# The Synthetic Longitudinal Business Database

Based on presentations by Kinney/Reiter/Jarmin/Miranda/Reznek[2]/Abowd

on July 31, 2009 at the

Census-NSF-IRS Synthetic Data Workshop

[link] [link]

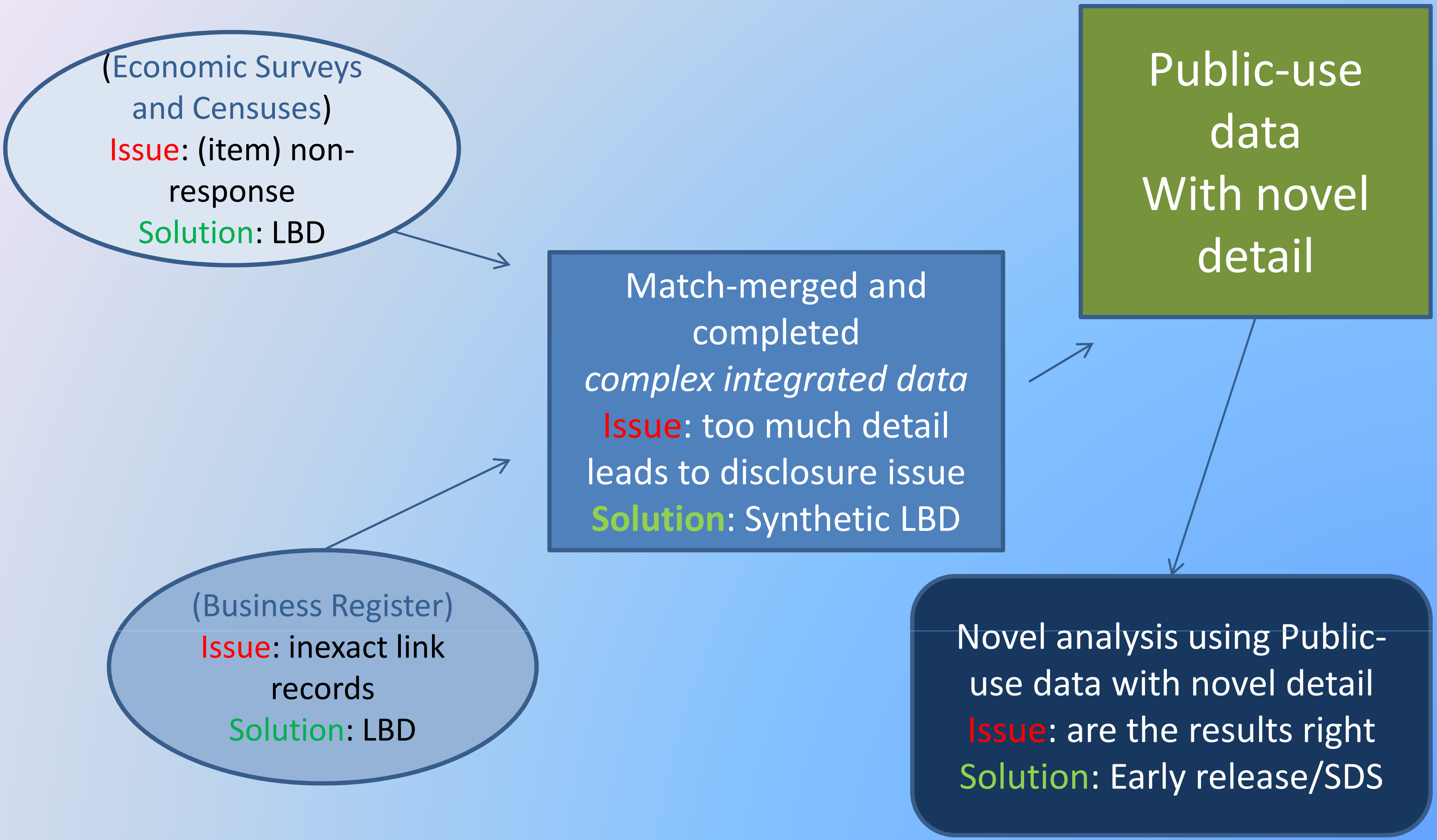
Kinney/Reiter/Jarmin/Miranda/Reznek/Abowd (2011) "Towards Unrestricted Public Use Microdata: The Synthetic Longitudinal Business Database.", CES-WP-11-04

*Work on the Synthetic LBD was supported by NSF Grant ITR-0427889, and ongoing work is supported by the Census Bureau. 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

# Elements



(Economic Surveys and Censuses)
Issue: (item) non-response
Solution: LBD

(Business Register)
Issue: inexact link records
Solution: LBD

Match-merged and completed
*complex integrated data*
Issue: too much detail leads to disclosure issue
**Solution**: Synthetic LBD

Public-use data
With novel detail

Novel analysis using Public-use data with novel detail
Issue: are the results right
Solution: Early release/SDS

# 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…

# Longitudinal Business Database(LBD)

- Detailed description in Jarmin and Miranda
- 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 (see Lecture 4)

# Longitudinal Business Database(LBD)

- CES constructed
  - longitudinal linkages (using probabilistic matching, see Lecture 10),
  - re-timed multi-unit births and
  - dealt with missing data

# Access to LBD data

- Different levels of access
  - Public use tabulations – *Business Dynamics Statistics* *http://www.ces.census.gov/index.php/bds*
  - “Gold Standard” confidential microdata available through the Census Research Data Center Network
    - (LBD in RDC)
    - Most used dataset in the RDCs

# Bridge between the two

- Synthetic data set
  - Available outside the Census RDC
  - Providing as much analytical validity as possible
  - 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

# Generic structure

- Gold standard: given by internal LBD (already completed)
- *Partially* synthetic:
  - Unsynthesized:
    - County (but not released!) [x1]
    - SIC [x2]
  - Synthesized
    - Birth [y1] and death [y2] year:
    - Multi-unit status [y3]
    - Employment (March 12) [y4]
    - Payroll [y5]

# Synthesis: General Approach

- Y=[y1|y2|y3|y4|y5]
- X=[x1|x2]
- Generate joint distribution of Y|X by sampling from conditionals
  - f(y1,y2,y3|X) = f(y1|X)·f(y2|y1,X)·f(y3|y1,y2,X)
- Use SIC as “by” group

# 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 $y_1,\ldots,y_{k-1}$

# SRMI approach

- Calendar:
  - Step1: Impute y1 | X
  - Step 2: Impute y2 | [y1| f(X)]
    - Where f(X) uses state [x1’] instead of county [x1]
- Type of firm
  - Step 3: Impute y3 | [y1|y2|X]
- Characteristics
  - Step 4: Impute y4(t)|[y1|y2|y3|y4(t-1)|x2]
  - Step 5: Impute y5(t)|[y1|y2|y3|y4(t)|y5(t-1)|x2]

# First Year

- Impute y1 (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 y2 (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

