Looking back on three years of Synthetic LBD Beta

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

August 2013, World Statistical Congress

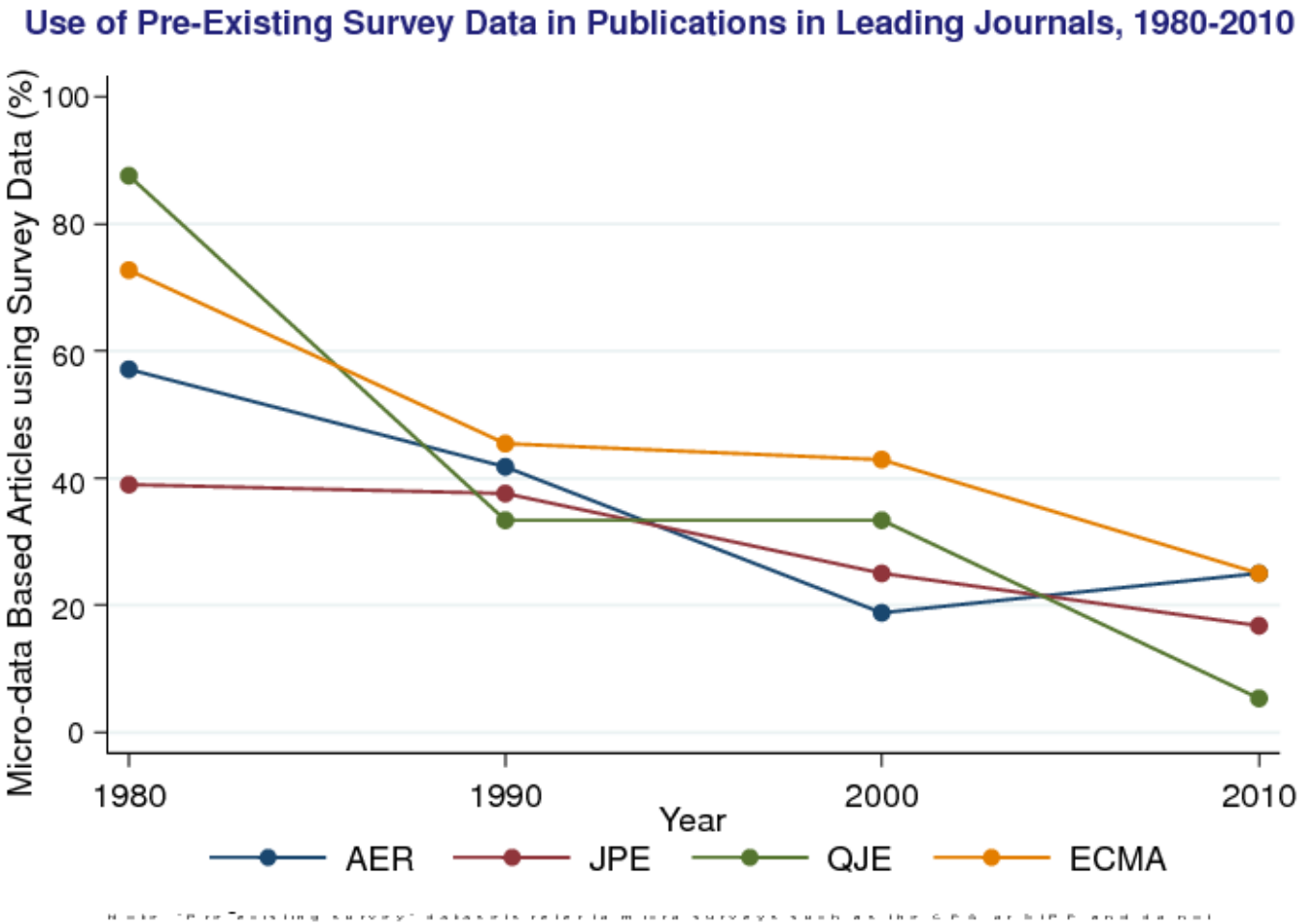
Disclaimer

- Vilhuber's work is partially funded by NSF Grant #1042181.
- For more information, see goo.gl/eyrv7w

