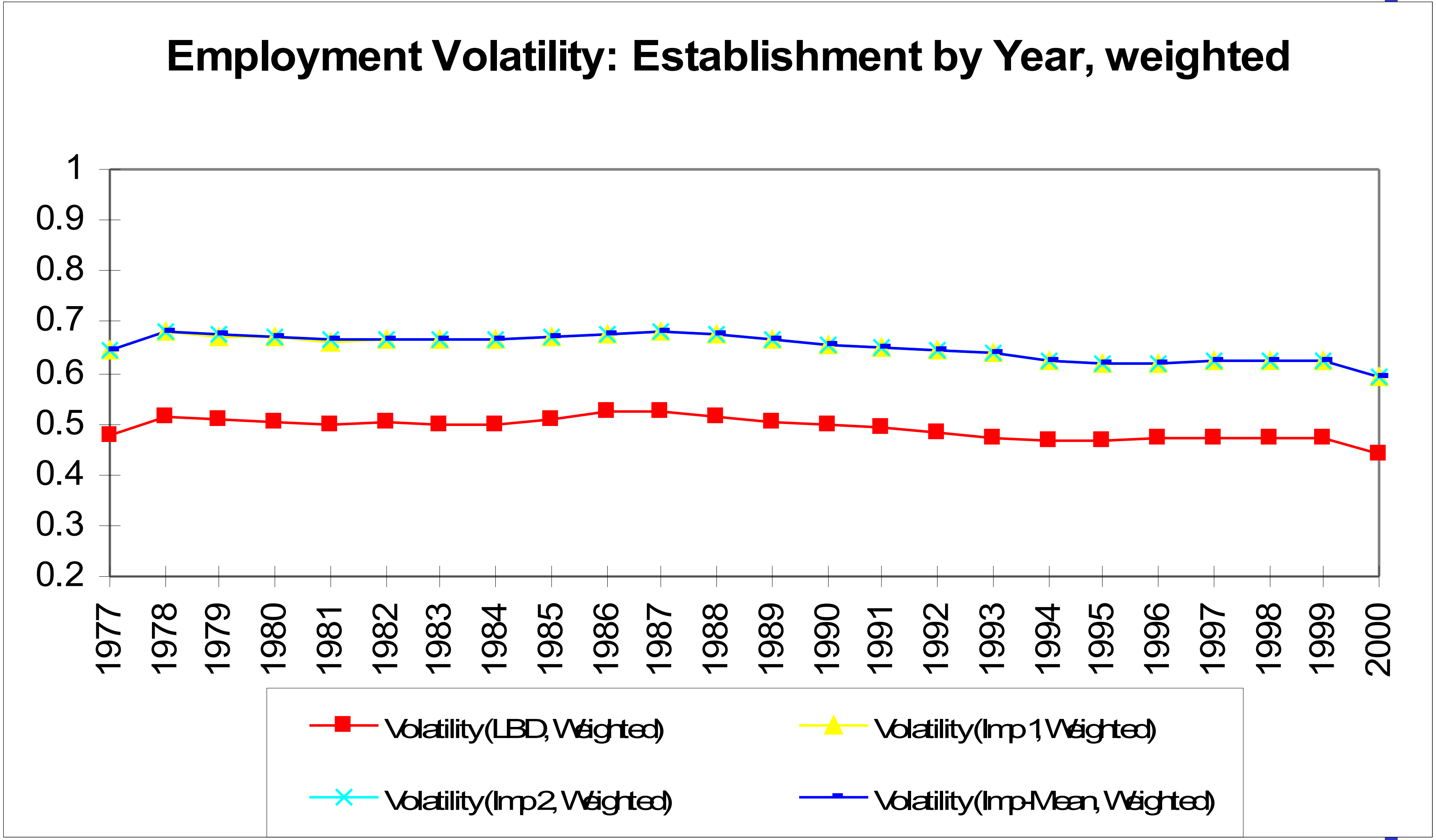
NISS *The Statistics Community Serving the Nation*

