SynLBD: Overview

Synthetic Longitudinal Business Data

International User Seminar

May 9, 2017

Longitudinal Business Database

- Jarmin and Miranda (2002)
- Research database as a longitudinal linkage of establishment records
- Successor to LRD (Longitudinal Research Database), and Longitudinal Establishment Database (LED)
- also available LDB (Longitudinal Database at BLS) and BITS (Business Information Tracking Series, also Census Bureau)
- Key value-added: using name and address matching to fix breaks of longitudinal links
- Coverage: entire economy, employers only

Uses of LBD

- Business dynamics (basis for Business Dynamics Statistics)
- Job flows, employment dynamics at establishment level
- Macro economic models of job dynamics

Structure of LBD

- Annual payroll (cumulative)
- Employment (March 12)
- Geography (county)
- Birth year
- Death year
- Firm structure (who owns the establishment)

Step back: Fundamental structure

- Longitudinal file on an (economic) entity
  - Which has a start and an end date
  - Which has a small number of key attributes evolving over time
- Other hypothetical examples
  - Jobs
  - Residency (in county, in country)
- Importantly:
  - No linkage between entities
  - In graph theoretic terms: only nodes of a network, with associated attributes, no edges

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

Longitudinal Business Database(LBD)

- Originally developed as a research dataset by U.S. Census Bureau Center for Economic Studies
  - Used for looking at business dynamics, job flows, market volatility, international comparisons…

The LBD

- Contains: Annual payroll, March 12 employment, SIC/NAICS, Geography (down to county), Entry year, Exit year, Firm structure
  - Covers all private non-farm business establishments with paid employees, for 1976 through 2010 (updated annually)
  - >30 million establishments

Project Goal

- Generate files for public release containing partially synthetic data
  - Allow researchers to obtain a range of valid inferences
  - Protect against re-identification of units or attributes

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
  - Aid users requiring RDC access
  - Reduce the number of requests for special tabulations
- Experiment in public use business microdata

SynLBD: Variables Used

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
y6	Firm ID	Categorical	Firm links	Yes
~~x1~~	~~Geography~~	~~Categorical~~	~~County or State~~	~~No~~
x2	NAICS	Categorical	3 digit Industry Code	No

Notes:

- There is also a randomly generated estab ID number, LBDnum
- Phase 1 Synth LBD contains one implicate, excludes geography, uses SIC instead of NAICS
- Additional firm variables constructed (firm employment, age, etc)

Why synthetic data?

- Concerns about confidentiality protection for *longitudinal*, *census* of *establishments*
  - Data are more disclosive than cross-sectional samples of people.
  - No actual values of confidential values may be released (i.e., swapping, etc. would provide insufficient protection)

Partially Synthetic Data

- Agency releases X and multiple implicates of Y (but not Y)
  - Y = variable(s) to be synthesized
  - X = variable(s) not synthesized
- Users apply standard data methods to each dataset, with simple combining rules to obtain combined inferences

Synthesis: General Approach

- Generate joint posterior predictive distribution of Y|X
  - $f(y1,y2,y3|X) = f(y1|X) \cdot f(y2|y1,X) \cdot f(y3|y1,y2,X)$
- Use industry (NAICS) as “by” group

Categorical Variable imputation

- Impute Firstyear | NAICS, County using variant of Dirichlet-Multinomial
  - Informative “prior” information is obtained by collapsing categories
- Impute Last Year| First Year, State, SIC
  - Dirichlet-multinomial with flat prior
- Similarly for Multiunit Status

Employment and Payroll

- Highly skewed longitudinal continuous variables
- Impute year by year, employment and then payroll
  - Phase 1: Normal linear models with kde transformation of response (Abowd & Woodcock)
  - Phase 2: CART models w/Bayesian boostrap (Reiter)