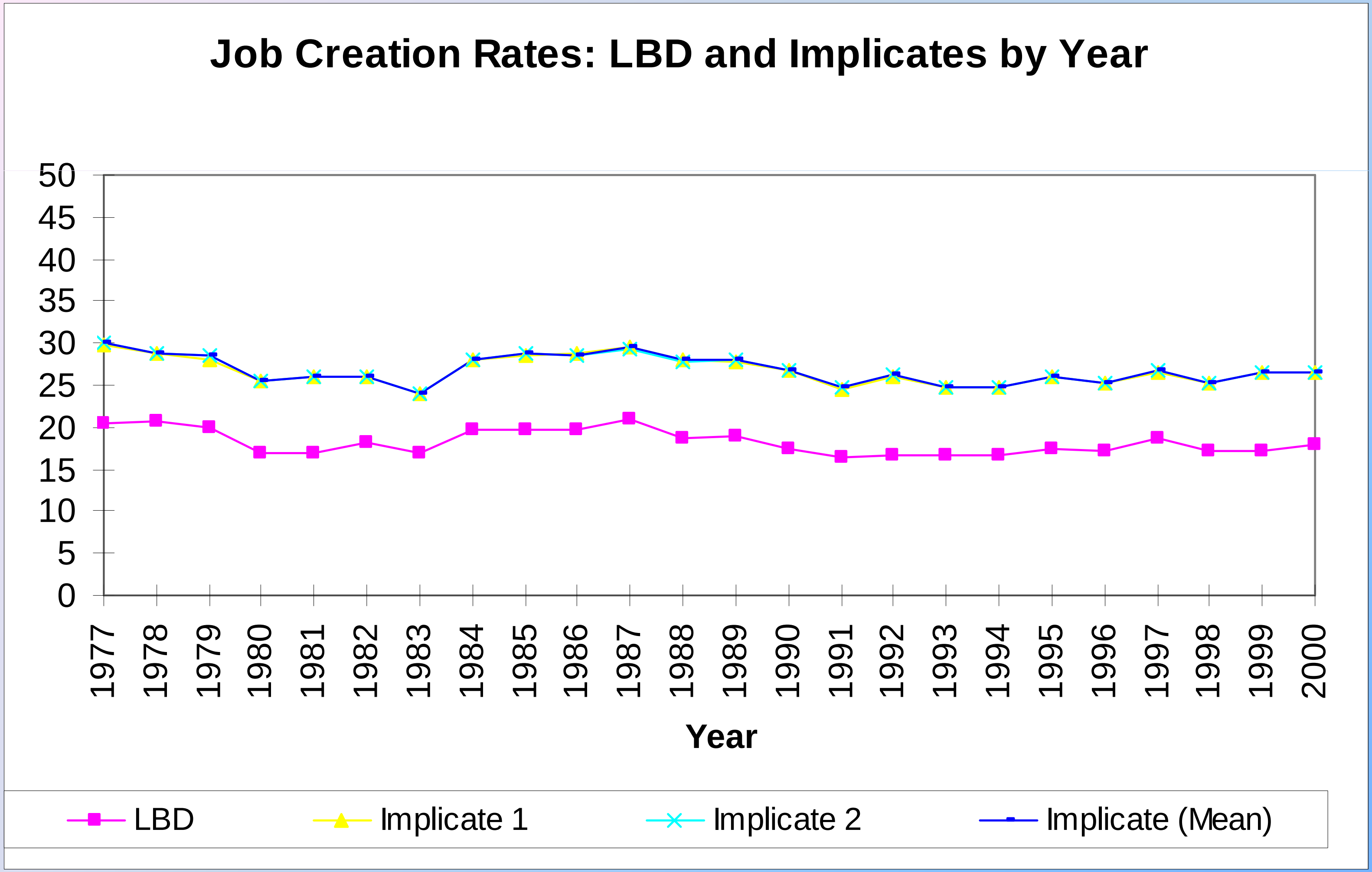
- Impute in two stages:
  - Categorical response: Always MU, sometimes MU, never MU
  - Imputed using simple multinomial approach
- Given change in status occurs, impute when change occurred (future)

# Employment and Payroll

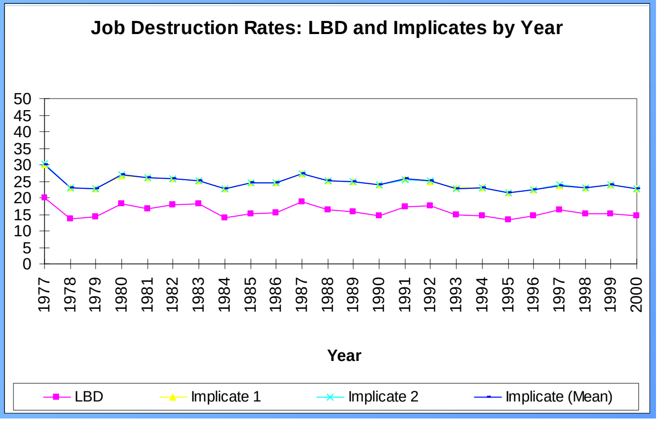
- Highly skewed longitudinal continuous variables
- Imputed using a set of normal linear models with kde transformation of response (Abowd and Woodcock, 2004)
- Impute year by year, employment and then payroll, based on groups
  - (3-digit SIC)
  - by (multiunit status)
  - by (continuer status)
  - by (top 5% status)
- If model too sparse, use 2-digit SIC as prior

# Analytic Validity Tests

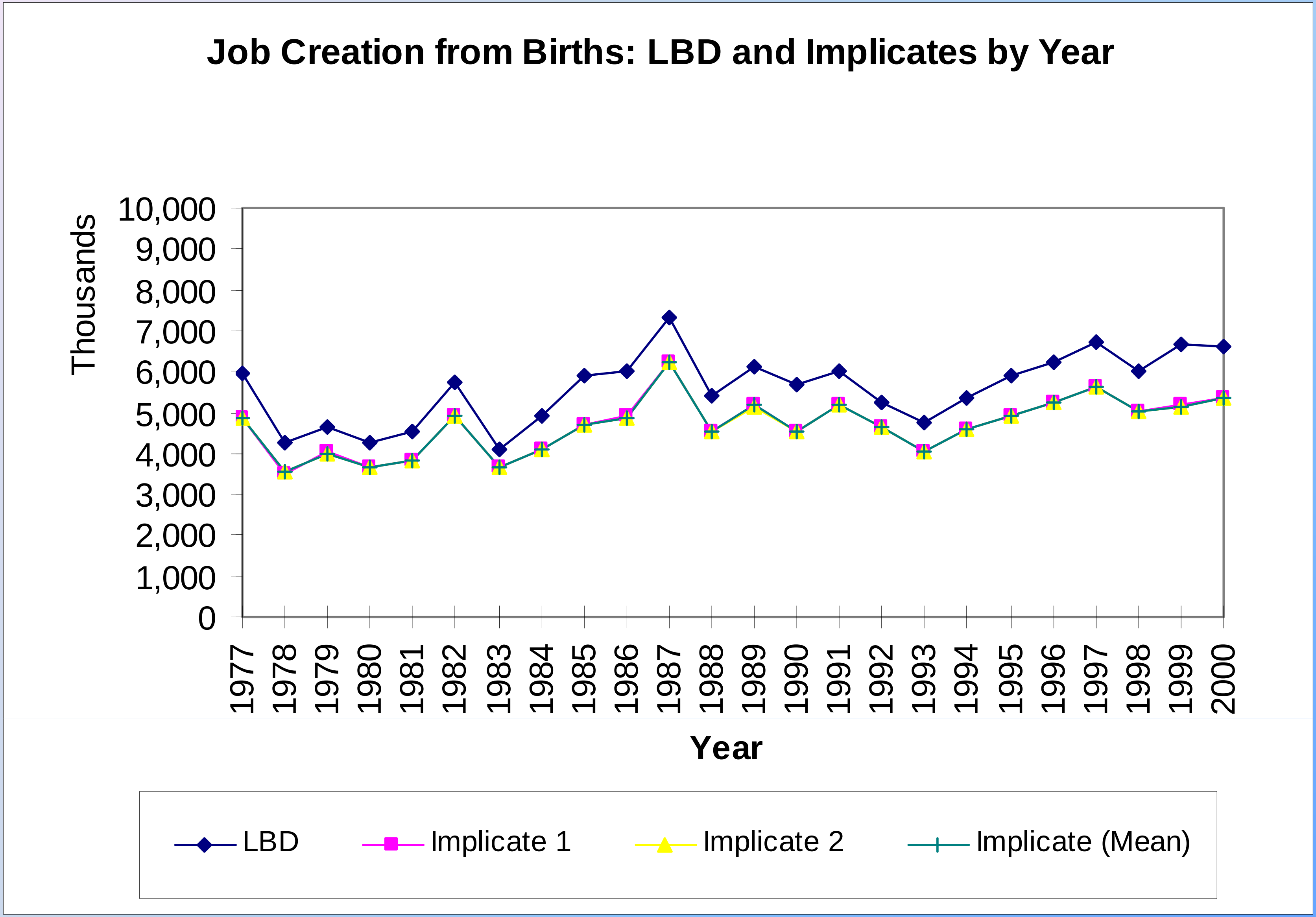
- Compare observed data and synthetic data for whole LBD
  - Job creation and destruction
  - Employment volatility
  - Gross employment levels



