# Using partially synthetic data to replace suppression in the Business Dynamics Statistics: early results

Javier Miranda[2] Lars Vilhuber[1]

[1] Labor Dynamics Institute, ILR, Cornell University, United States

[2] Center for Economic Studies, U.S. Census Bureau, United States

September 2014, PSD2014, Eivissa

## Funding

- Vilhuber's work is partially funded by NSF Grant #1042181.

## Disclaimer

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion. Any findings, conclusions or opinions are those of the authors. They do not necessarily reflect those of the Center for Economic Studies, the U.S. Census Bureau, or the National Science Foundation.

## Business Dynamics

”The U.S. economy is comprised of over 6 million establishments with paid employees. The population of these businesses is constantly churning – some businesses grow, others decline and yet others close. New businesses are constantly replenishing this pool.”[*]

## Questions

- Small businesses’ contribution to job and productivity growth
- ... or is it young businesses’ contribution?
- Dynamics of businesses in their early (post-founding) years

# Data for Business Statistics in the United States

## Provided by the US Census Bureau

- County Business Patterns (CBP)
- Annual Survey of Manufactures (ASM)
- and over 100 separate additional surveys..
- Economic Census
- Business Dynamic Statistics (BDS)
- Quarterly Workforce Indicators (QWI)

# Data for Business Statistics in the United States

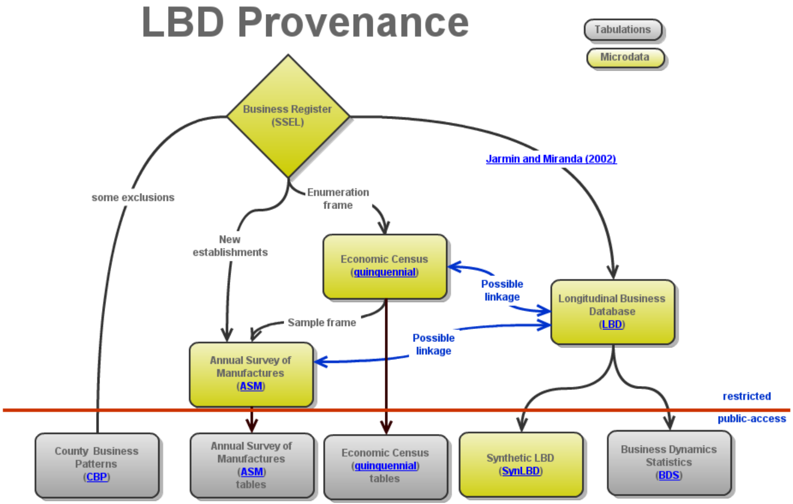
## Provided by the US Census Bureau

- County Business Patterns (CBP)
- Annual Survey of Manufactures (ASM)
- and over 100 separate additional surveys..
- Economic Census
- Business Dynamic Statistics (BDS)
- Quarterly Workforce Indicators (QWI)

## Provided by others

- Quarterly Census of Employment and Wages (QCEW, by BLS)
- Compustat (S & P)

# Business Microdata at the Census Bureau



LBD Provenance

Tabulations

Microdata

Business Register (SSEL)

some exclusions

Enumeration frame

New establishments

Jarmin and Miranda (2002)

Economic Census (quinquennial)

Possible linkage

Longitudinal Business Database (LBD)

Sample frame

Possible linkage

Annual Survey of Manufactures (ASM)

restricted

public-access

County Business Patterns (CBP)

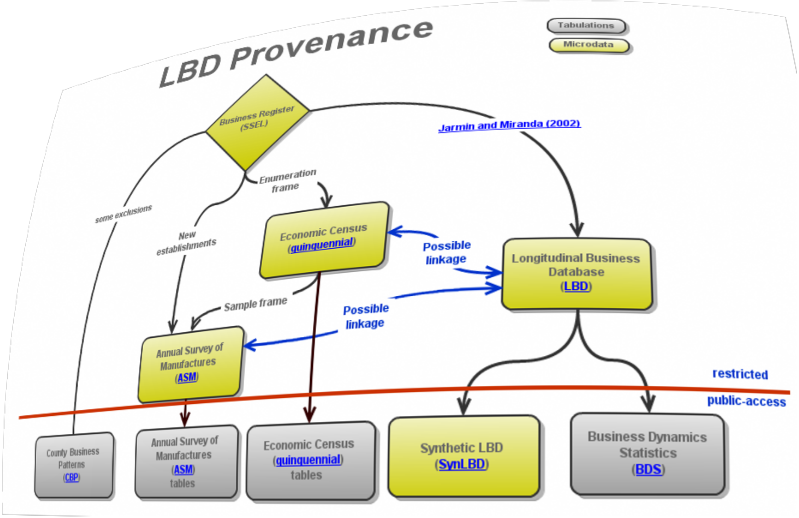
Annual Survey of Manufactures (ASM) tables