CART synthesis method

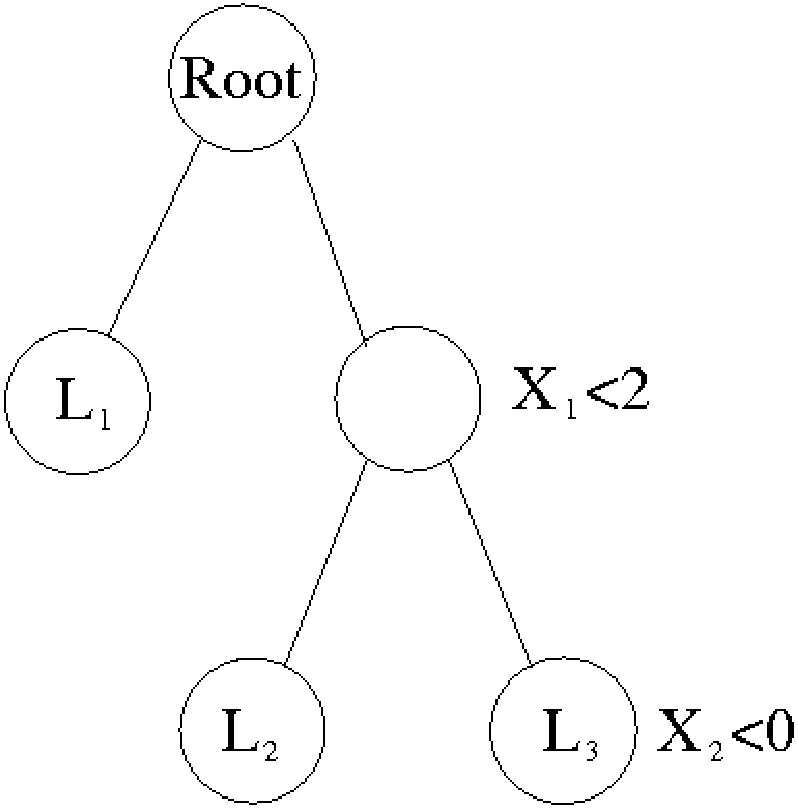
Goal: Synthesize Y | X.

-- Grow maximum tree.

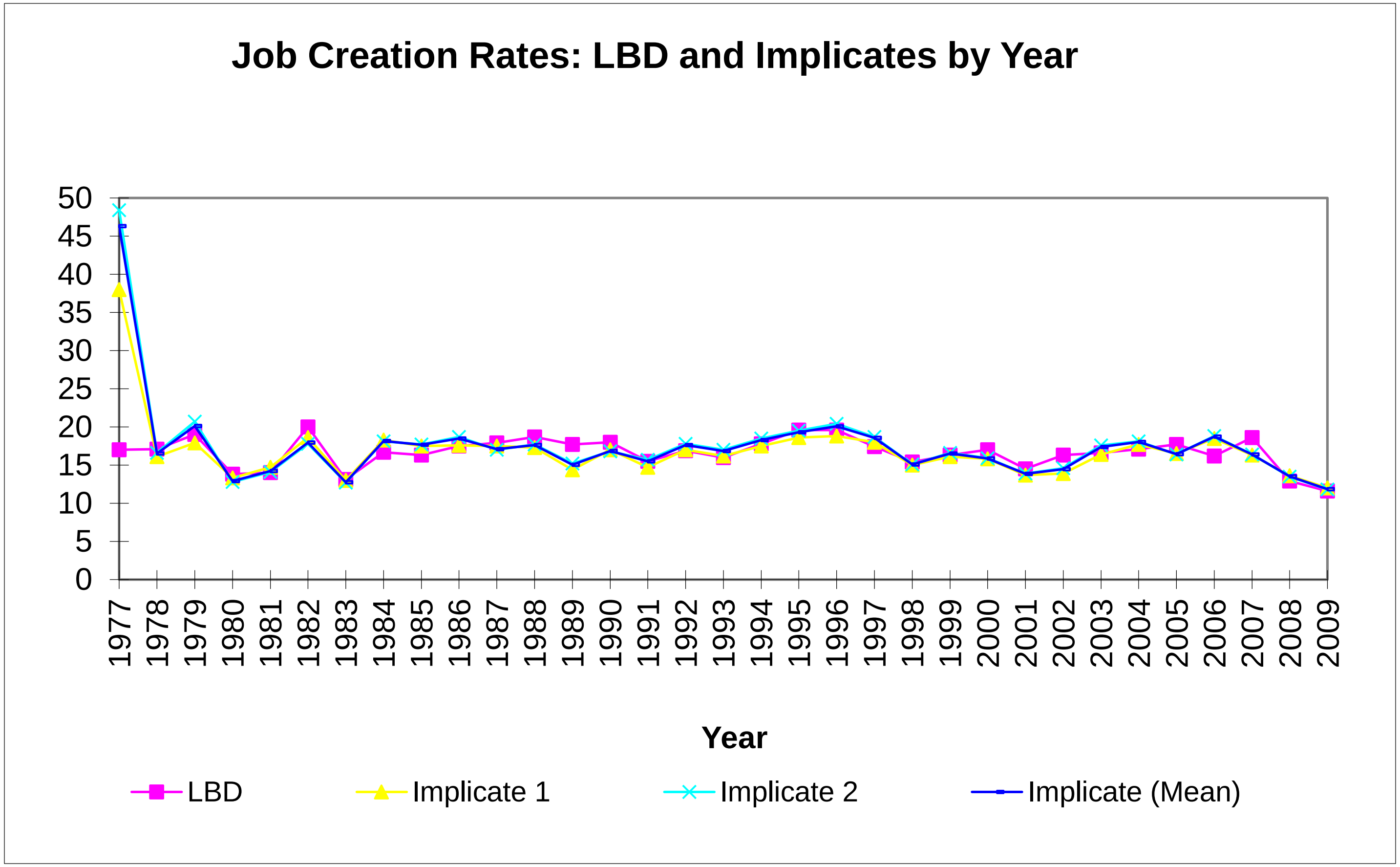
-- Prune for confidentiality.