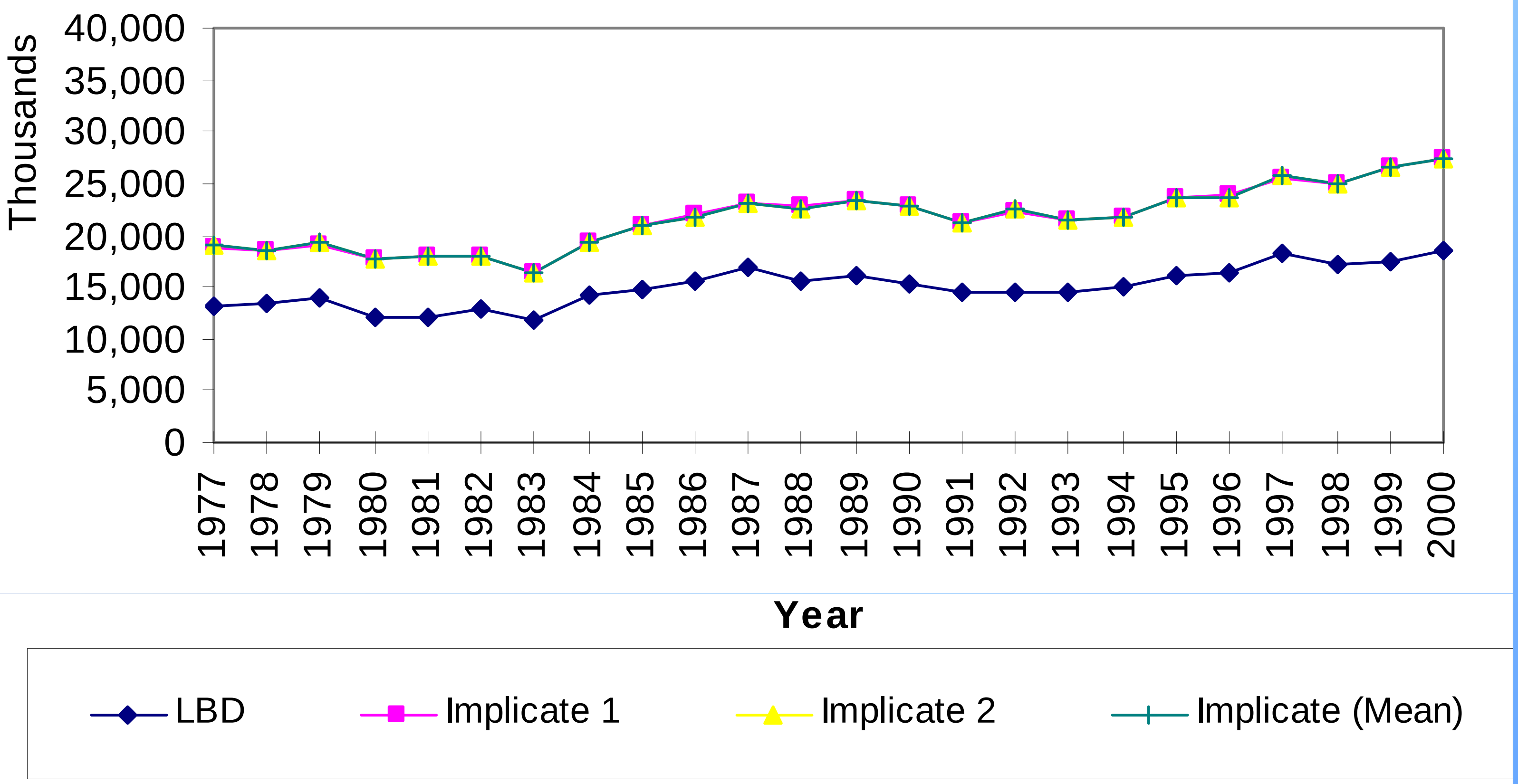
Job Creation Rates: LBD and Implicates by Year