Disclaimer

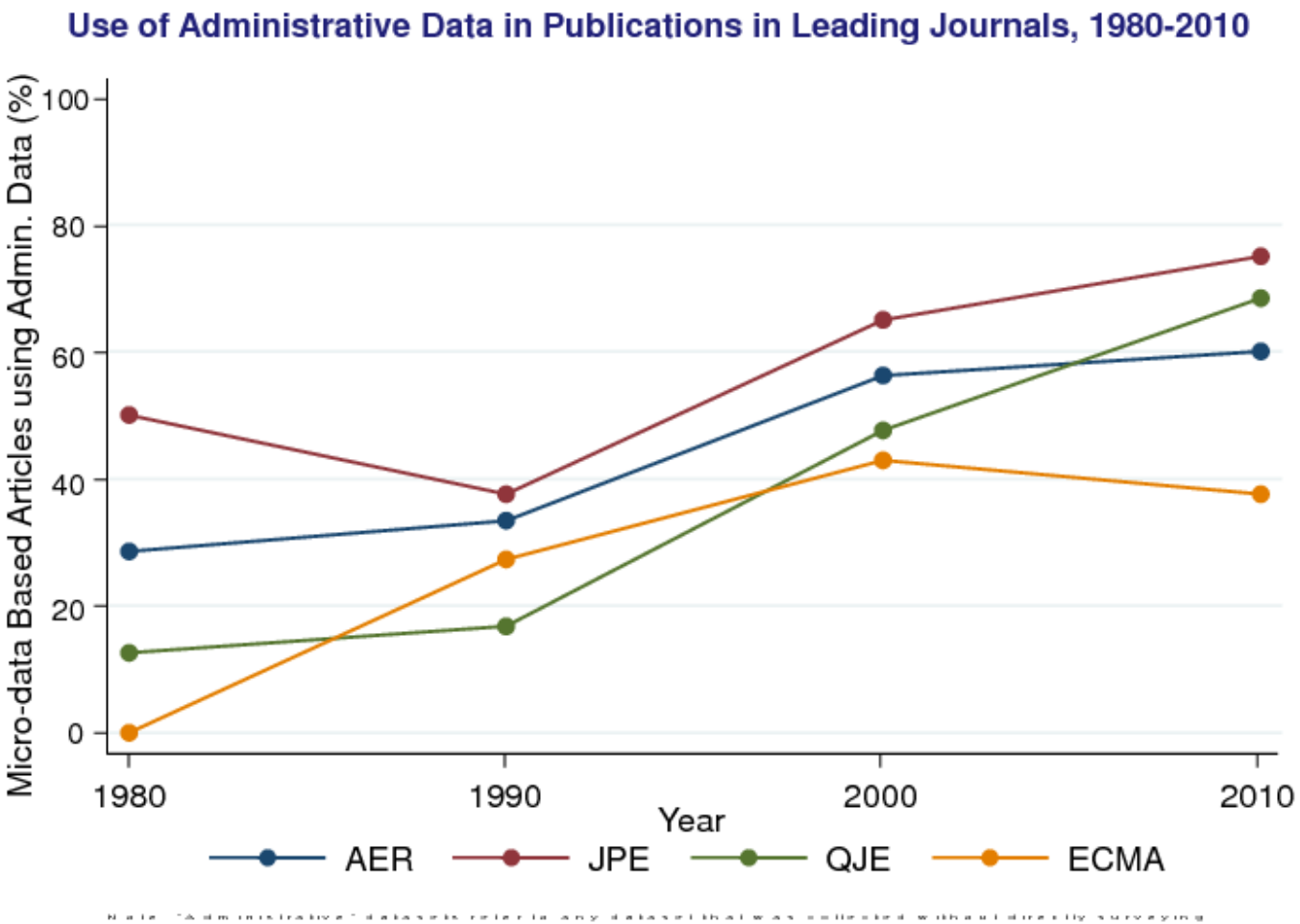
This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion. Any findings, conclusions or opinions 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.

Decline in the use of classic public-use data



include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.

Increase in the use of administrative data in economics



individuals (e.g., scanner data, stock prices, school district records, social security records).
Sample excludes studies whose primary data source is from developing countries.

A problem

Increased use of restricted-access data

- Today's young scholars pursue research programs that mandate inherently identifiable data
  - Geospatial relations,
  - Exact genome data,
  - Networks of all sorts,
  - Linked administrative records
- These researchers acquire authorized, generally unfettered, restricted access to the confidential, identifiable data and perform their analyses in secure environments.

