Issues in Disclosure Avoidance at the Census Bureau (for Synthetic Data)

Arnold Reznek, U.S. Census Bureau

Presented at
Joint NSF-Census-IRS Workshop
on
Synthetic Data and Confidentiality Protection
July 31, 2009

Grant ITR-0427889 – Info Tech Challenges for Secure Access to Confidential Social Science Data

Introduction

Census Bureau has released or is developing several data sets discussed at this workshop:

- LEHD
- SIPP Synthetic Beta
- Decennial Census
- OnTheMap
- Synthetic LBD
- ACS Group Quarters

This presentation summarizes issues related to confidentiality protection of synthetic data at the Census Bureau

Structure of This Talk

The Issues:

- What are the disclosure risks?
- How do we balance confidentiality and data utility?
- What disclosure analysis methods should we use?

The talk summarizes these issues, emphasizing the first and third, and with reference to the Synthetic LBD

The Synthetic LBD – A Review

Synthesized Variables (Y):

- First Year (1975-2001)
- Last Year (1976-2005)
- Multi-unit status (1-5)
- March 12 Employment (1976-2001)
- Annual Payroll (1976-2001)

Unsynthesized variables (X):

- County/State, SIC (3 digit)

Other (not used)

- EIN, NAICS, firm structure, legal form

Synthesis methods discussed this morning

What are the disclosure risks?

For business data:

- Disclosure risk is highest at the firm level
- Attribute disclosures more problematic than identity disclosure
- Geography-by-industry combinations are the most risky for attribute disclosure
- Presence of only one establishment (not firm) in industry-county is shown in County Business Patterns
- Only employment intervals shown in CBP

How do we protect the data?

"Move" the data a lot

If someone claims the synthetic datafile contains "my data", have plausible argument that

- This is an accident
- The probability is very low it is "my data."

For Synthetic LBD:

- We were very cautious on confidentiality protection
  - First try at releasing a synthetic file of business microdata
  - We want to get a version out relatively quickly to get feedback
- Discussion follows of the disclosure issues and the protection methods

Decisions to Reduce Risk in this Beta Version

We release no geography released within SIC3

- Geography is used in synthesis
- We anticipate releasing geography in future versions

We Break Firm Links

- No firm characteristics synthesized, except multi-unit firm membership
- Attackers cannot determine establishments owned by particular firms, aggregate synthetic data to firm level, or identify these firms
- We anticipate including more geography in future versions

Decisions to Reduce Risk (cont'd)

We release only one implicate in this beta version

- Anticipate releasing more implicates (from same synthesizer) in the future

Properties of “Randomized Sanitizer”

- For all synthesized variables, domain of possible synthesized data is same as domain of all possible (not just observed) confidential data
- All possible values of synthesized data can occur with positive probability
- Any point in possible confidential data can map to any point in synthetic data
- There are no exact disclosures, in “differential privacy” sense