Economic Census (quinquennial) tables

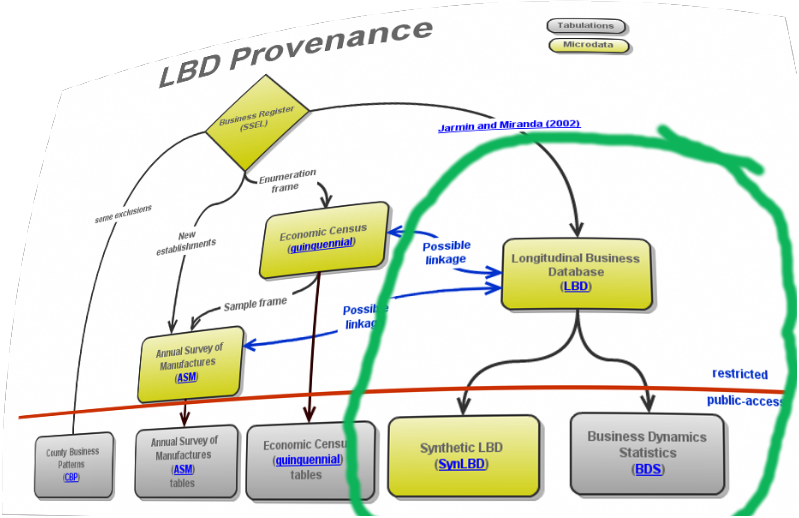
Synthetic LBD (SynLBD)

Business Dynamics Statistics (BDS)

# Business Microdata at the Census Bureau


LBD Provenance
Tabulations
Microdata
Business Register (SSEL)
Jarmin and Miranda (2002)
some exclusions
Enumeration frame
New establishments
Economic Census (quinquennial)
Possible linkage
Longitudinal Business Database (LBD)
Sample frame
Possible linkage
Annual Survey of Manufactures (ASM)
restricted
public-access
County Business Patterns (CBP)
Annual Survey of Manufactures (ASM) tables
Economic Census (quinquennial) tables
Synthetic LBD (SynLBD)
Business Dynamics Statistics (BDS)

# Business Microdata at the Census Bureau


LBD Provenance
Tabulations
Microdata
Business Register (SSEL)
Jarmin and Miranda (2002)
Enumeration frame
some exclusions
New establishments
Economic Census (quinquennial)
Possible linkage
Longitudinal Business Database (LBD)
Sample frame
Possible linkage
Annual Survey of Manufactures (ASM)
restricted
public-access
County Business Patterns (CBP)
Annual Survey of Manufactures (ASM) tables
Economic Census (quinquennial) tables
Synthetic LBD (SynLBD)
Business Dynamics Statistics (BDS)

# Business Dynamic Statistics

## Annual data series

- Establishment - level business dynamics: by firm age and firm size
- Employment - job creation and destruction
- Job expansions and contractions
- Number of establishments
- Establishment openings and closings
- Number of startups and firm shutdowns

More info: www.census.gov/ces/dataproducts/bds/

# Available BDS tabulations

## Firm and Establishment Characteristics

- Sector
- Firm Size
- Firm Age
- Initial Firm Size
- Geography (State, Metro/Non-metro, MSA)
- Cross-tabulations by up to three of these characteristics

## Lots of detail

62 very detailed tables

# Disclosure avoidance in the BDS

## P-percent rule with secondary suppressions

- Cells where the top 2 firms account for more than *P* percent of the total value of the cell are flagged for suppression

# Disclosure avoidance in the BDS

## P-percent rule with secondary suppressions

- Cells where the top 2 firms account for more than *P* percent of the total value of the cell are flagged for suppression
- *P* value is not disclosed

# Disclosure avoidance in the BDS

## P-percent rule with secondary suppressions

- Cells where the top 2 firms account for more than *P* percent of the total value of the cell are flagged for suppression
- *P* value is not disclosed
- Trivially: cells with fewer than 3 firms represented are always suppressed

# Disclosure avoidance in the BDS

## P-percent rule with secondary suppressions

- Cells where the top 2 firms account for more than *P* percent of the total value of the cell are flagged for suppression
- *P* value is not disclosed
- Trivially: cells with fewer than 3 firms represented are always suppressed
- Secondary suppressions: “minimize the amount of information loss in a given table row or column”.

