This document was generated: December 26, 2018 at 11:59 PM

Synthetic Longitudinal Business Database

(28 variables)
Last update to metadata: 2016-11-11 11:47:26 (auto-generated)
Document Date: January 6th, 2014

Codebook prepared by: Lars Vilhuber

Data prepared by: United States Department of Commerce, Bureau of the Census , Duke University , and Cornell University, Labor Dynamics Institute

Principal Investigator(s): United States Department of Commerce. Bureau of the Census. , Internal Revenue Service. , and Cornell University. Labor Dynamics Institute.

Citation

Please cite this codebook as:
Comprehensive Extensible Data Documentation and Access Repository. Codebook for the Synthetic LBD Version 2.0 [Codebook file]. Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, 2013

Please cite this dataset as:
U.S. Census Bureau. Synthetic Longitudinal Business Database: Version 2.0 [Computer file]. Washington DC; Cornell University, Synthetic Data Server [distributor], Ithaca, NY, 2013

Abstract

In most countries, national statistical agencies do not release establishment-level business microdata, because doing so represents too large a risk to establishments' confidentiality. One approach with the potential for overcoming these risks is to release synthetic data; that is, the released establishment data are simulated from statistical models designed to mimic the distributions of the underlying real microdata. The Synthetic Longitudinal Business Database (SynLBD) is the synthetic data version of the Longitudinal Business Database (LBD), an annual economic census of establishments in the United States comprising more than 20 million records dating back to 1976. More information is available at https://www.census.gov/ces/dataproducts/synlbd/index.html. In this codebook, variables are noted as "blanked" if they are available on the confidential version but have been removed from the synthetic version; "synthetic" if the confidential values have been synthesized and released on the synthetic version.

Datasets

synlbd1997c.dta synlbd1997c.dta (Incomplete URL provided - synlbd1997c.dta) (Stata)

Terms of Use

Access Levels

restricted

No description given

released

No description given

Access Restrictions (Default)

The data can only be used on the VirtualRDC Synthetic Data Server http://www.vrdc.cornell.edu/sds/ at Cornell University. While no SynLBD data downloads are permitted at this time, users do not have to operate behind the Census Bureau firewall to access this server.

Access Requirements

In order to access the Synthetic LBD, users should apply for a free account on the Synthetic Data Server (SDS) housed at the VirtualRDC at Cornell University. Application forms can be found at https://www.census.gov/ces/dataproducts/synlbd/accesslbd.html. Application decisions are based solely on feasibility, determined by evaluating whether the data necessary to conduct the analysis are included on the SynLBD Beta file. Decisions generally occur within 10 business days.
Additional information: https://www.census.gov/ces/dataproducts/synlbd/accesslbd.html

Access Permission Requirements

The SynLBD files have been cleared by the Census Bureau Disclosure Review Board and IRS for use by individuals wihtout Census Bureau Special Sworn Status and outside of Census Bureau facilities. Establishments in the SynLBD are fully synthesized using statistical models, and the SynLBD contains no data from actual establishments. Comparison at the establishment level shows SynLBD data differ substantially from the actual data. Modeling preserves variable relationships while protecting establishment identity.

Citation Requirements

Please use the following language in published work that make use of this dataset: "The creation of the Synthetic LBD was made possible through NSF Grant #0427889. Access to the Synthetic LBD was made possible through NSF Grant #1042181." Please also cite Kinney et al (2011) and use the bibliographic citation for the dataset provided in this document.

Disclaimer

Establishments in the SynLBD are fully synthesized using statistical models, and the SynLBD contains no data from actual establishments. Comparison at the establishment level shows SynLBD data differ substantially from the actual data. Modeling preserves variable relationships while protecting establishment identity. Because the SynLBD has not been fully validated, relationships between SynLBD variables may not correspond to the relationships in the

underlying confidential microdata. Unless validated, there is no guarantee results from the SynLBD reflect results from the underlying confidential data. Researchers are strongly encouraged to request result validation prior to publishing results based on the SynLBD. Validation occurs as part of an internal Census Bureau process to improve current beta data products, and is free, as resources permit. (See https://www.census.gov/ces/dataproducts/synlbd/validatingresults.html)