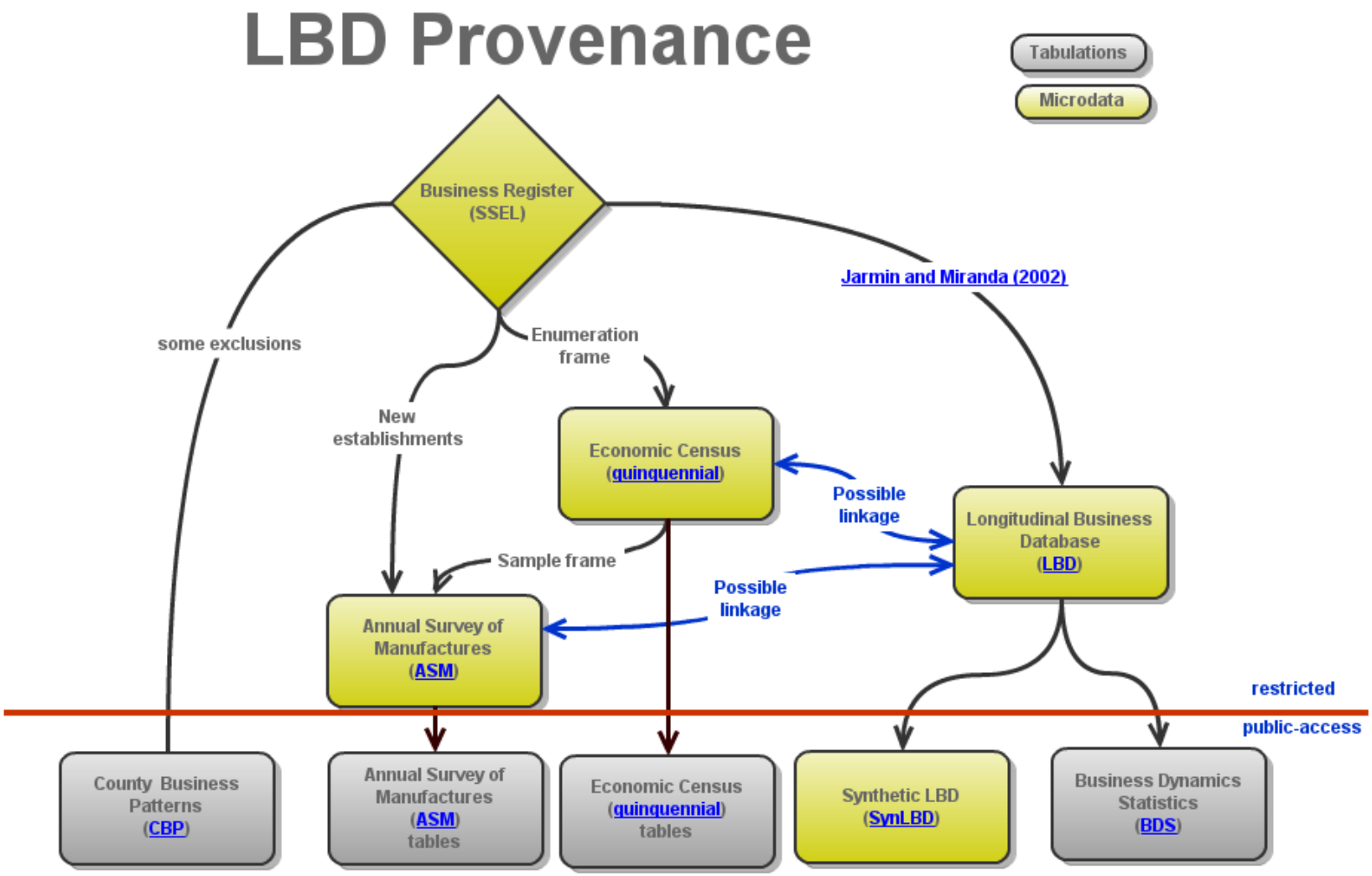
Access to restricted-access data

- Archiving (curation) of input data is complicated
- Knowledge discovery is complicated
- Access is complicated

Economic (business) datasets

- 71% of datasets used by research projects in the U.S. Census Bureau's RDC are business (economic) datasets
- Primarily establishment-based records from the Economic Censuses and Surveys, the Business Register, and the Longitudinal Business Database (LBD)
- They form the core of the modern industrial organization studies [3, 8] as well as modern gross job creation and destruction in macroeconomics [2, 4].
- But there are no public-use micro-data for these establishment-based products
- Exception: Synthetic LBD [1, 6]

Business Microdata in the United States



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

Purpose of SynLBD

The SynLBD is

- designed to facilitate researcher access to establishment microdata

Purpose of SynLBD

The SynLBD is

- designed to facilitate researcher access to establishment microdata
- while preserving the confidentiality of establishment/business data.