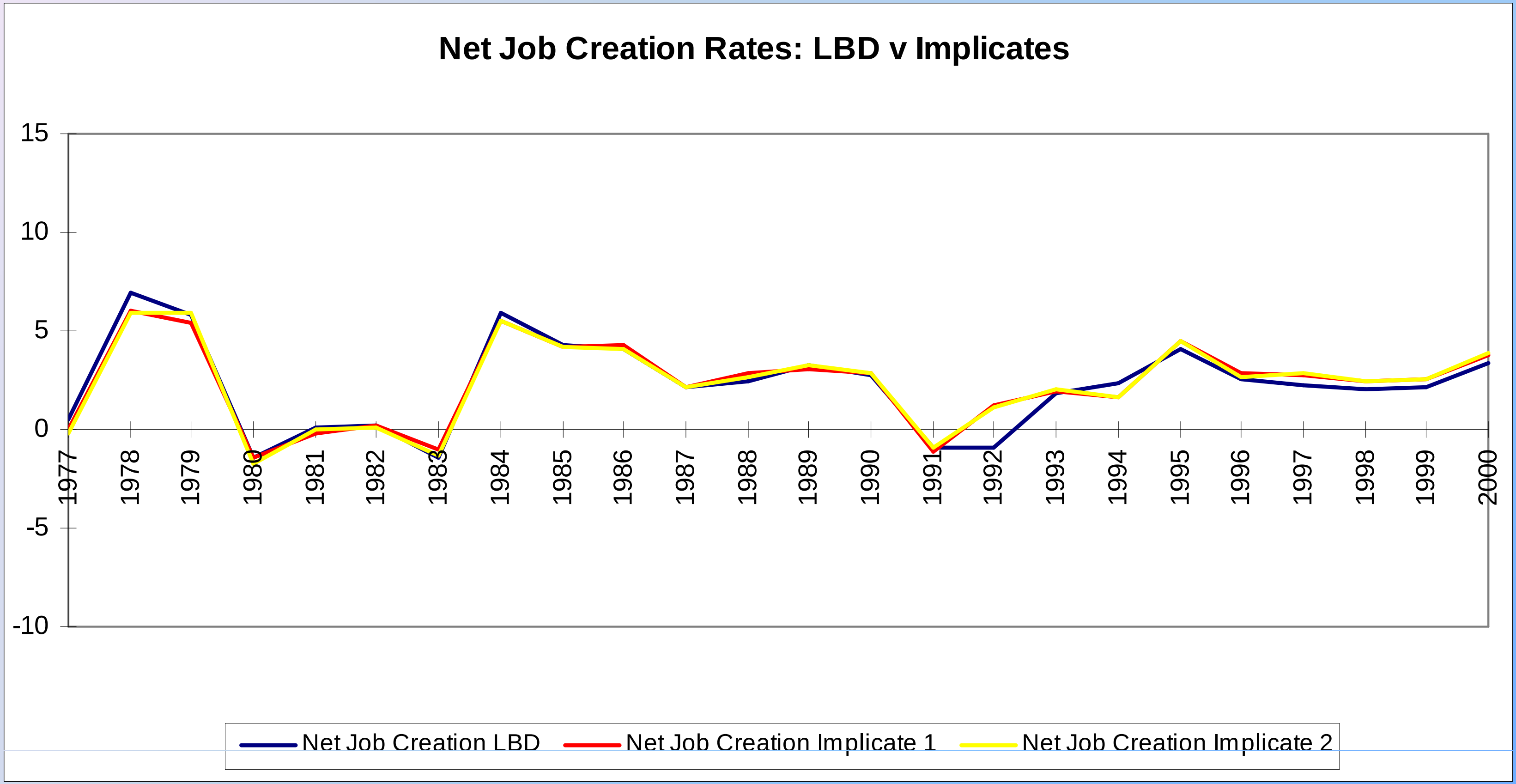


Job Destruction Rates: LBD and Implicates by Year

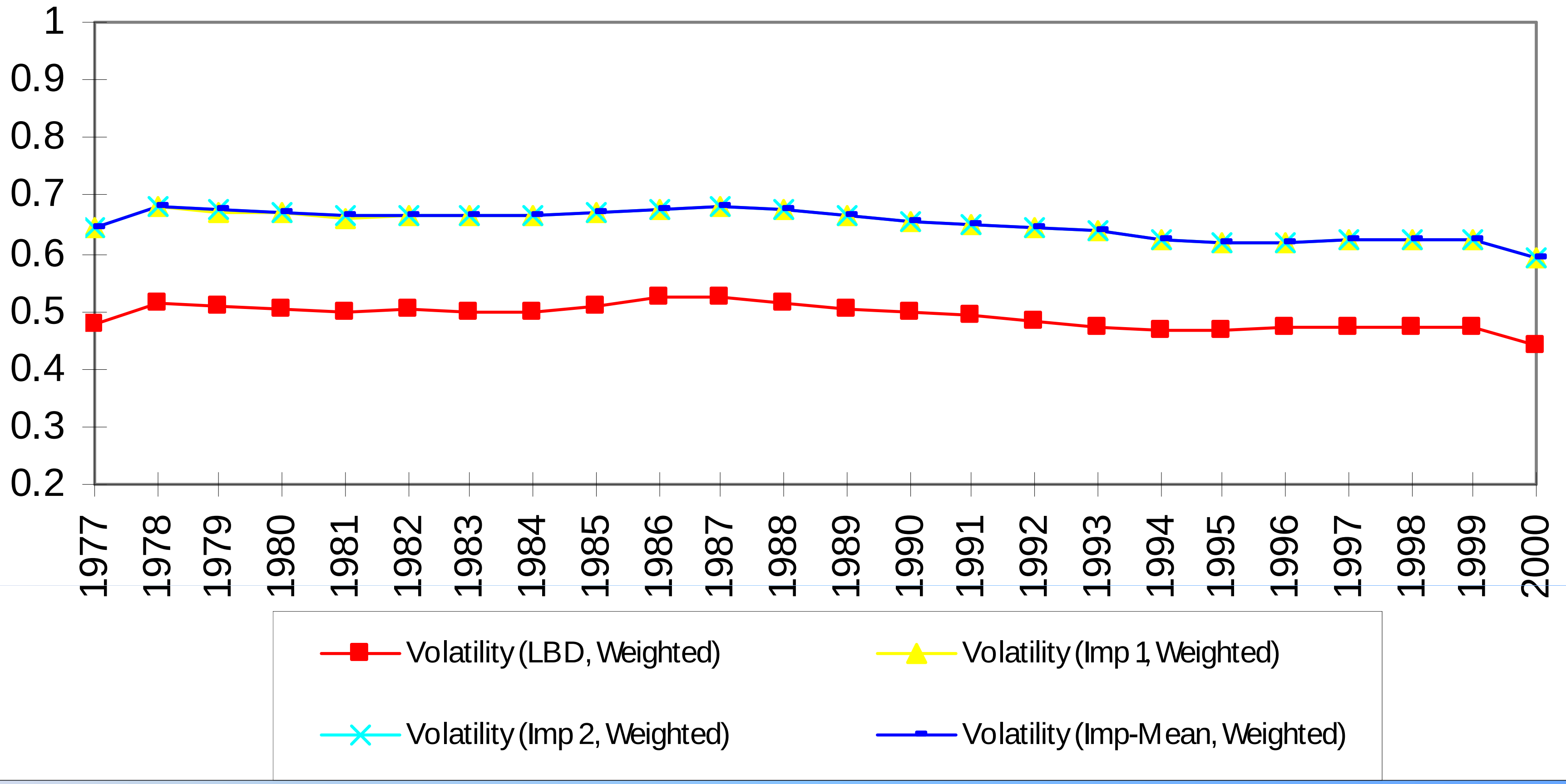

Job Creation from Births: LBD and Implicates by Year
Thousands
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
Year
LBD
Implicate 1
Implicate 2
Implicate (Mean)

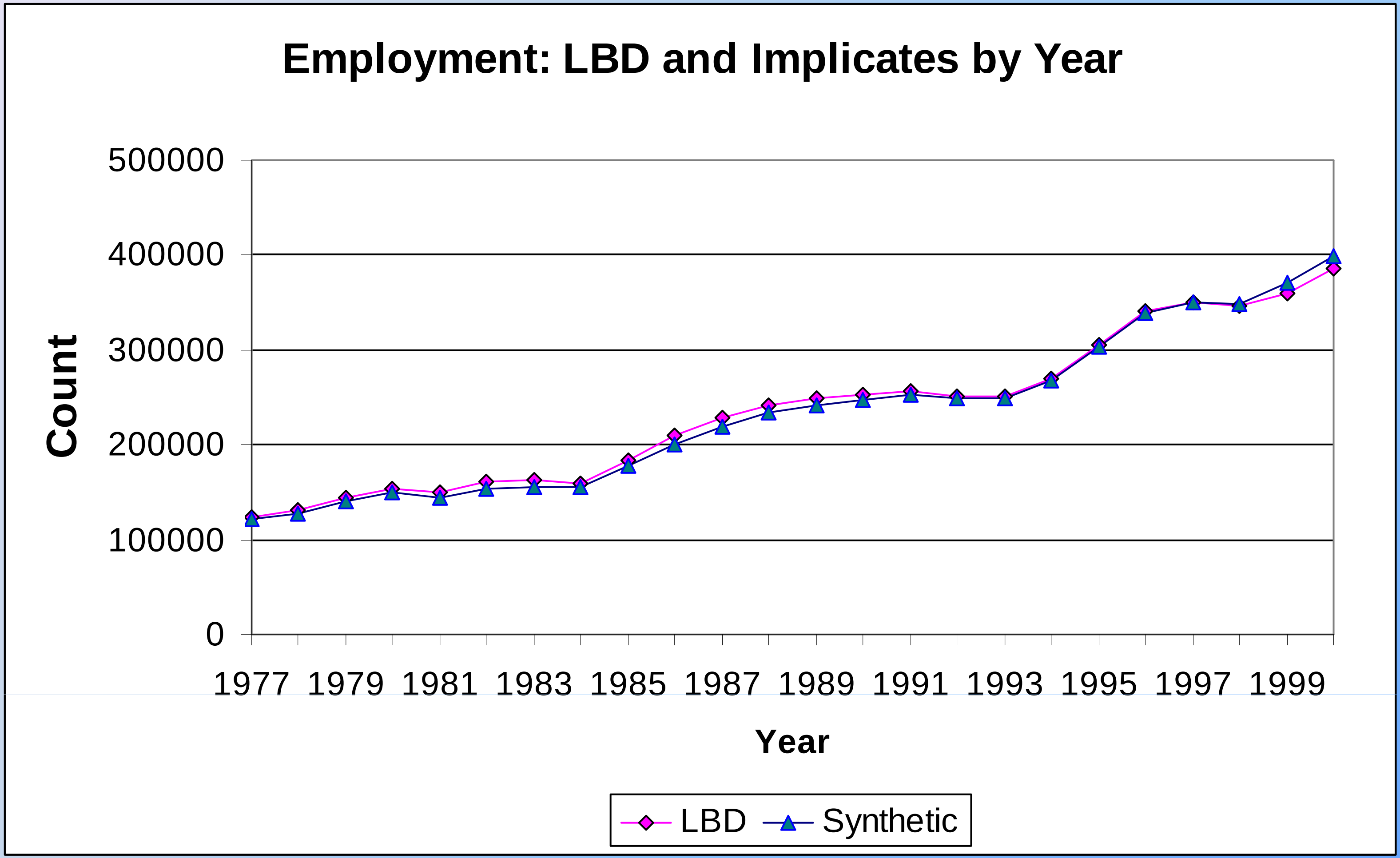
# Job Creation from Births and Expansions: LBD and Implicates by Year




Net Job Creation Rates: LBD v Implicates
15
10
5
0
-5
-10
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
Net Job Creation LBD
Net Job Creation Implicate 1
Net Job Creation Implicate 2