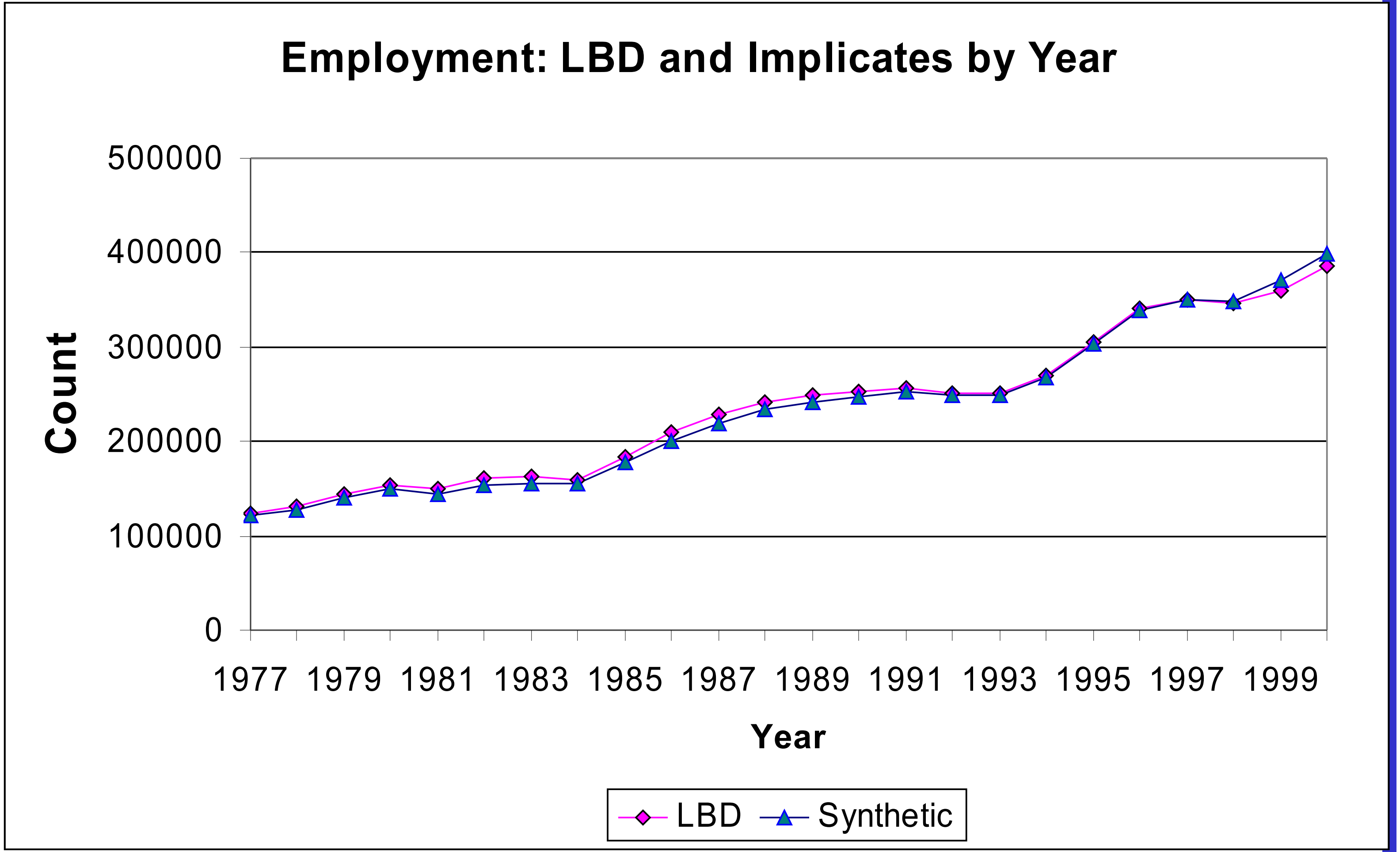
Job Creation from Births and Expansions: LBD and Implicates by Year





NISS *The Statistics Community Serving the Nation*


Employment Volatility: Establishment by Year, weighted
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
Volatility(LBD, Weighted)
Volatility(Imp1, Weighted)
Volatility(Imp2, Weighted)
Volatility(Imp-Mean, Weighted)


Employment: LBD and Implicates by Year
500000
400000
300000
200000
100000
0
Count
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
Year
LBD
Synthetic

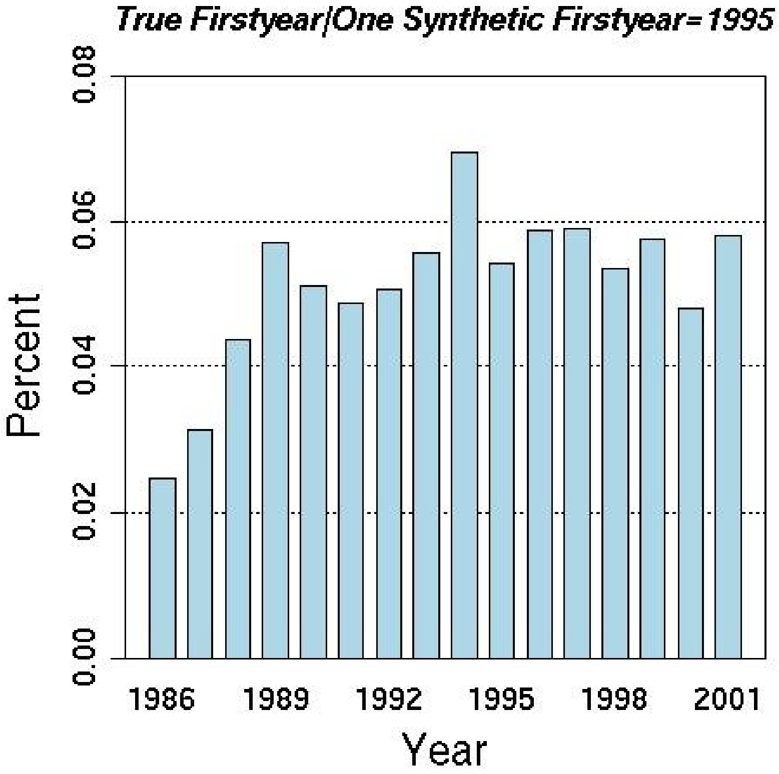
Confidentiality Protection

- Firm structure, firm linkages, geography unavailable in current release
- Several layers of protection from replacing sensitive values of with draws from probability distributions
- Can't link estabs across implicates

Disclosure analysis

- High probability that an individual establishment’s synthetic birth/death year is different from its actual birth/death year
- Synthetic maxima not necessarily near actual
- High between-imputation variability at establishment level
- More in disclosure session (Reznek)

Example: Synthetic First Year



Conclusions and Plans

- Analytical validity supported for broad analyses
  - Obtain user feedback to inform future refinements
- Sufficient confidentiality protection
  - Expected satisfy stringent requirements of differential privacy protection
- Provide training to users on computations from synthetic implicates

Conclusions and Plans (cont.)

- Future Synthetic LBD
  - Include NAICS, geography, changes in multiunit status, firm age & size
  - Multiple Imputations
  - Address bias in job creation/destruction
  - Additional years

Great! Now how do I get it?

- Access to be granted, at least initially, via Cornell Virtual RDC
  - Obtain user account
  - Conduct analyses on VRDC
- Details TBA at vrdc.ciser.cornell.edu