Using partially synthetic microdata to protect sensitive cells in business statistics

Javier Miranda\textsuperscript{2} Lars Vilhuber\textsuperscript{1}

\textsuperscript{1}Labor Dynamics Institute, ILR, Cornell University, United States
\textsuperscript{2}Center for Economic Studies, U.S. Census Bureau, United States

October 2015
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Helsinki
Funding

- Vilhuber’s work is partially funded by NSF Grant #1042181 and #0941226.
- This work is part of the Census Bureau’s LBD Initiative.

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.”[*]

Statistics at great detail on

- job creation and destruction
- establishment births and deaths
- firm startups and shutdowns

by establishment and firm characteristics (age, size, location)
Business Dynamic Statistics (BDS)

www.census.gov/ces/dataproducts/bds/

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
Currently 62 very detailed tables
LBD-BDS complex
Business Microdata at the Census Bureau

LBD Provenance

- Business Register (SSEL)
  - Enumeration frame
  - New establishments
  - Sample frame
  - Economic Census (quinquennial)
    - Possible linkage
  - Longitudinal Business Database (LBD)
    - Possible linkage
    - Reserved
  - Restricted
  - County Business Patterns (CBP)
  - Annual Survey of Manufactures (ASM) tables
  - Economic Census (quinquennial) tables
  - Synthetic LBD (SynLBD)
  - Business Dynamics Statistics (BDS)

Jarmin and Miranda (2002)
Business Microdata at the Census Bureau

LBD Provenance

Business Register (SSEL)

Economic Census (quinquennial)

Annual Survey of Manufactures (ASM)

County Business Patterns (CBP)

Annual Survey of Manufactures (ASM) tables

Possible linkage

Sample frame

Possible linkage

New establishments

Enumeration frame

Possible linkage

same exclusions

Jarmin and Miranda (2002)

Longitudinal Business Database (LBD)

Synthetic LBD (SynLBD)

Business Dynamics Statistics (BDS)

restricted public-access

Tabulations

Microdata
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

The graph displays the extent of suppression for different variables.

- **Y-axis**: Percent suppressed cells
- **X-axis**: Variables (all, isz, sic, sz, age, st, szsc, iszsc, ageisz, agesc, agesz, iszst, szst, agest)

Legend:
- **Blue**: Any suppression
- **Red**: JC by entrants
- **Green**: JC by continuers
Business Microdata at the Census Bureau

LBD Provenance

Business Register (SSEL)

Economic Census (quinquennial)

Annual Survey of Manufactures (ASM)

County Business Patterns (CBP)

Annual Survey of Manufactures (ASM) tables

Economic Census (quinquennial) tables

Synthetic LBD (SynLBD)

Longitudinal Business Database (LBD)

Restricted

Public-access

Jarmin and Miranda (2002)

Some exclusions

Enumeration frame

New establishments

Sample frame

Possible linkage
Purpose of SynLBD

The SynLBD is

- synthetic establishment (and soon firm) microdata
- derived from confidential Longitudinal Business Database (LBD, [5])
- designed to facilitate researcher access to establishment microdata (LBD) (see http://vrdc.cornell.edu/sds)
- 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 [9]
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]
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
Goal is two-fold

Retro-active utility
A mechanism that can fill in existing suppressions.

Improving disclosure avoidance going forward
Evaluate future disclosure avoidance mechanisms:
  ▶ Suppression
  ▶ This proposition
  ▶ Noise infusion (not here)
Analytic validity

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

Figure 3: Share of Employment by Industry Sector and Year, 1976-2000
Figure 8: Job Creation Rate by Year, LBD vs Synthetic

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

Establishment employment $e_{jt}$.

**Example**

$$\begin{align*}
    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} \quad (1) \\

    jcbirth_{jt} &= \begin{cases} 
        e_{jt} - e_{jt-1} & \text{if } e_{jt} > 0 \text{ and } e_{jt-s} = 0 \forall s \geq 1 \\
        0 & \text{otherwise}
    \end{cases} \quad (2)
\end{align*}$$
Synthetic values
Synthesized version of variable \( x_{jt} \) is denoted \( \tilde{x}_{jt} \).

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 \( k_t(j) = k' \).
Aggregations
Generically in capital letters:

\[ E_t = \sum_{j=1}^{J} e_{jt}, \]  

(3)

Aggregations across establishments having characteristics \( k' \) at time \( t \)

\[ X_{k't} = \sum_{j \in K'_t} x_{jt} \]  

(4)
Suppression rules

for (aggregate) variable $X$ are captured by $I_t^X$, such that the releasable variable $X^{(0)}$ under the current regime can be described by

$$X_{k't}^{(0)} = \begin{cases} X_{k't} & \text{if } I_{kt}^X = 1 \\ \text{missing} & \text{otherwise} \end{cases}$$