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

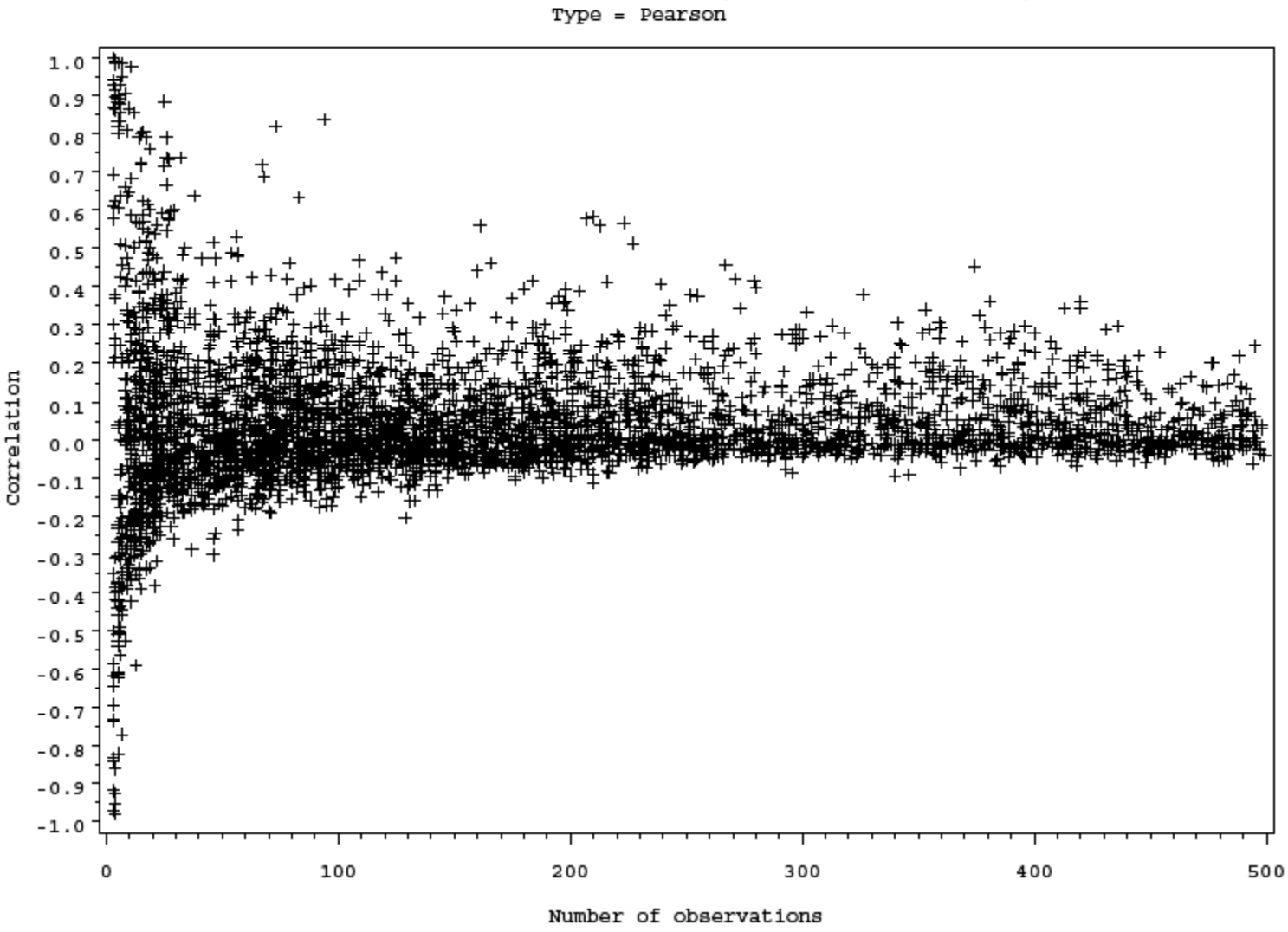
Synthsizing Births and Deaths Moves Establishment Lifetimes

Presence of Actual and Synthetic Establishments in a Year				
		Synthetic Present		
		Yes	No	
Actual Present	Yes	A	B	Actual
	No	C	D	
		Synthetic		