-- For any X, trace down tree until reach appropriate leaf.

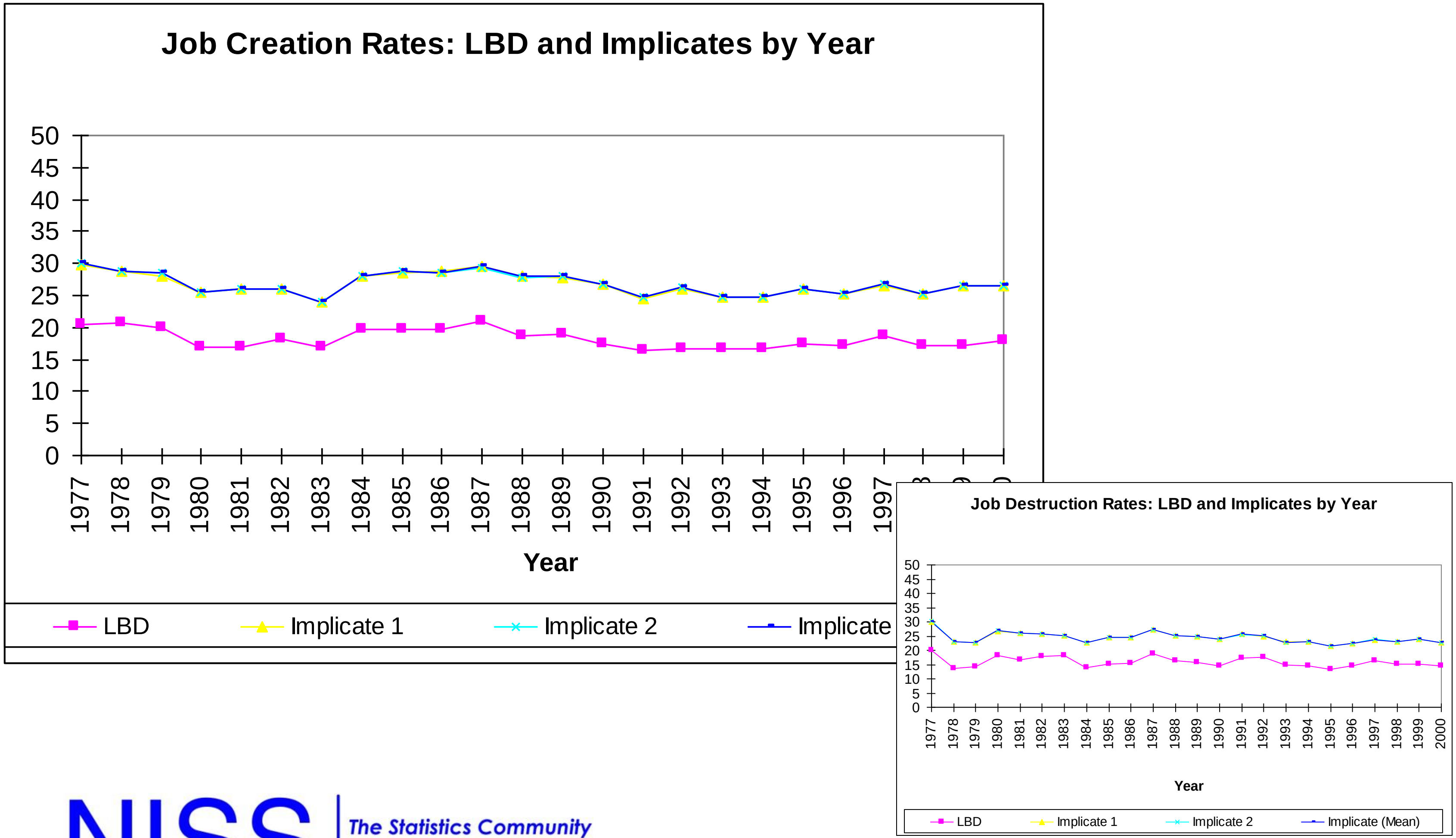
-- Draw Y from Bayes bootstrap followed by smoothed density estimate if needed (with agency-specified bandwidth).