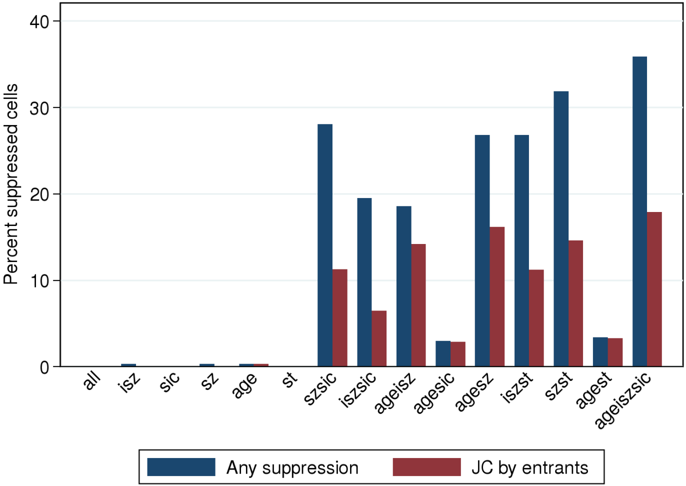
# Extent of suppression

Table : Suppressions in establishment-level BDS

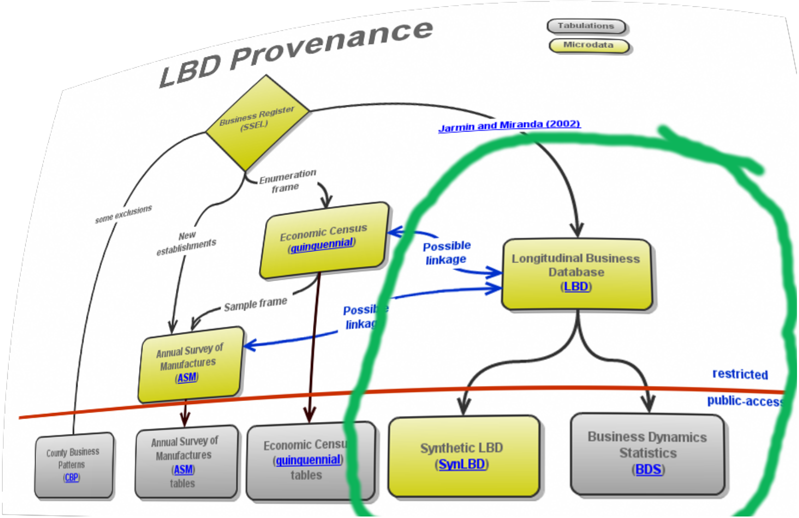
| Type | Level | Number of cells | Suppressions (%) Any | Suppressions (%) Job creation by entrants |
|---|---|---|---|---|
| Age | e | 325 | 0.3 | 0.3 |
| Age-Initial Size | e | 2925 | 18.6 | 14.2 |
| Age-Initial Size-SIC | e | 25994 | 35.9 | 17.9 |
| Age-SIC | e | 2925 | 3 | 2.9 |
| Age-State | e | 18360 | 3.4 | 3.3 |
| Age-Size | e | 2925 | 26.8 | 16.2 |
| All | e | 35 | 0 | 0 |
| Initial Size | e | 315 | 0.3 | 0 |
| Initial Size-SIC | e | 2835 | 19.5 | 6.5 |
| Initial Size-State | e | 17847 | 26.8 | 11.2 |
| SIC | e | 315 | 0 | 0 |
| State | e | 1785 | 0 | 0 |
| Size | e | 315 | 0.3 | 0 |
| Size-SIC | e | 2834 | 28.1 | 11.3 |
| Size-State | e | 17848 | 31.9 | 14.6 |

Note: Cells are year *x* categories, where the number of categories varies by published table.