# Employment Volatility: Establishment by Year, 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

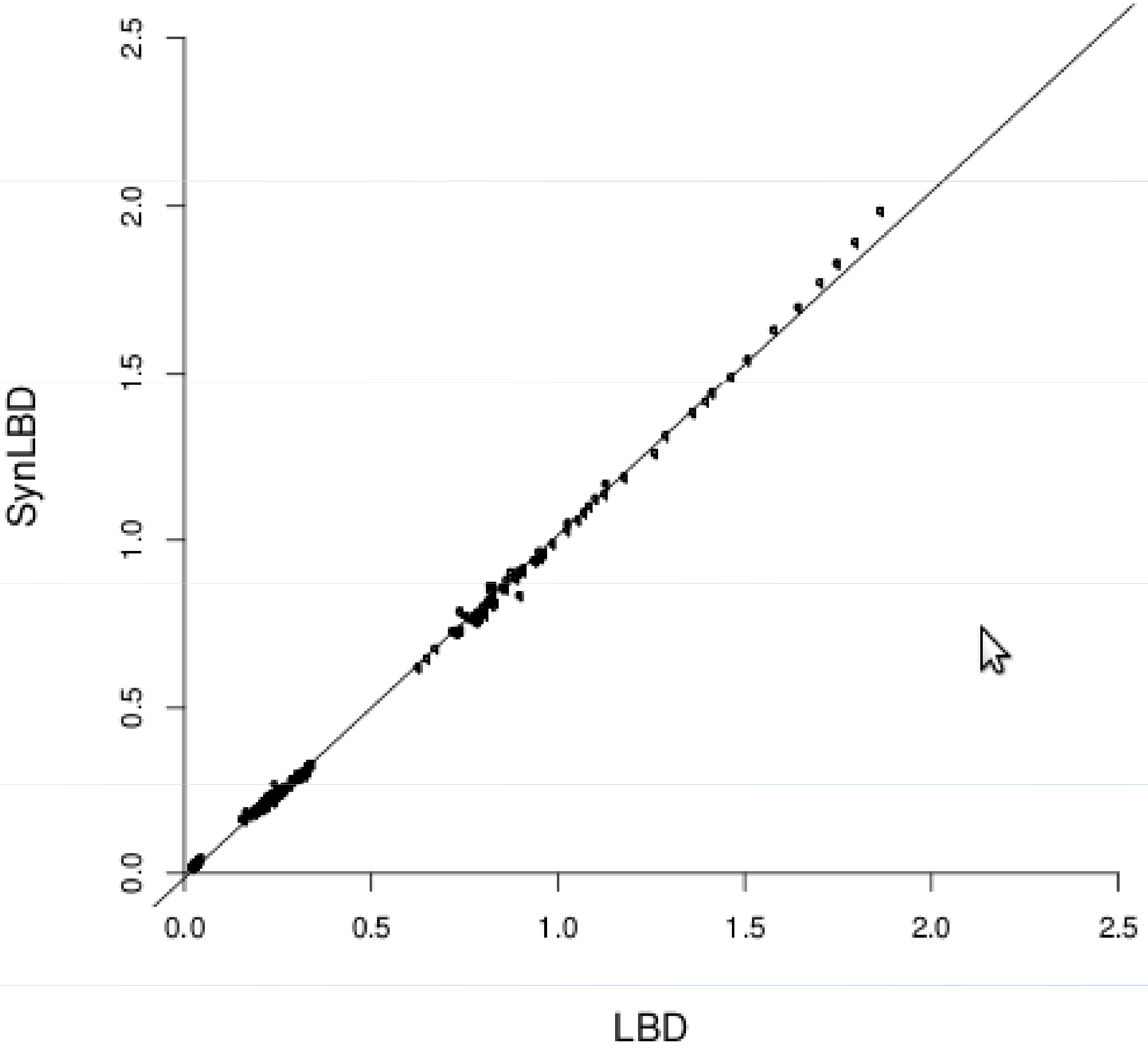


Figure 3: Share of Employment by Industry Sector and Year, 1976-2000

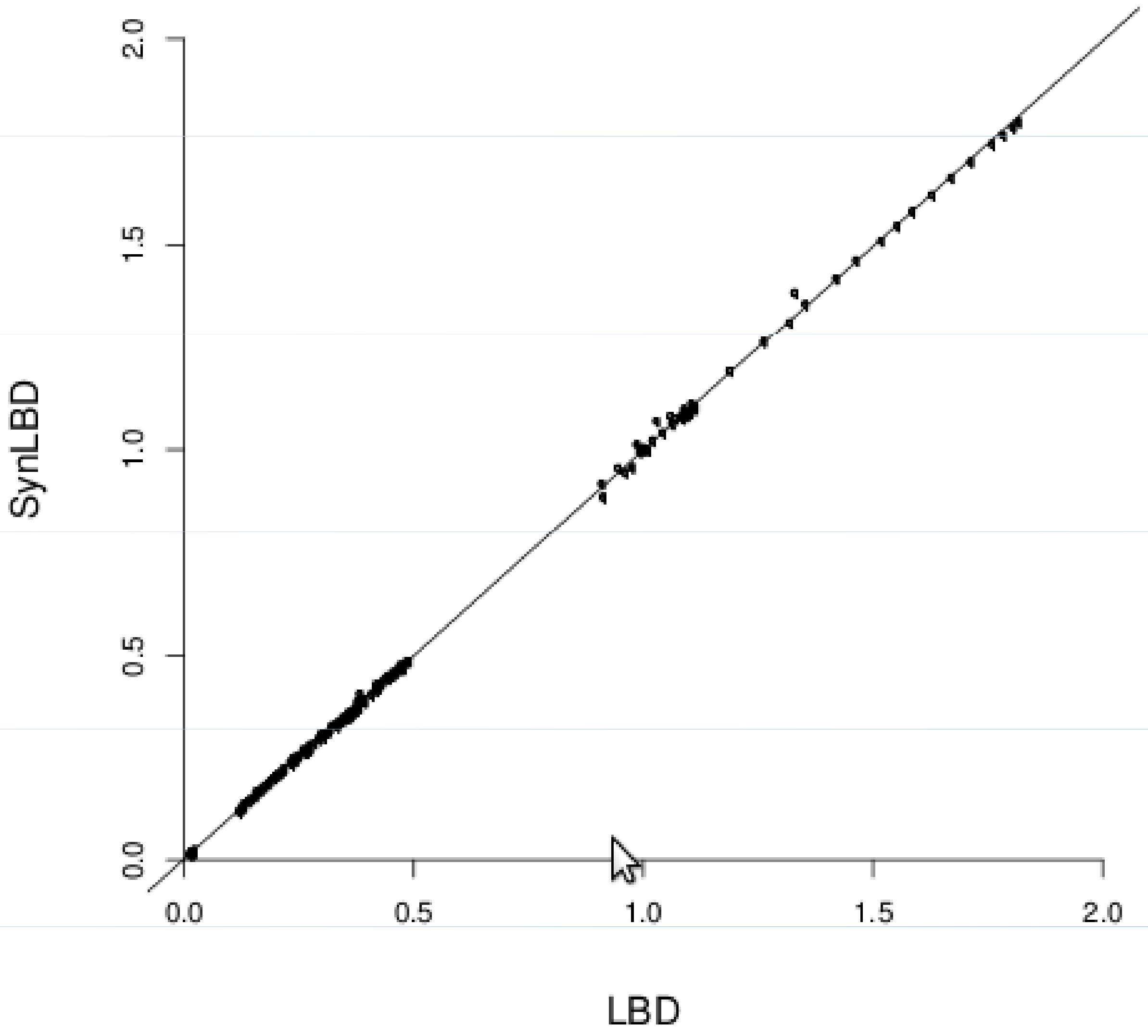


Figure 2: Share of Establishments by Industry Sector and Year, 1976-2000.