Contact

For questions regarding this data collection, please contact: ces.synthetic.data.use@census.gov

Additional Information

Methodology

Sampling from posterior predictive distribution

Related Material

I. https://www.census.gov/ces/pdf/SynLBD_Codebook.pdf

Related Publications

I. Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin and John M. Abowd. 2011. CES WP-11-04 In most countries, national statistical agencies do not release establishment-level business microdata, because doing so represents too large a risk to establishments' confidentiality. One approach with the potential for overcoming these risks is to release synthetic data; that is, the released establishment data are simulated from statistical models designed to mimic the distributions of the underlying real microdata. In this article, we describe an application of this strategy to create a public use file for the Longitudinal Business Database, an annual economic census of establishments in the United States comprising more than 20 million records dating back to 1976. The U.S. Bureau of the Census and the Internal Revenue Service recently approved the release of these synthetic microdata for public use, making the synthetic Longitudinal Business Database the first-ever business microdata set publicly released in the United States. We describe how we created the synthetic data, evaluated analytical validity, and assessed disclosure risk.

Variable Groups - Synthetic Longitudinal Business Database

Blanked Variables

Identifiers

Synthetic Variables

Variable Name

lbdnum

Label

(synthetic) LBD Number

Concept

Type

character

Files

F1

Full Description

Longitudinal establishment identifier. Can be used to track establishment units over time.

Groups

Identifiers

Variable Name

emp

Label

(synthetic) March 12 Employment

Concept

Type

numeric

Files

F1

Full Description

Paid employment consists of full and part-time employees, including salaried officers and executives of corporations, who were on the payroll in the pay period including March 12. Included are employees on sick leave, holidays, and vacations; not included are proprietors and partners of unincorporated businesses. Employment refers to paid employment at the establishment where business is conducted. Note, this will correspond to firm employment only for single-unit establishment firms. Reported to the Internal Revenue Service (IRS) on Form 941. In some cases this value is imputed due to missing or invalid data.

Groups

Synthetic Variables

Variable Name

pay

Label

(synthetic) Reported Annual Payroll (in $1,000)

Concept

Type

numeric

Files

F1

Full Description

Total annual payroll includes all forms of compensation, such as salaries, wages, commissions, bonuses, vacation allowances, sick-leave pay, and the value of payments in kind (e.g., free meals and lodgings) paid during the year to all employees. Sum of quarterly IRS Form 941 payroll for the year. Missing or invalid 941 data is replaced with imputed values.

Groups

Synthetic Variables

Variable Name

sic3

Label

(observed, computed) SIC3 code

Concept

Standard Industrial Classification

Type

character

Files

F1

Full Description

Three digit Standard Industrial Classification code.

Summary Statistics

Valid values	all
Invalid values	0

Groups

Identifiers

Variable Name

mu

Label

(synthetic) Single-Multi Identifier

Concept

Type

numeric

Files

F1

Full Description

Indicator for whether the establishment belongs to a firm composed of two or more establishments. A value of 1 indicates the establishment is a member of a firm composed of two or more establishments. A value of 0 indicates the establishment is the only member ofthe firm.

Values (2 total)

0	Establishment is the only member of the firm
1	Establishment is a member of a firm composed of two or more establishments

Summary Statistics

Valid values	all
Invalid values	0

Value Ranges

Value Range

Range: [0 , 1]

Groups

Synthetic Variables

Variable Name

firstyear

Label

(synthetic) First Year Establishment is Observed

Concept

Type

numeric

Files

F1

Full Description

Indicator for the first year the establishment is observed in the data (birth year). This variable is left censored at 1976. In conjunction with LASTYEAR, allows users to quickly determine the tenure of an establishment from any point in the data series.

Summary Statistics

Valid values	all
Invalid values	0
Minimum	1975
Maximum	2000

Value Ranges

Value Range

Range: [1975 , 2000]

Groups

Synthetic Variables

Variable Name

lastyear

Label

(synthetic) Last Year Establishment is Observed

Concept

Type

numeric

Files

F1

Full Description

Indicator for the last year the establishment is observed (death year). This variable is right censored at the last year of the data. In conjunction with FIRSTYEAR, allows users to quickly determine the tenure of an establishment from any point in the data series.