# Extent of suppression



# Business Microdata at the Census Bureau


LBD Provenance
Tabulations
Microdata
Business Register (SSEL)
Jarmin and Miranda (2002)
Enumeration frame
some exclusions
New establishments
Economic Census (quinquennial)
Possible linkage
Longitudinal Business Database (LBD)
Sample frame
Possible linkage
Annual Survey of Manufactures (ASM)
restricted
public-access
County Business Patterns (CBP)
Annual Survey of Manufactures (ASM) tables
Economic Census (quinquennial) tables
Synthetic LBD (SynLBD)
Business Dynamics Statistics (BDS)

# Purpose of SynLBD

The SynLBD is

- derived from confidential Longitudinal Business Database (LBD, [5])

# Purpose of SynLBD

The SynLBD is

- derived from confidential Longitudinal Business Database (LBD, [5])
- designed to facilitate researcher access to establishment microdata (LBD)

# Purpose of SynLBD

The SynLBD is

- derived from confidential Longitudinal Business Database (LBD, [5])
- designed to facilitate researcher access to establishment microdata (LBD)
- while preserving the confidentiality of establishment/business data.

# Purpose of SynLBD

The SynLBD is

- derived from confidential Longitudinal Business Database (LBD, [5])
- designed to facilitate researcher access to establishment microdata (LBD)
- while preserving the confidentiality of establishment/business data.
- part of a larger strategy by the Census Bureau to provide *better statistics on business dynamics* CNSTAT [8]

# Contents of (Syn)LBD

## Data elements

- longitudinal establishment identifiers (created using probabilistic matching [5])
- information on birth, death
- employment and payroll over time
- location
- industry
- firm affiliation of employer establishments

# Contents of (Syn)LBD

## Data elements

- longitudinal establishment identifiers (created using probabilistic matching [5]) Masked
- information on birth, death Synthesized
- employment and payroll over time Synthesized
- location Suppressed
- industry Released
- firm affiliation of employer establishments → next version

# Contents of (Syn)LBD

## Data elements

- longitudinal establishment identifiers (created using probabilistic matching [5]) Masked
- information on birth, death Synthesized
- employment and payroll over time Synthesized
- location Suppressed
- industry Released
- firm affiliation of employer establishments → next version

## Complete description

Kinney et al [7]

[more]

## Putting two and two together...

V2.0 of SynLBD released by Census Bureau’s Disclosure Review Board in 2011

## Putting two and two together...

V2.0 of SynLBD released by Census Bureau's Disclosure Review Board in 2011

Let's combine public-use data to fill in suppressions

# Combining synthetic and protected data

## Initially...

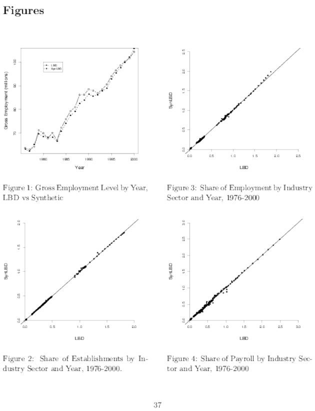
... explored as part of Kaj Gitting's thesis [3]

... expanded as part of our FCSM paper [1]

## Could it work for BDS?

- LBD underlies BDS
- SynLBD derived from LBD
- SynLBD proven analytic validity

# Analytic validity



**Figures**

Figure 1: Gross Employment Level by Year, LBD vs 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