(5)
We can now express the simple “drop-in” algorithm, leading to the released variable $X^{(i)}$, as:

\begin{align*}
\text{BDS}^{(in)}
\text{if}\ I_t^X = 0 & \quad \text{then} \\
X^{(i)}_{k't} & = \tilde{X}_{k't} \\
\text{else} & \\
X^{(i)}_{k't} & = X_{k't} \\
\text{end if}
\end{align*}
Algorithm 1: Weighted Drop-in

\[ s^* = \min_{s \in [0, n]} \text{s.t. } l_{t-s}^X = 0 \]

if \( n > 0 \) and \( \exists s^* \) then

\[ X_{k't}^{(i)} = \frac{s^*}{n} X_{k't} + (1 - \frac{s^*}{n}) \tilde{X}_{k't} \]

else if \( n = 0 \) and \( l_t^X = 0 \) then

\[ X_{k't}^{(i)} = \tilde{X}_{k't} \]

else

\[ X_{k't}^{(i)} = X_{k't} \]

end if
Algorithm 2

Similar idea, at microdata level
Replace sensitive establishments with synthetic establishments.

Smooth the replacement

- per-establishment weight $w_{js} \in [0, 1]$, applied to the observed data, that increases from 0 in $t$ to 1 in $t + n$,
- a per-establishment weight $\tilde{w}_{js}$, applied to the synthetic data, that decreases from 1 in $t$ to 0 in $t + n$,
Algorithm 2

BDS\(^{(ii)}\)

**Algorithm 2: Forward-longitudinal**

Compute: \(X_{k't} = \sum_{j \in K_{t}'} x_{jt}\)

Compute: \(I_{t}^{X}\)

if \(I_{t}^{X} = 0\) then

// Suppression condition met for cell \(k'\)

Assign all \(j \in K_{t}'\) to \(J_{k't}^{-}\) for \(t \leq s \leq t + n\)

Assign all \(j \in \tilde{K}_{t}'\) to \(J_{k't}^{+}\) for \(t \leq s \leq t + n\)

end if

Compute:

\[X_{k't}^{(iiw)} = \sum_{j \in \{K_{t}' \cap J_{k't}^{+}\}} \tilde{w}_{jt} \tilde{x}_{jt} + \sum_{j \in K_{t}' \land j \in J_{k't}^{-}} w_{jt} x_{jt} + \sum_{j \in K_{t}' \land j \notin J_{k't}^{-}} x_{jt}\]
Careful treatment of border cases

- Setting $n = 0$ is similar to the "Drop-in" case, but margins add up.
- Setting $w_{js} = 0$ for $s \in (t, t + n]$ simply replaces real establishments with synthetic establishments, no phase-in.
- Synthetic establishments that are in cell $k'$ in $t$ but are in cell $k''$ in $t + 1$: should they receive $\tilde{w}_{jt+1} > 0$?
Analysis

- We implemented Algorithm 1 and 2 for Business Dynamics Statistics (BDS) tabulations by establishment age and size ($bds_e_agesz$).
- Variations of $w$ and $n$
- About 26% of all cells have some suppression
- Here: variable, “Job Creation by establishment births” ($job\_creation\_births$) and “Job Creation by establishment continuers” ($job\_creation\_continuers$)
Protection: From Kinney et al

The comparison is for individual establishments, not within cells.

Figure 13: Histogram: Percent Distance Between Actual and Synthetic Employment
Cell-wise comparison

Criteria for cell-wise comparison

- Differences in count of establishment in a cell
- Differences in values of cells
Cell-wise comparison

Released vs. Synthetic BDS: pct_job_creation

Miranda, Vilhuber
SynBDS
Cell-wise comparison

Released vs. Synthetic BDS: pct_job_creation_births

year2
Cell-wise comparison

Released vs. Synthetic BDS: `pct_job_creation_continuers`
Analytic validity: time-series

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

Metrics

- number of missing time-series estimates/feasible regressions
- the number of significant coefficients for the first lag $\rho_1$ 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 $\rho_1^{(s)}$ and $\rho_1^{(i)}$,
- interval overlap measure $J_k$ [6]
Consider the overlap of confidence intervals $(L, U)$ for $\rho_1$ (estimated from the confidential data) and $(L^*, U^*)$ for $\rho_1^*$. Let $L^{\text{over}} = \max(L, L^*)$ and $U^{\text{over}} = \min(U, U^*)$. Then the average overlap in confidence intervals is

$$J_k^* = \frac{1}{2} \left[ \frac{U^{\text{over}} - L^{\text{over}}}{U - L} + \frac{U^{\text{over}} - L^{\text{over}}}{U^* - L^*} \right]$$

We then average $J_k^*$ over all estimated AR(2) regressions.
### Analytic validity: Percent missing