Value Ranges

Value Range

Range: [1976 , 2000]

Groups

Synthetic Variables

Variable Name

act

Label

(blanked) Activity Code.

Concept

Type

character

Files

F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD

Groups

Blanked Variables

Variable Name

bestsic

Label

(blanked) Best SIC code

Concept

Type

character

Files

F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD

Groups

Blanked Variables

Variable Name

bestnaics

Label

(blanked) Best NAICS code

Concept

Type

character

Files

F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD

Groups

Blanked Variables

Variable Name

cbp

Label

(blanked) Indicator: used in County Business Patterns

Concept

Type

numeric

Files

F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD. Indicates that an observation was used in the tabulation of the County Business Patterns

Values (1 total)

Sysmiss

Groups

Blanked Variables

Variable Name

cfn

Label

(blanked) Census File Number

Concept

Type

character

Files

F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD. Links to other Economic microdata

Values (2 total)

0123456789
1234567890

Groups

Blanked Variables

Identifiers

Variable Name	county
Label	(blanked) County FIPS codes
Concept	FIPS code
Type	character
Files	F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD. County FIPS code. It is not possible to compute statistics at the state or county level.

Groups

Blanked Variables

Variable Name

firstflag

Label

(blanked) First Link Flag

Concept

Type

character

Files

F1

Full Description

Indicator for first link. Only available on confidential LBD.

Groups

Blanked Variables

Variable Name

flaga

Label

(blanked) Type of Link Flag

Concept

Type

character

Files

F1

Full Description

Identifies the type of link. Only available on confidential LBD.

Groups

Blanked Variables

Variable Name

flagb

Label

(blanked) Birth-Death-Continuer Link Flag

Concept

Type

character

Files

F1

Full Description

Identifies if the link is for birth/death/continuing establishment. Only available on confidential LBD.

Groups

Blanked Variables

Variable Name

lastflag

Label

(blanked) Last Link Flag

Concept

Type

character

Files

F1

Full Description

Only available on confidential LBD.

Groups

Blanked Variables

Variable Name

lfo

Label

(blanked) Legal Form of Organization

Concept

Type

character

Files

F1

Full Description

Identifies the legal form of the organization. Only available on confidential LBD.

Groups

Blanked Variables

Variable Name

lfo1

Label

(blanked) LFO1

Concept

Type

character

Files

F1

Full Description

Processing variable. Only available on confidential LBD.

Groups

Blanked Variables

Variable Name

mfsic1

Label

(blanked) Most Frequent SIC 1

Concept

Type

character

Files

F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD.

Groups

Blanked Variables

Variable Name

pdiv

Label

(blanked) Processing (Economic) Division Code

Concept

Type

character

Files

F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD.

Groups

Blanked Variables

Variable Name

rapf

Label

(blanked) Reported Annual Payroll Flag

Concept

Type

character

Files

F1

Full Description

Only available on confidential LBD.

Groups

Blanked Variables

Variable Name

recnum

Label

(blanked) SSEL Record Number

Concept

Type

numeric

Files

F1

Full Description

Links to BR. Only available on confidential LBD, Variable not present on Synthetic LBD.

Values (1 total)

Sysmiss

Groups

Blanked Variables

Variable Name	sic
Label	(blanked) Standard Industrial Classification Code
Concept	Standard Industrial Classification
Type	character
Files	F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD. Detailed SIC code

Groups

Blanked Variables

Variable Name

sone

Label

processing variable (to be dropped in future version)

Concept

Type

character

Files

F1

Groups

Blanked Variables

Variable Name	state
Label	(blanked) State FIPS codes
Concept	FIPS state code
Type	character
Files	F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD. It is not possible to compute statistics at the state or county level.

Values (1 total)

Groups

Blanked Variables

Variable Name

toc

Label

(blanked) Type of Operation Code

Concept

Type

character

Files

F1

Full Description

Variable not present on Synthetic LBD, only available on confidential LBD.

Groups

Blanked Variables

Variable Name

yr

Label

(computed) Year

Concept

Type

numeric

Files

F1

Full Description

Implicit in file name, was added on SynLBD 2.0.2

Value Ranges

Value Range

Range: [1976 , 2001]

Groups

Identifiers