37

# Analytic validity

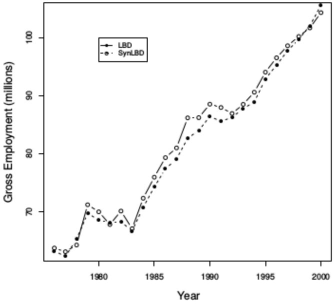


Figure 1: Gross Employment Level by Year, LBD vs Synthetic

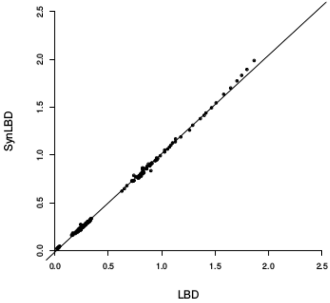


Figure 3: Share of Employment by Industry Sector and Year, 1976-2000

# Analytic validity

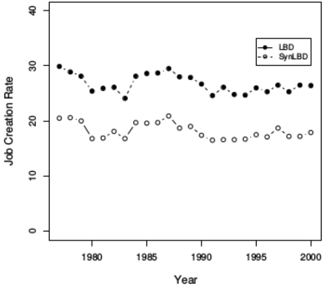


Figure 8: Job Creation Rate by Year, LBD vs Synthetic

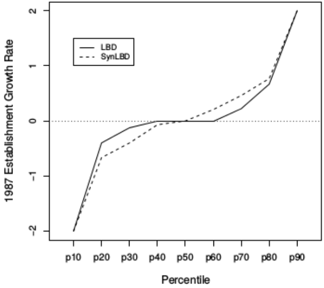


Figure 9: Distribution of Job Creation Rates, LBD vs Synthetic

# Notation

## Base variable

Establishment employment $e_{jt}$.

## Example

$$birth_{jt} = \begin{cases} 1 & \text{if } e_{jt} > 0 \text{ and } e_{jt-s} = 0 \ \forall s \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$jcbirth_{jt} = \begin{cases} e_{jt} - ejt - 1 & \text{if } e_{jt} > 0 \text{ and } e_{jt-s} = 0 \ \forall s \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

# Notation

## Synthetic values

Synthesized version of variable $x_{jt}$ is denoted $\tilde{x}_j t$.

## Cells

Collections of characteristics $k_t(j)$ (industry, geography, establishment or firm age and size)

$j \in K'_t$ describes the set of firms at time $t$ such that $k_t(j) = k'$.

# Notation

## Aggregations

Generically in capital letters:

$$E_{\cdot t} = \sum_{i=1}^{J} e_{it}, \tag{3}$$

Aggregations across establishments having characteristics $k'$ at time $t$

$$X_{k't} = \sum_{j \in K'_t} x_{jt} \tag{4}$$

# Suppression rules

## Suppression rules

for (aggregate) variable $X$ are captured by $I_t^X$, such that the releasable variable $X^o$ under the current regime can be described by

$$X^o_{k't} = \begin{cases} X_{k't} & \text{if } I^X_{kt} = 1 \\ \text{missing} & \text{otherwise} \end{cases} \tag{5}$$

# Algorithm 1

We can now express the “drop-in” algorithm, leading to the released variable $X^{(i)}$, as:

BDS$^{(i)}$

```
if $I_t^X = 1$ then
    $X_{k't}^{(i)} = X_{k't}$
else
    $X_{k't}^{(i)} = \tilde{X}_{k't}$
end if
```

# Analysis

## Analysis

- We implemented Algorithm 1 for Business Dynamics Statistics (BDS) tabulations by establishment age and size (`bds_e_agesz`).

# Analysis

## Analysis

- We implemented Algorithm 1 for BDS tabulations by establishment age and size (`bds_e_agesz`).
- About 26% of all cells have some suppression

# Analysis

## Analysis

- We implemented Algorithm 1 for BDS tabulations by establishment age and size (`bds_e_agesz`).
- About 26% of all cells have some suppression
- Here: variable, “Job Creation by establishment births” (`job_creation_births`).

# Protection

## Protection through synthesis

- Cells are filled in with data available to a wide audience (public-use)

# Protection

## Protection through synthesis

- Cells are filled in with data available to a wide audience (public-use)
- ....(but which typically cannot create tabulations)

# Protection

## Protection through synthesis

- Cells are filled in with data available to a wide audience (public-use)
- ....(but which typically cannot create tabulations)
- ....(future tables will contain variables which are not currently available on the synthetic data file)

# Protection

## Protection through synthesis

- Cells are filled in with data available to a wide audience (public-use)
- ....(but which typically cannot create tabulations)
- ....(future tables will contain variables which are not currently available on the synthetic data file)
- Structurally: the synthetic data are ... fully synthetic (discussed in Kinney et al, 2011)

# Protection

## Protection through synthesis

- Cells are filled in with data available to a wide audience (public-use)
- ....(but which typically cannot create tabulations)
- ....(future tables will contain variables which are not currently available on the synthetic data file)
- Structurally: the synthetic data are ... fully synthetic (discussed in Kinney et al, 2011)
- Additional comparison: differences in each cell