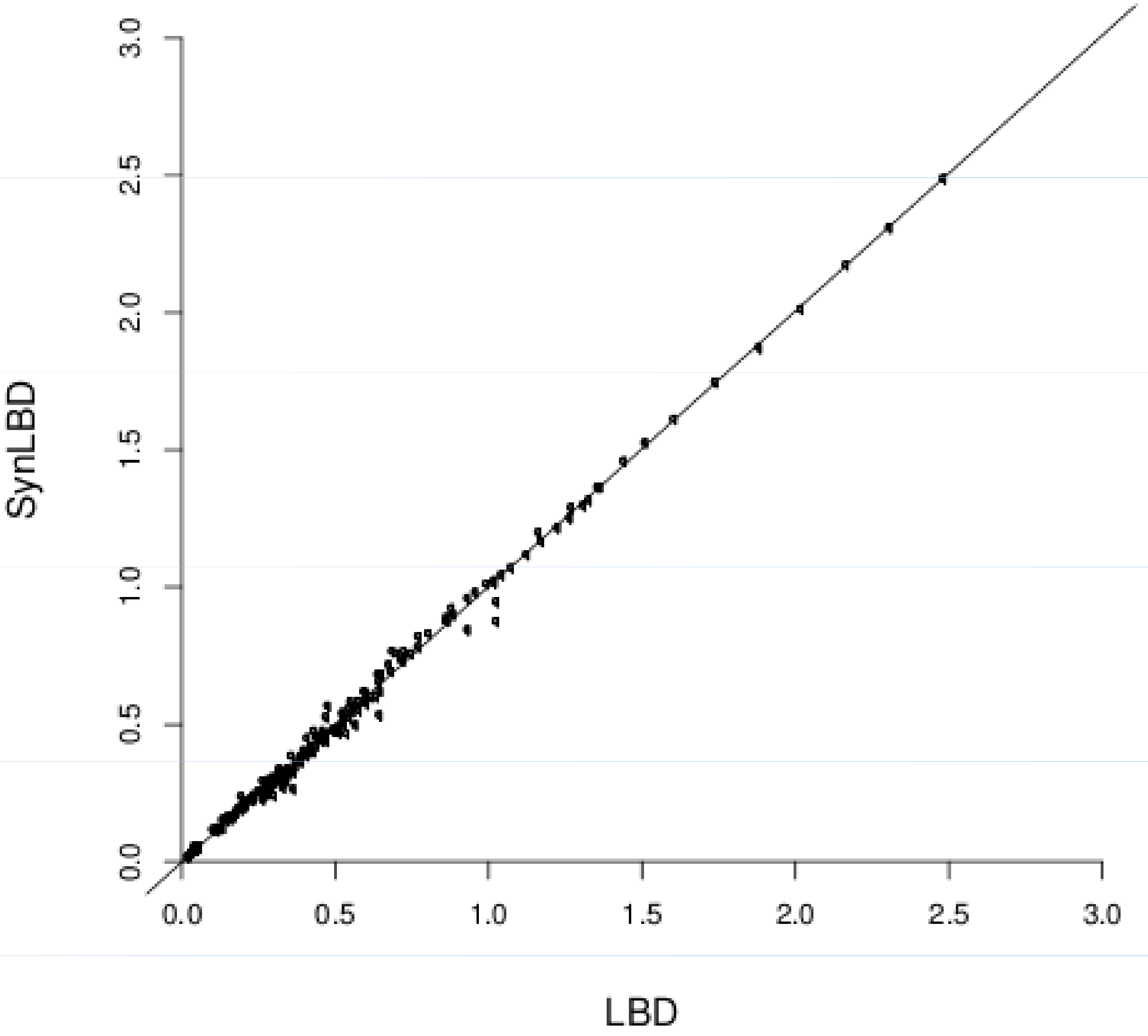


Figure 4: Share of Payroll by Industry Sector and Year, 1976-2000

$$EMP_i = \alpha + \beta EMP_{i-1} + \delta PAY_i + \theta IND_i + \psi STATE_i + \vartheta AGE_i + \gamma MU_i + \epsilon$$

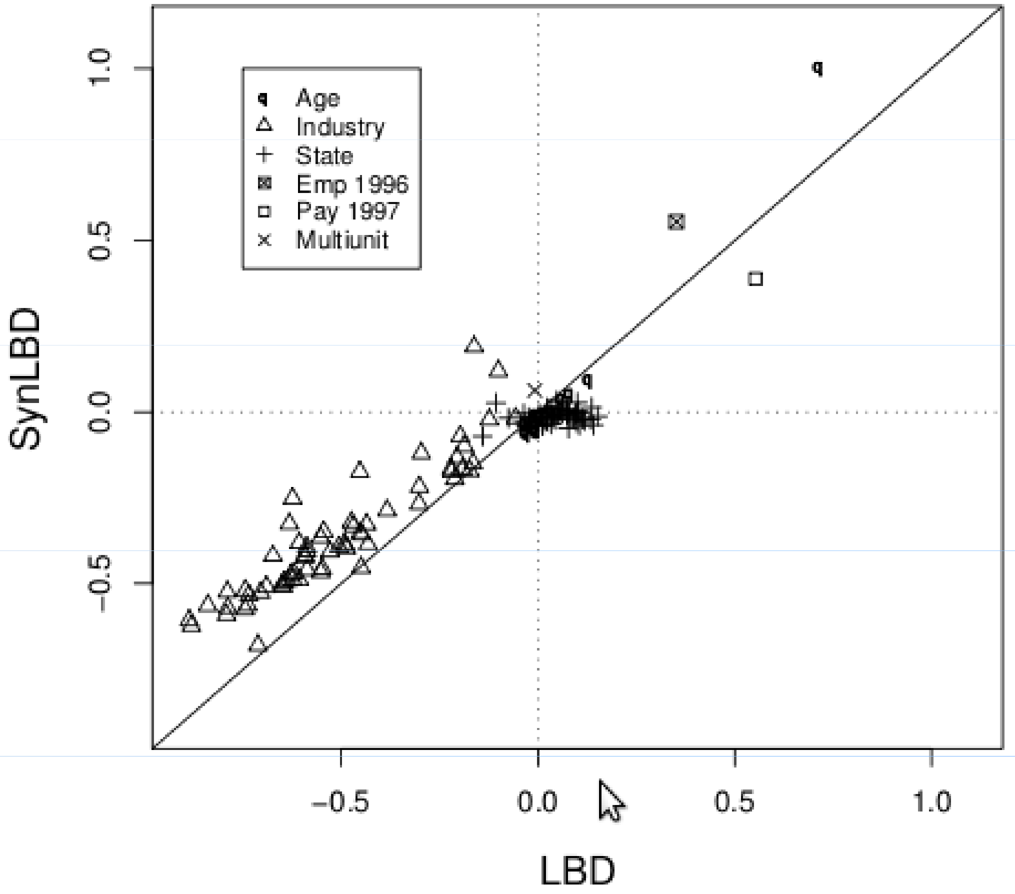


Figure 11: Regression Coefficients, LBD vs Synthetic

# Confidentiality Protection

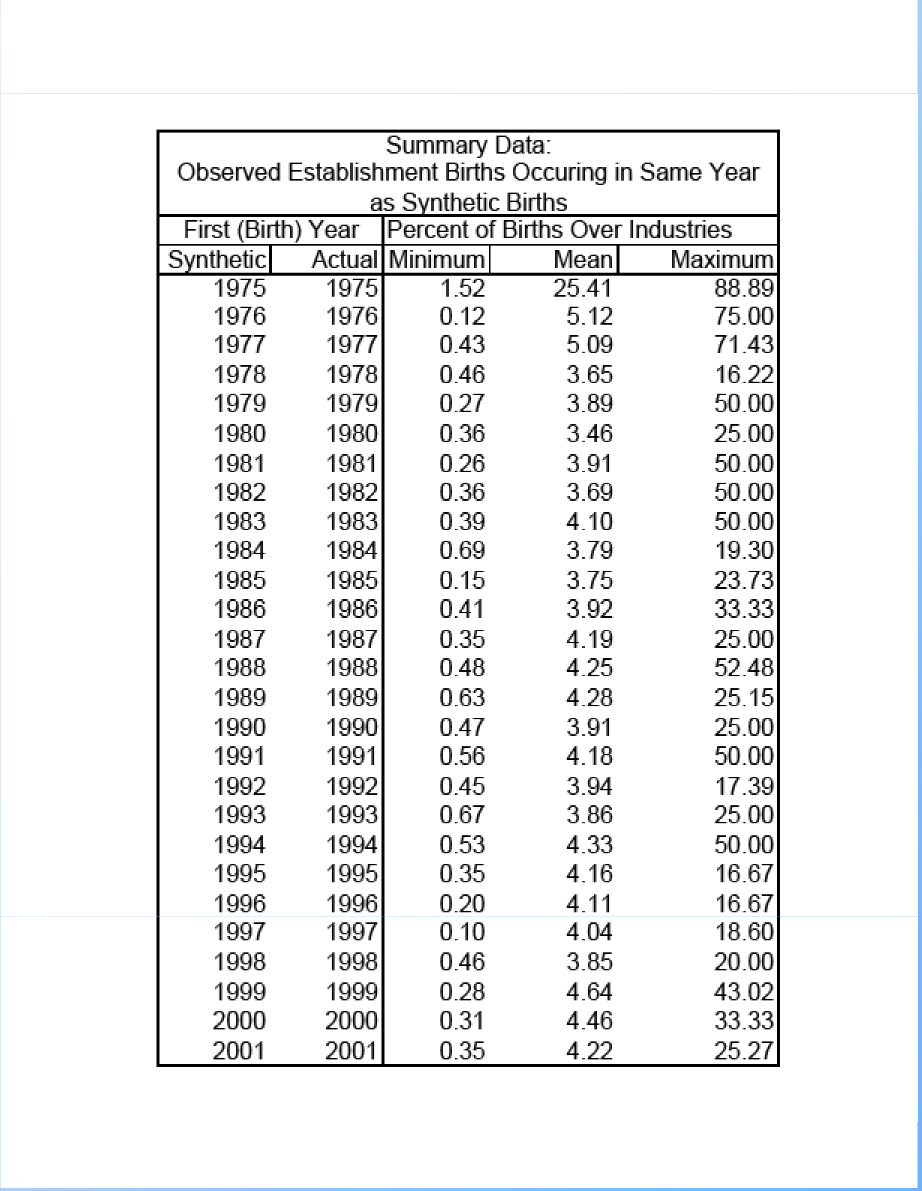
- Unavailable in SynLBD v2
  - Firm structure
  - firm linkages (across time, across implicates)
  - Geography
- Basic protection
  - replacing sensitive values of with draws from probability distributions