Table: Analytic validity: Feasibility of AR(2) regressions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number feasible</th>
<th>$X_{k't}^{(s)}$</th>
<th>$X_{k't}^{(0)}$</th>
<th>$X_{k't}^{(i)}$</th>
<th>$X_{k't}^{(in)}$</th>
<th>$X_{k't}^{(ii)}$</th>
<th>$X_{k't}^{(iiw)}$</th>
<th>$X_{k't}^{(*)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>estabs</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>estabsentry</td>
<td>64</td>
<td>59.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>jobcreation</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>jobcreationbirths</td>
<td>90</td>
<td>25.6</td>
<td>18.9</td>
<td>13.3</td>
<td>13.3</td>
<td>1.1</td>
<td>2.2</td>
<td>1.1</td>
</tr>
<tr>
<td>jobcreationcontinuers</td>
<td>81</td>
<td>0</td>
<td>6.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
### Analytic validity: Percent significant

**Table:** Analytic validity: AR(2) regressions with significant parameters

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\rho_1$</th>
<th>$\rho_1^{(s)}$</th>
<th>$\rho_1^{(0)}$</th>
<th>$\rho_1^{(i)}$</th>
<th>$\rho_1^{(in)}$</th>
<th>$\rho_1^{(ii)}$</th>
<th>$\rho_1^{(iwi)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp</td>
<td>0.256</td>
<td>0.2</td>
<td>0.256</td>
<td>0.256</td>
<td>0.256</td>
<td>0.256</td>
<td>0.256</td>
</tr>
<tr>
<td>estabs</td>
<td>0.267</td>
<td>0.178</td>
<td>0.267</td>
<td>0.267</td>
<td>0.267</td>
<td>0.267</td>
<td>0.267</td>
</tr>
<tr>
<td>estabsentry</td>
<td>0.109</td>
<td>0</td>
<td>0.063</td>
<td>0.078</td>
<td>0.078</td>
<td>0.109</td>
<td>0.109</td>
</tr>
<tr>
<td>jobcreation</td>
<td>0.178</td>
<td>0.1</td>
<td>0.178</td>
<td>0.178</td>
<td>0.178</td>
<td>0.178</td>
<td>0.178</td>
</tr>
<tr>
<td>jobcreationbirths</td>
<td>0.078</td>
<td>0.015</td>
<td>0.068</td>
<td>0.09</td>
<td>0.115</td>
<td>0.067</td>
<td>0.08</td>
</tr>
<tr>
<td>jobcreationcontinuers</td>
<td>0.21</td>
<td>0.111</td>
<td>0.184</td>
<td>0.16</td>
<td>0.247</td>
<td>0.173</td>
<td>0.173</td>
</tr>
</tbody>
</table>
### Analytic validity: Coverage

**Table:** Analytic validity: AR(2) regressions: Coverage

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\rho_1^{(s)}$</th>
<th>$\rho_1^{(0)}$</th>
<th>$\rho_1^{(i)}$</th>
<th>$\rho_1^{(in)}$</th>
<th>$\rho_1^{(ii)}$</th>
<th>$\rho_1^{(iwi)}$</th>
<th>$\rho_1^{(iin)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp</td>
<td>88.9</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>estabs</td>
<td>88.9</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>estabsentry</td>
<td>92.3</td>
<td>90.6</td>
<td>90.6</td>
<td>90.6</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>jobcreation</td>
<td>82.2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>jobcreationbirths</td>
<td>89.6</td>
<td>91.8</td>
<td>91</td>
<td>89.7</td>
<td>97.8</td>
<td>97.7</td>
<td>98.9</td>
</tr>
<tr>
<td>jobcreationcontinuers</td>
<td>76.5</td>
<td>100</td>
<td>81.5</td>
<td>87.7</td>
<td>87.7</td>
<td>88.9</td>
<td>86.4</td>
</tr>
</tbody>
</table>
### Analytic validity: Overlap

**Table: Analytic validity: AR(2) regressions: Interval overlap**

<table>
<thead>
<tr>
<th>Variable</th>
<th>$J_k^{(s)}$</th>
<th>$J_k^{(0)}$</th>
<th>$J_k^{(i)}$</th>
<th>$J_k^{(in)}$</th>
<th>$J_k^{(ii)}$</th>
<th>$J_k^{(iin)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp</td>
<td>83.4</td>
<td>99.4</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>estabs</td>
<td>80.4</td>
<td>97.6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>estabsentry</td>
<td>78.7</td>
<td>82.6</td>
<td>82.6</td>
<td>82.6</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>jobcreation</td>
<td>73.3</td>
<td>94.4</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>jobcreationbirths</td>
<td>72.9</td>
<td>80.9</td>
<td>81.5</td>
<td>79.9</td>
<td>91.9</td>
<td>91.9</td>
</tr>
<tr>
<td>jobcreationcontinuers</td>
<td>70.7</td>
<td>92.6</td>
<td>77.5</td>
<td>81.6</td>
<td>85.1</td>
<td>85.3</td>
</tr>
</tbody>
</table>
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])
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
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


BDS  Business Dynamics Statistics