# From Kinney et al

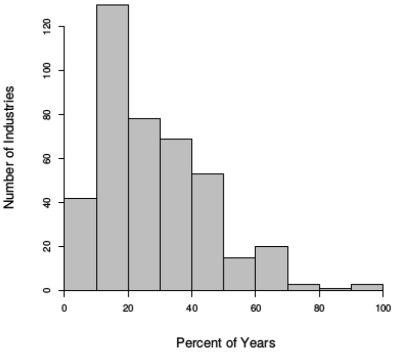


Figure 13: Histogram: Percent Distance Between Actual and Synthetic Employment

The comparison is for individual establishments, not within cells

# Cell-wise comparison

## Criteria for cell-wise comparison

- Differences in count of establishment in a cell
- Differences in values of cells

Not done yet.

# Analytic validity: time-series

## Setup

Estimate an AR(2) process for each of $X_{k't}$, $X^{s}_{k't}$, and $X^{(i)}_{k't}$

## Metrics

- number of missing time-series estimates
- the number of significant coefficients for the first lag of the AR(2)
- *coverage*, the percentage of regressions where the true $\rho_1$ lies within the confidence band around the coefficient estimated from the comparison $\rho^{s}_{1}$ and $\rho^{(i)}_{1}$,
- interval overlap measure $J_k$ [6]

# Analytic validity

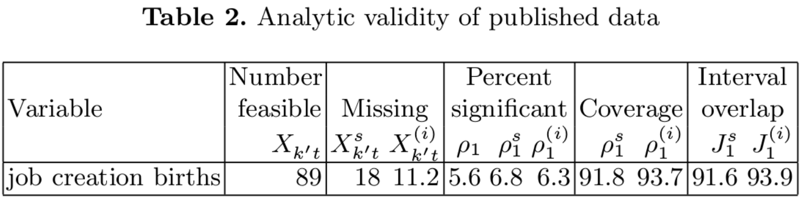


**Table 2.** Analytic validity of published data

| Variable | Number feasible | Missing | | Percent significant | | | Coverage | | Interval overlap | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $X_{k't}$ | $X^s_{k't}$ | $X^{(i)}_{k't}$ | $\rho_1$ | $\rho^s_1$ | $\rho^{(i)}_1$ | $\rho^s_1$ | $\rho^{(i)}_1$ | $J^s_1$ | $J^{(i)}_1$ |
| job creation births | 89 | 18 | 11.2 | 5.6 | 6.8 | 6.3 | 91.8 | 93.7 | 91.6 | 93.9 |

# Open issues

## Unexplored issues

- SynLBD is synthesized independently within industry

# Open issues

## Unexplored issues

- SynLBD is synthesized independently within industry
- Geography is not synthesized, not considered within synthesis process (and not released) - unclear how geography subtabulations will fare, what the disclosure avoidance implications are

# Open issues

## Unexplored issues

- SynLBD is synthesized independently within industry
- Geography is not synthesized, not considered within synthesis process (and not released) - unclear how geography subtabulations will fare, what the disclosure avoidance implications are
- Firm-level characteristics go into a bit more detail, and require availability of SynLBD v3

# Open issues

## Unexplored issues

- SynLBD is synthesized independently within industry
- Geography is not synthesized, not considered within synthesis process (and not released) - unclear how geography subtabulations will fare, what the disclosure avoidance implications are
- Firm-level characteristics go into a bit more detail, and require availability of SynLBD v3
- Time consistency of the series

# Open issues

## Unexplored issues

- SynLBD is synthesized independently within industry
- Geography is not synthesized, not considered within synthesis process (and not released) - unclear how geography subtabulations will fare, what the disclosure avoidance implications are
- Firm-level characteristics go into a bit more detail, and require availability of SynLBD v3
- Time consistency of the series
- Comparison to alternative “outside-the-firewall” imputation mechanisms ([4, 2])

# Postulated alternative algorithm

## *BDS*$^{(ii)}$

```
Compute: $X_{k't} = \sum_{j \in K'_t} x_{jt}$
Compute: $I_t^X$
**if** $I_t^X = 0$ **then**
    Assign all $j \in K'_t$ to $J^-_{k't}$
    Assign all $j \in J^-_{k's}$ to $J^-_{k't}$ for $t > s > t - n$
**end if**
Compute:
```

$$X_{k't}^{(ii)} = \sum_{j \in \left\{K'_t \cap J^-_{k't}\right\}} \tilde{x}_{jt} + \sum_{j \in K'_t \wedge j \notin J^-_{k't}} x_{jt}$$

For $n = \infty$, $J_t$ is an absorbing set, which seems undesirable.
For $n = 1$, this reduces to Algorithm 1.