# 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

# Synthesizing Firstyear (Birth) and Lastyear (Death)

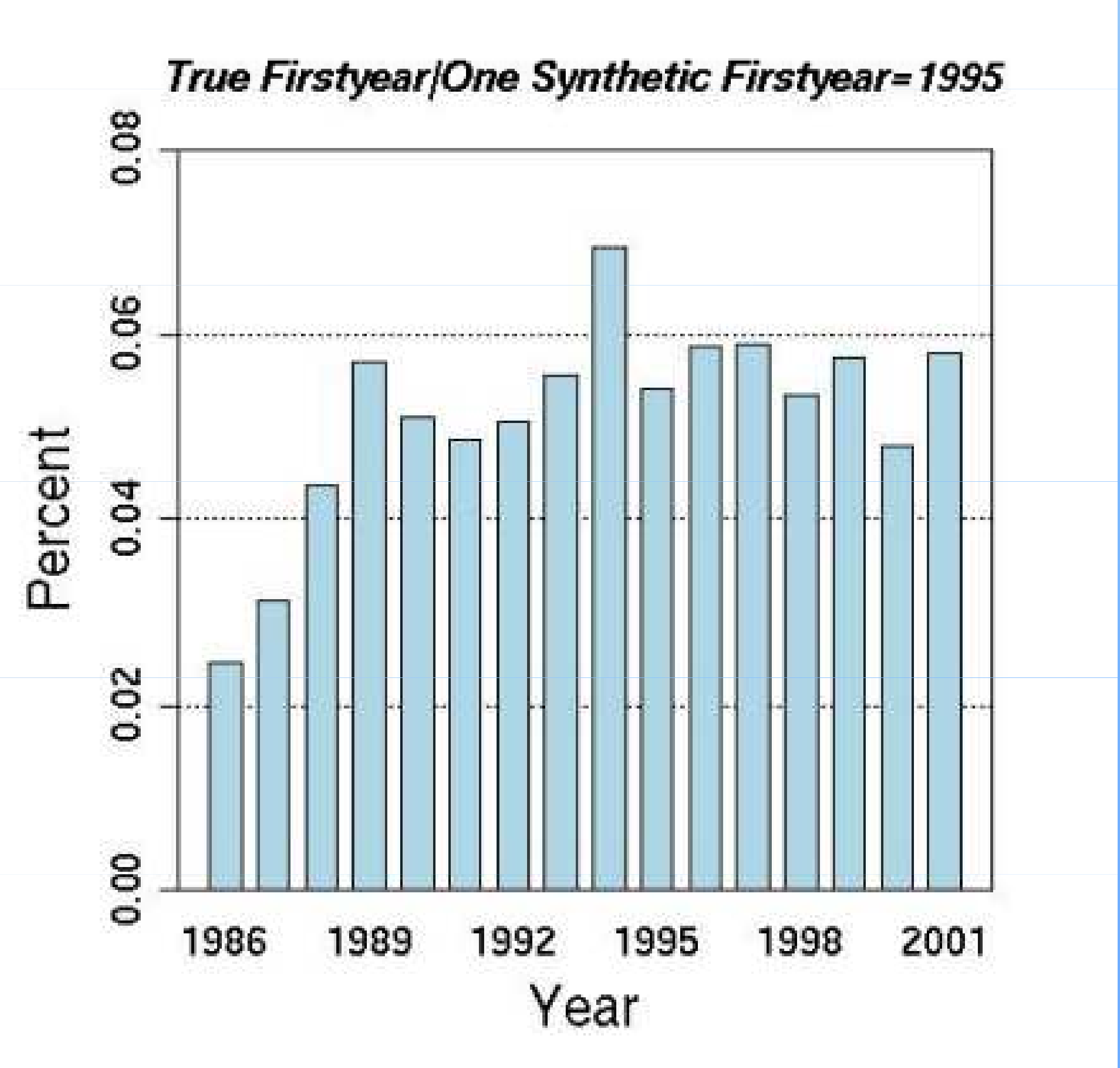
- Positive probability exists of producing any feasible birth year, and substantial probability exists that synthesized firstyear is not the actual firstyear
- Table on next slide shows this:  prob(actual birth year=synthetic birth year l synthetic birth year) is low
- Similar results hold for deaths
- Conclusions:  establishment lifetimes are random, so users can’t accurately attach establishment identifications to them



Summary Data:
Observed Establishment Births Occuring in Same Year
as Synthetic Births

| First (Birth) Year | | Percent of Births Over Industries | | |
|---|---|---|---|---|
| Synthetic | Actual | Minimum | Mean | Maximum |
| 1975 | 1975 | 1.52 | 25.41 | 88.89 |
| 1976 | 1976 | 0.12 | 5.12 | 75.00 |
| 1977 | 1977 | 0.43 | 5.09 | 71.43 |
| 1978 | 1978 | 0.46 | 3.65 | 16.22 |
| 1979 | 1979 | 0.27 | 3.89 | 50.00 |
| 1980 | 1980 | 0.36 | 3.46 | 25.00 |
| 1981 | 1981 | 0.26 | 3.91 | 50.00 |
| 1982 | 1982 | 0.36 | 3.69 | 50.00 |
| 1983 | 1983 | 0.39 | 4.10 | 50.00 |
| 1984 | 1984 | 0.69 | 3.79 | 19.30 |
| 1985 | 1985 | 0.15 | 3.75 | 23.73 |
| 1986 | 1986 | 0.41 | 3.92 | 33.33 |
| 1987 | 1987 | 0.35 | 4.19 | 25.00 |
| 1988 | 1988 | 0.48 | 4.25 | 52.48 |
| 1989 | 1989 | 0.63 | 4.28 | 25.15 |
| 1990 | 1990 | 0.47 | 3.91 | 25.00 |
| 1991 | 1991 | 0.56 | 4.18 | 50.00 |
| 1992 | 1992 | 0.45 | 3.94 | 17.39 |
| 1993 | 1993 | 0.67 | 3.86 | 25.00 |
| 1994 | 1994 | 0.53 | 4.33 | 50.00 |
| 1995 | 1995 | 0.35 | 4.16 | 16.67 |
| 1996 | 1996 | 0.20 | 4.11 | 16.67 |
| 1997 | 1997 | 0.10 | 4.04 | 18.60 |
| 1998 | 1998 | 0.46 | 3.85 | 20.00 |
| 1999 | 1999 | 0.28 | 4.64 | 43.02 |
| 2000 | 2000 | 0.31 | 4.46 | 33.33 |
| 2001 | 2001 | 0.35 | 4.22 | 25.27 |

# Example: Year of birth



# Confidentiality Protection: Breaking Firm Links

- Firm characteristics not synthesized
- Firm characteristics more skewed than establishment characteristics
- Cannot link multi-unit establishments to their firms

# Confidentiality Protection: Breaking Links Across Implicates

- Synthetic observations with the same LBDnum across implicates are not generated from the same LBD establishment
- Can’t group (across implicates within year) observations generated from same establishment

# Confidentiality Protection: Synthesizing Employment and Payroll

- Synthesis models are essentially regressions with transformed variables
- Synthesis captures low-dimensional relationships and sacrifices higher-dimensional ones
- Synthesized employment and payroll vary substantially around regression lines
- Synthesized employment and payroll vary significantly from observed values