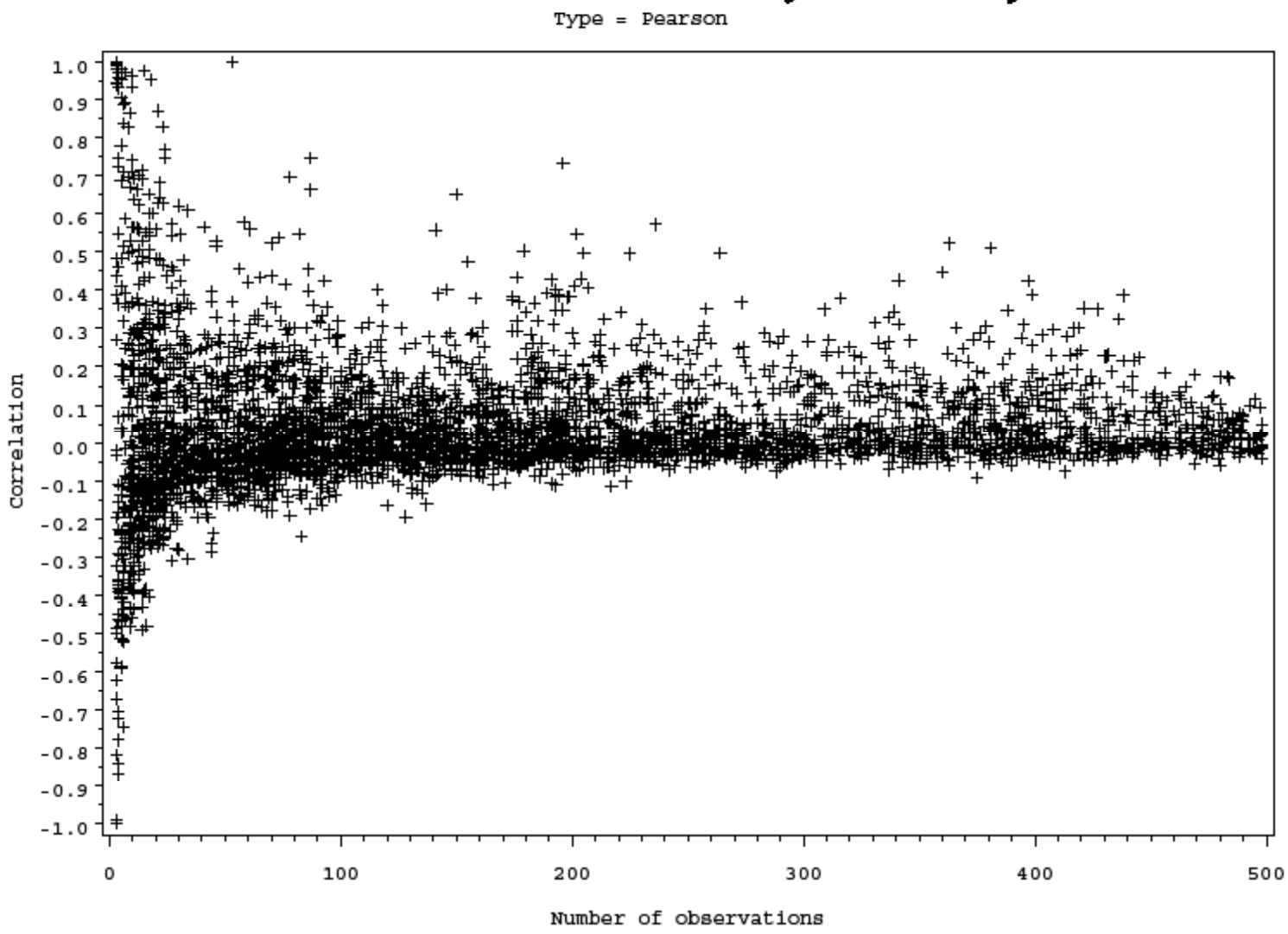
Synthesizing Employment and Payroll

- 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
- Next two slides show this is true