Purpose of SynLBD

The SynLBD is

- designed to facilitate researcher access to establishment microdata
- 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 [5]

Purpose of SynLBD

The SynLBD is

- designed to facilitate researcher access to establishment microdata
- 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 [5]
  - development of new public-use data products

Purpose of SynLBD

The SynLBD is

- designed to facilitate researcher access to establishment microdata
- 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 [5]
  - development of new public-use data products
  - expanded research access to business microdata

Construction of SynLBD

Complete description

Kinney et al [6]

Challenges

- Distributions of business data are typically much more skewed than those for household or individual data
- public knowledge of the underlying units is greater

→ More difficult to release data that is informative and non-disclosive

- LBD is longitudinally linked, further increasing the difficulty.

Construction of SynLBD

Data elements

- longitudinal establishment identifiers (created using probabilistic matching [7])
- information on birth, death
- employment and payroll over time
- location
- industry
- firm affiliation of employer establishments

Construction of SynLBD

Data elements

- longitudinal establishment identifiers (created using probabilistic matching [7])
- information on birth, death Synthesized
- employment and payroll over time Synthesized
- location
- industry
- firm affiliation of employer establishments

Construction of SynLBD

Data elements

- longitudinal establishment identifiers (created using probabilistic matching [7]) Masked
- information on birth, death Synthesized
- employment and payroll over time Synthesized
- location
- industry
- firm affiliation of employer establishments

Construction of SynLBD

Data elements

- longitudinal establishment identifiers (created using probabilistic matching [7]) Masked
- information on birth, death Synthesized
- employment and payroll over time Synthesized
- location Suppressed
- industry Released
- firm affiliation of employer establishments

Construction of SynLBD

Data elements

- longitudinal establishment identifiers (created using probabilistic matching [7]) Masked
- information on birth, death Synthesized
- employment and payroll over time Synthesized
- location Suppressed
- industry Released
- firm affiliation of employer establishments $\rightarrow$ next version

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).

Success so far

- **25** researchers from **21** different institutions in **3** different countries have requested access (May 2013)
- Researchers include doctoral students, and researchers who later intend to apply for a more complex project in the Census Bureau's RDC
- 3 researchers have requested validation (May 2013), but more by now
- Among the rejected projects, quite a few asked for **firm** identifiers, **NAICS** codes, or a **longer time** series, all of which will be addressed in forthcoming 2.2 or 3.0 releases

Conclusion

Early in the process

- Synthetic datsets open up new opportunities for access to microdata
- Novel but slowly gaining acceptance: 25% of applications received in the 3 months preceding the paper

Limitations

- Impossible to model all possible relationships in the data
- Synthetic data will only be as good as the models that used to synthesize them
- In this regard its use as a stand alone analytical tool should be discouraged unless validated results are provided.

Conclusion

Part of a menu of access options

- Needs to be part of a menu of access options
- Needs input/feedback from users
- Needs to provide validation as part of the process

Thank you.
More info: www.vrdc.cornell.edu/sds/

$Id: Presentation-ISI-SynLBD-subdoc.tex 1298 2013-0


J. M. Abowd and L. Vilhuber. (2010) Synthetic data server. [Online]. Available: http://www.vrdc.cornell.edu/sds/


S. J. Davis, J. C. Haltiwanger, and S. Schuh, *Job creation and destruction*. Cambridge, MA: MIT Press, 1996.


T. Dunne, M. J. Roberts, and L. Samuelson, "The Growth and Failure of U.S. Manufacturing Plants," *Quarterly Journal of Economics*, vol. 104, no. 4, pp. 671–698, 1989.


J. Haltiwanger, R. S. Jarmin, and J. Miranda, "Who creates jobs? Small vs. large vs. young," Center for Economic Studies, U.S. Census Bureau, Working Papers 10-17, Aug. 2010. [Online]. Available: http://ideas.repec.org/p/cen/wpaper/10-17.html


J. Haltiwanger, L. Lynch, and C. Mackie, Eds., *Understanding Business Dynamics: An Integrated Data System for America's Future*. The National Academies Press for the National Research Council, 2007.


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


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


G. S. Olley and A. Pakes, "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, vol. 64, no. 6, pp. 1263–1297, November 1996. [Online]. Available: http://www.jstor.org/stable/2171831