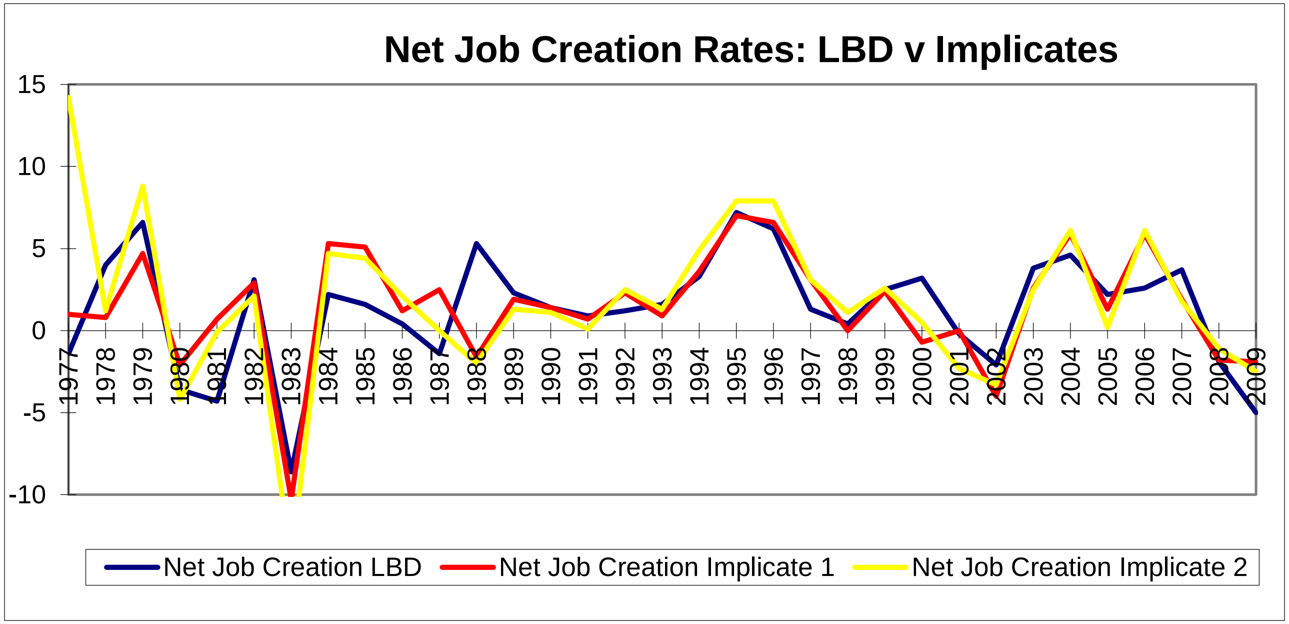
Employment Results



Bias observed in Phase 1

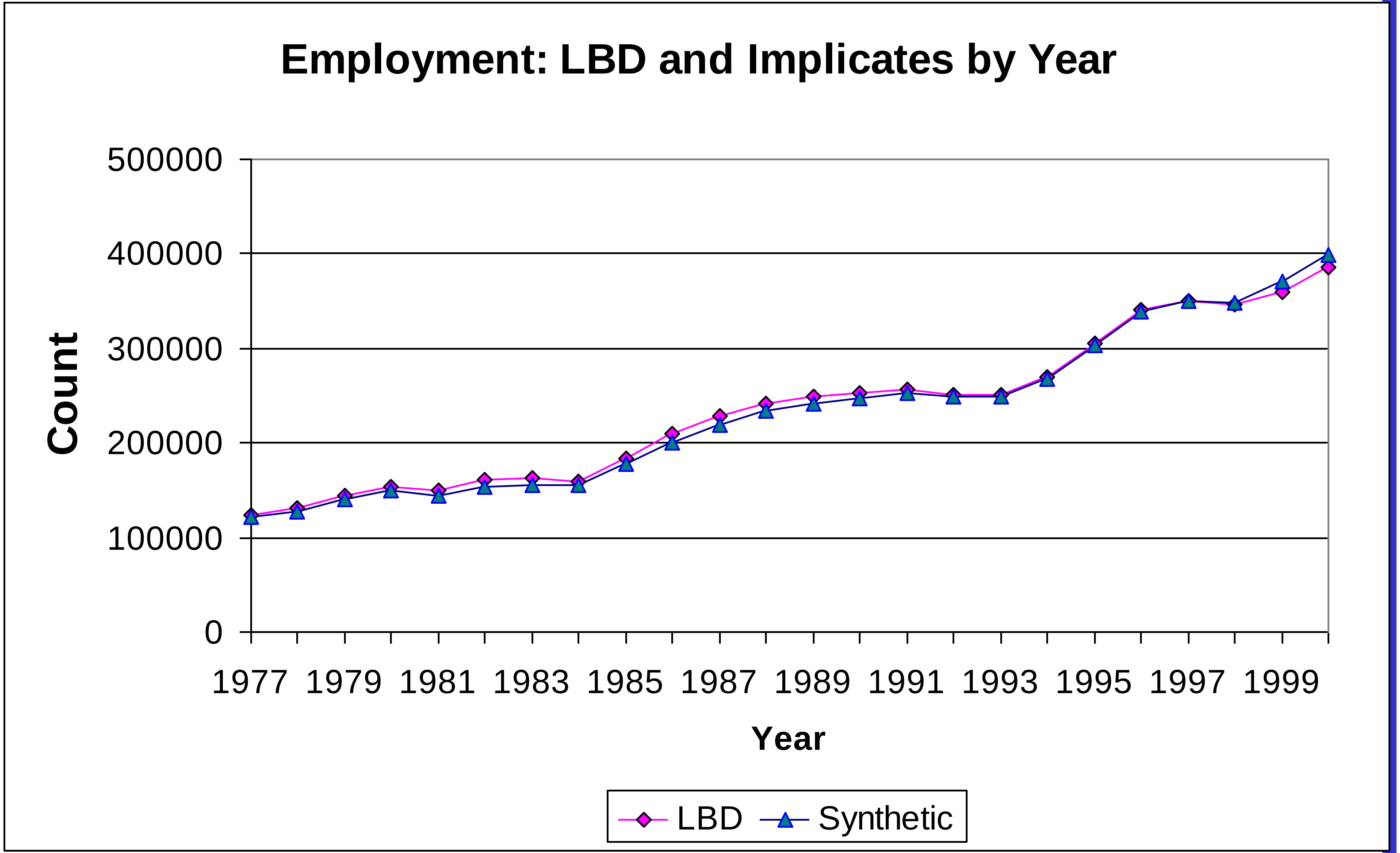


Net Job Creation



Net job creation = Job creation rate – job destruction rate (bias cancels out)

Phase 1 results


Employment: LBD and Implicates by Year
Count
Year
LBD
Synthetic