Correlations of observed vs synthetic Employment

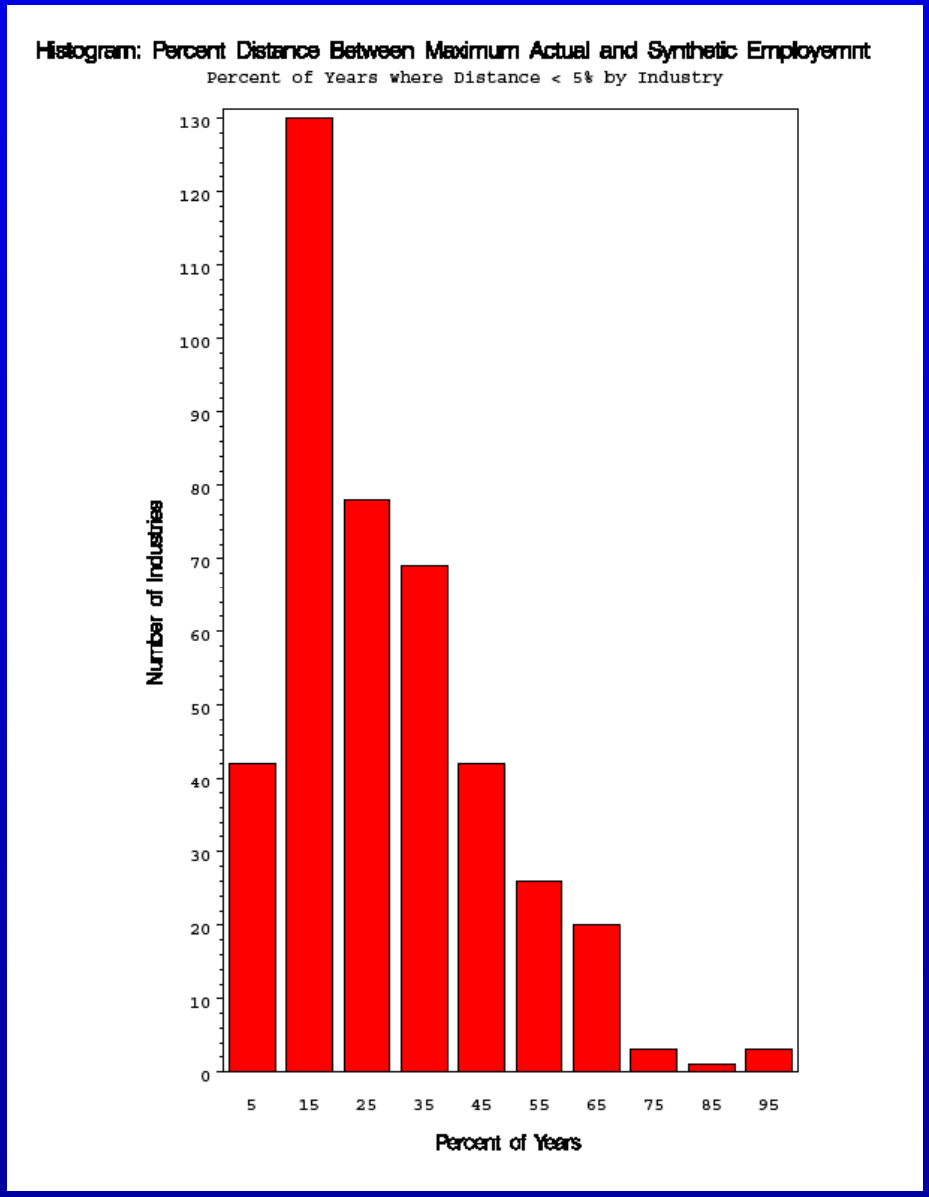


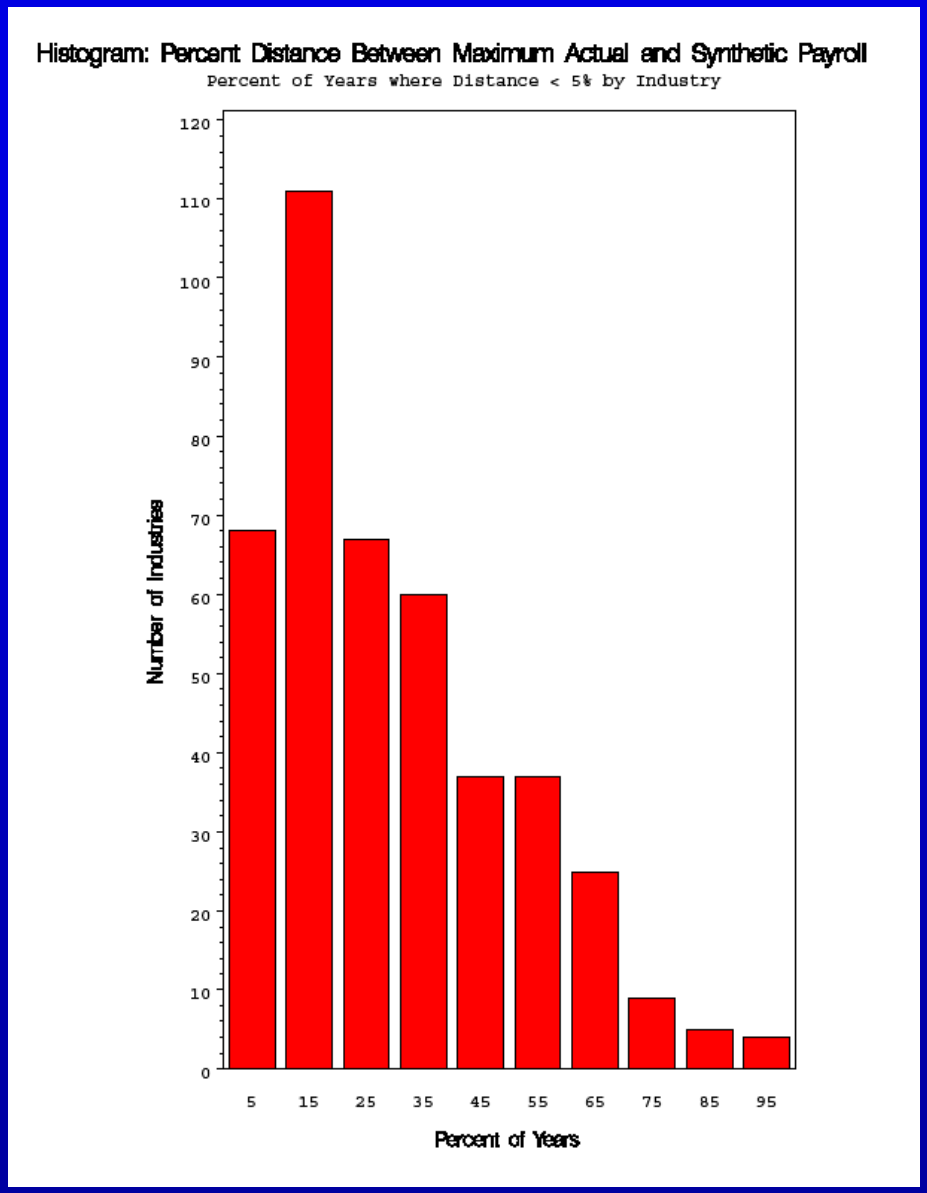
Correlations of observed vs synthetic Payroll



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 the upper tail, but not necessarily close to real ones
- Next two slides support this contention


Histogram: Percent Distance Between Maximum Actual and Synthetic Employemnt
Percent of Years where Distance < 5% by Industry
Number of Industries
130
120
110
100
90
80
70
60
50
40
30
20
10
0
5
15
25
35
45
55
65
75
85
95
Percent of Years


Histogram: Percent Distance Between Maximum Actual and Synthetic Payroll
Percent of Years where Distance < 5% by Industry
Number of Industries
120
110
100
90
80
70
60
50
40
30
20
10
0
5
15
25
35
45
55
65
75
85
95
Percent of Years

Conclusions on Disclosure Risk

- Synthetic firstyear and lastyear values differ from actual firstyear and last year values
- Very little correlation between synthetic employment and real employment values, and between synthetic payroll and real payroll values
- Synthetic maxima within any industry can vary widely from the actual maxima within the industry
- Synthetic LBD expected to satisfy the stringent requirements of differential privacy protection

Overall Conclusions

- We have approval from Census Bureau and IRS to release LBD synthetic beta
- This talk summarized decisions made and disclosure methods used to get approval
- It illustrates the issues the Census Bureau faces in developing and releasing synthetic data
- We want to get feedback to improve future versions