# Conclusion

## Early in the process

- Desirable a-priori properties (use of public-use data to fill in blanks)
- May not work for other variables
- Assumes suppression as primary disclosure avoidance mechanism...

Thank you

$Id: Presentation-subdoc.tex 1580 2014-09-16 22:38:54Z lv39 $

More info:

- For information on the SynLBD, see goo.gl/eyrv7w
- Access through the Synthetic Data Server, www.vrdc.cornell.edu/sds/

# Extra slides

# Bibliography

J. M. Abowd, K. Gittings, K. L. McKinney, B. E. Stephens, L. Vilhuber, and S. Woodcock, “Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series,” Federal Committee on Statistical Methodology, Tech. Rep., January 2012. [Online]. Available: http://www.fcsm.gov/events/papers2012.html

J. R. Bradley, S. H. Holan, and C. K. Wikle, “Mixed Effects Modeling for Areal Data that Exhibit Multivariate-Spatio-Temporal Dependencies,” *ArXiv e-prints*, Jul. 2014.

R. K. Gittings, “Essays in labor economics and synthetic data methods,” Ph.D., Cornell University, 2009.

S. H. Holan, D. Toth, M. A. R. Ferreira, and A. F. Karr, “Bayesian multiscale multiple imputation with implications for data confidentiality,” *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 564–577, 2010. [Online]. Available: http://dx.doi.org/10.1198/jasa.2009.ap08629

R. Jarmin and J. Miranda, “The Longitudinal Business Database,” U.S. Census Bureau, Center for Economic Studies, Discussion Paper CES-WP-02-17, 2002.

A. F. KARR, C. N. KOHNEN, A. OGANIAN, J. P. REITER, and A. P. SANIL, “A framework for evaluating the utility of data altered to protect confidentiality,” vol. 60, no. 3, pp. 1–9, 2006.

S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd, “Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database,” *International Statistical Review*, vol. 79, no. 3, pp. 362–384, December 2011. [Online]. Available: http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html

Panel on Measuring Business Formation, Dynamics, and Performance:, J. Haltiwanger, L. M. Lynch, and C. Mackie, *Understanding Business Dynamics: An Integrated Data System for America’s Future*. National Research Council, The National Academies Press, 2007. [Online]. Available: http://www.nap.edu/openbook.php?record_id=11844

# Acronyms

BDS Business Dynamics Statistics

# Feedback loop

## Critical element

- Not just “release and forget”

## Closing the loop

- Researchers access the data on a special server (open internet, no RDC)
- No disclosure-avoidance analysis done on results created from SynLBD
- Validation server allows to request validation, release of results using confidential data (offline submission, full disclosure-avoidance)

# Feedback loop

## Critical element

- Not just “release and forget”
- First attempt, needs feedback

## Closing the loop

- Researchers access the data on a special server (open internet, no RDC)
- No disclosure-avoidance analysis done on results created from SynLBD
- Validation server allows to request validation, release of results using confidential data (offline submission, full disclosure-avoidance)

# Feedback loop

## Critical element

- Not just “release and forget”
- First attempt, needs feedback
- Researchers want reassurance

## Closing the loop

- Researchers access the data on a special server (open internet, no RDC)
- No disclosure-avoidance analysis done on results created from SynLBD
- Validation server allows to request validation, release of results using confidential data (offline submission, full disclosure-avoidance)

# Access to SynLBD

## Key goals

- Easier (very easy) access for researchers: average project approval within 2 (TWO) week
- Quick turnaround on validation (depends on complexity)
- See also SIPP Synthetic Beta (SSB)

# Application

## Process to gain access

- Abstract of a project
- Description of the variables needed
- Application decisions based solely on feasibility

# Validation

### Validation is easy

if the analysis runs error-free on the SDS, then researchers can request that programs be run against the confidential data. All such analyses are reviewed by Census Bureau Disclosure Review Officers, and approved output is provided to both the researchers as well as to the Statistics of Income (SOI) Program at the United States Internal Revenue Service (IRS).