# Example: Correlations Among Actual and Synthetic Data

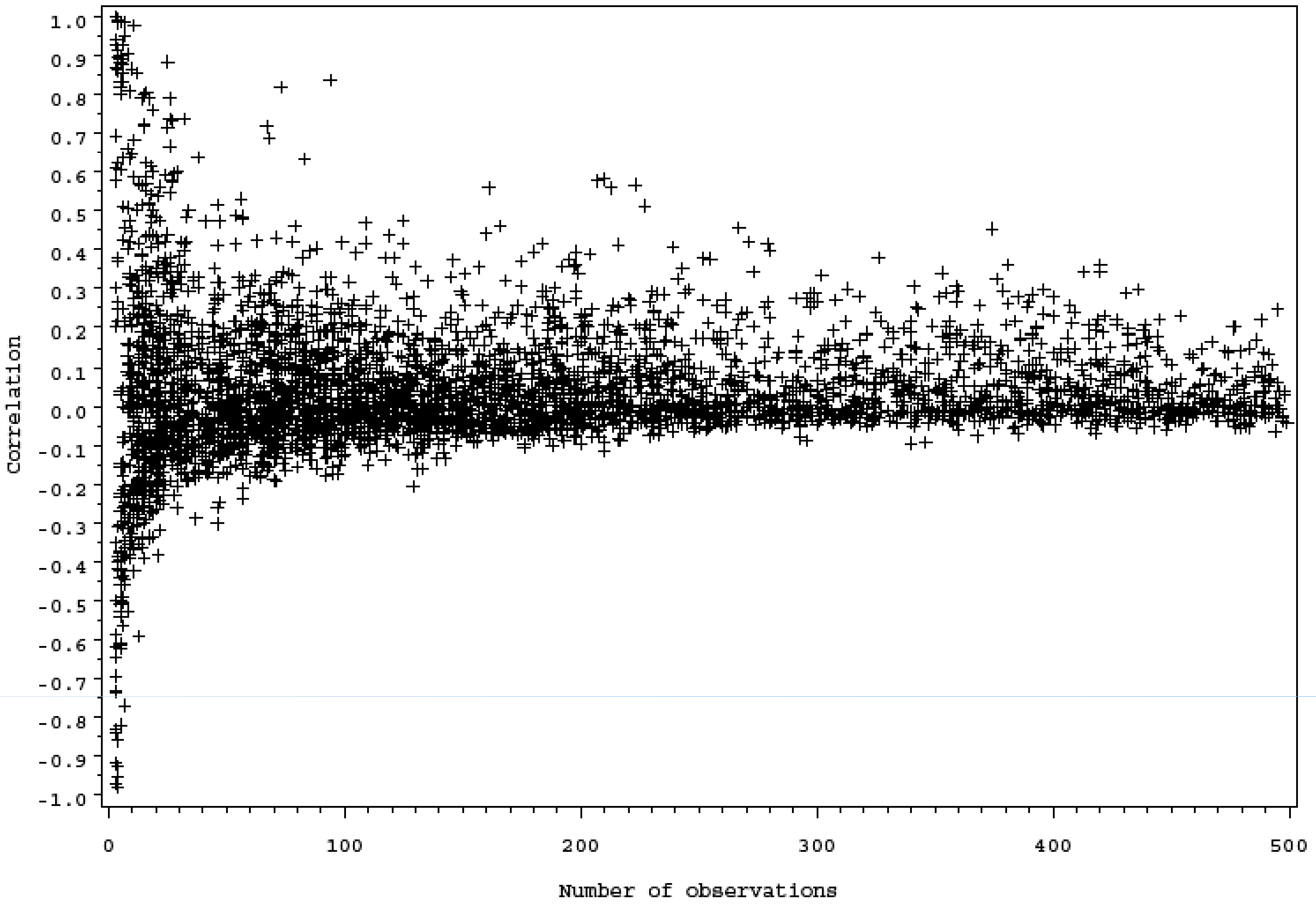
- SIC 573 - year 2000

Pearson Correlation Coefficients
SIC 573
Year: 2000

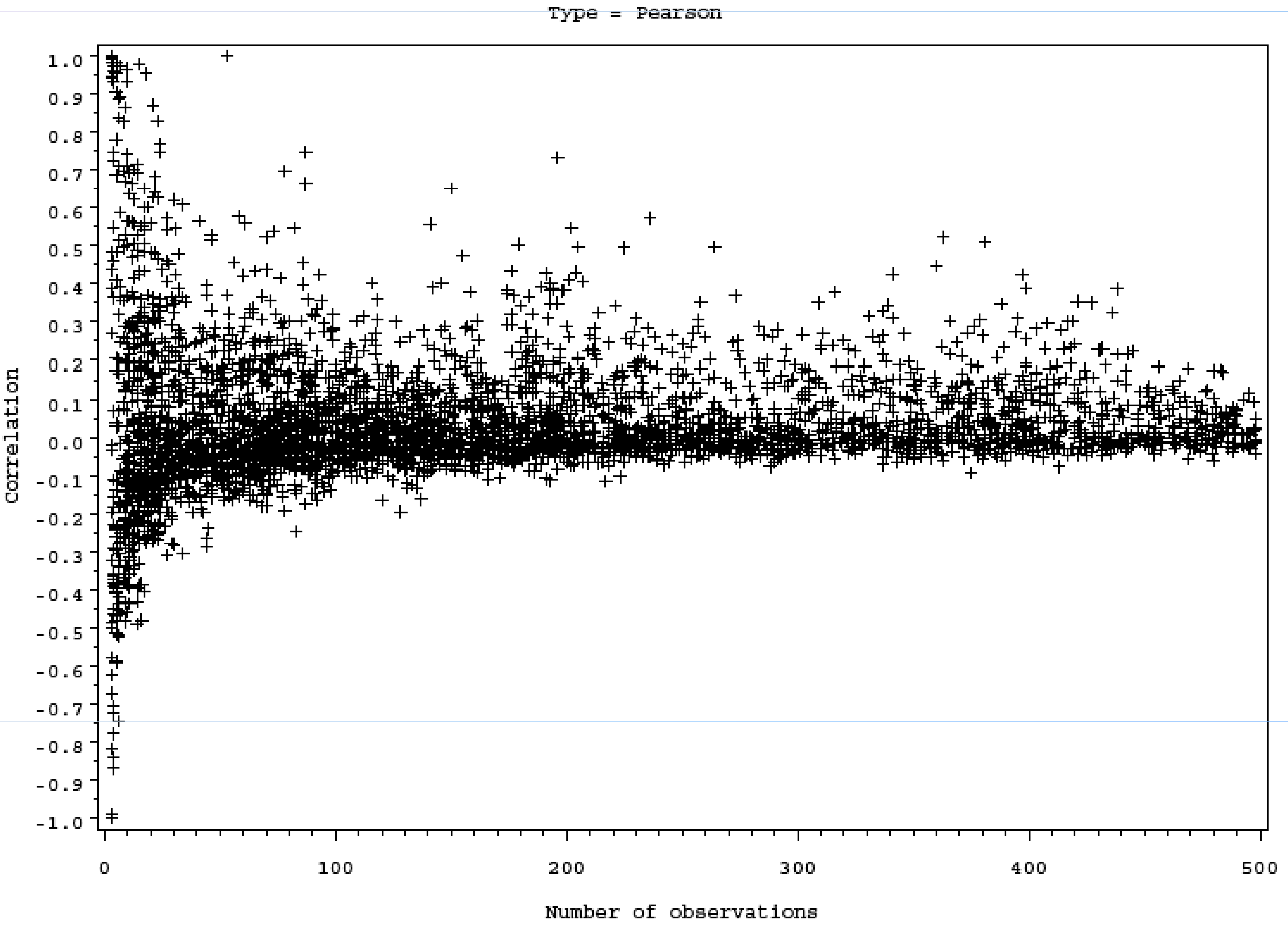
| | Employment | Synthetic Employment | Payroll | Synthetic Payroll |
|---|---|---|---|---|
| Employment | 1<br>41000 | | | |
| Synthetic Employment | 0.003<br>21100 | 1<br>41000 | | |
| Payroll | 0.712<br>41000 | -0.012<br>21100 | 1<br>41000 | |
| Synthetic Payroll | 0.007<br>21100 | 0.444<br>41000 | 0.004<br>21100 | 1<br>41000 |

# Correlations of observed vs synthetic Employment

Type = Pearson



# Correlations of observed vs synthetic Payroll



# Conclusions

- Analytical validity supported for broad analyses
  - Issues with some details
  - Obtain user feedback to inform future refinements
- Sufficient confidentiality protection
  - Basic metrics show strong protection
  - Differential privacy protection not yet verified

# Ongoing work at Census

- Include NAICS, geography, changes in multiunit status, firm age & size
- Multiple Imputations for release
- Address bias in job creation/destruction
- Extend time series