Institute for Employment Research

The Research Institute of the Federal Employment Agency

Replicating the synthetic LBD with German establishment data

59th World Statistics Congress

28. August 2013, Hong Kong

Jörg Drechsler

Institute for Employment Research

&

Lars Vilhuber

Cornell University

Synthetic data for save data dissemination

- offers a very high level of data protection
- can preserve analytical validity if models are chosen carefully
- especially attractive for sensitive data
- dissemination of data on businesses is often considered too risky:

  - skewed distributions make identification of single units easy
  - information on businesses in the public domain
  - high benefits from identifying a single unit
  - high probability of inclusion for large establishments

- synthetic data as a viable solution

Criticism against synthetic data

- quality of the data strongly depends on the quality of the models
- generating synthetic data is burdensome
- needs modeling experts that really know the data
- constraints between the variables can further complicate the modeling
- generating the data might take so long that the data is outdated by the time of release
- especially small agencies are currently reluctant to invest in the approach

Some arguments for synthetic data

- sometimes the only realistic solution
- first implementations of the idea started 10 years ago
- everything needed to be developed from scratch
- a lot has been learned in the meantime
- new projects can build on this
- nonparametric modeling tools such as CART can further simplify the modeling task

Project idea

- general idea: illustrate that synthesis code might be reused again on another dataset if the two datasets are similar in structure
- basic setup is simple:

  - Find/Construct a dataset that is comparable to the LBD
  - Run the synthesis models from the LBD on this data and see what happens

- not that simple in practice

  - Difficult to construct the same variables from the German register data
  - SAS package that is a core element of the synthesis code not available at the IAB

- this talk will only focus on the construction of a German version of the LBD (GLBD)

The Longitudinal Business Database

- created from the U.S. Census Bureau's Business Register
- data available from 1976 to 2011
- contains information on:
  - birth
  - death
  - industry
  - location
  - payroll
  - firm affiliation
- in the synthetic version location is not available

The German Employment History Panel (BHP)

- no business register available at the IAB
- all establishment level information is derived by aggregating the German Social Security Data via the establishment id
- BHP is one of the data products derived from the GSSD
- BHP will be the main data source to build the German LBD
- data available from 1976/1992 (Western Germany/Eastern Germany)
- contains detailed information on the personnel structure
- not all the variables available in the LBD are also available in the BHP

Differences between the LBD and BHP

- information whether establishment belongs to multi-unit firm not available
- until 1999 the BHP only contains establishments that had at least one employee covered by social security
- payroll information in the BHP is for the reference date June 30 for each year
- LBD contains yearly payroll

Building the German LBD (GLBD)

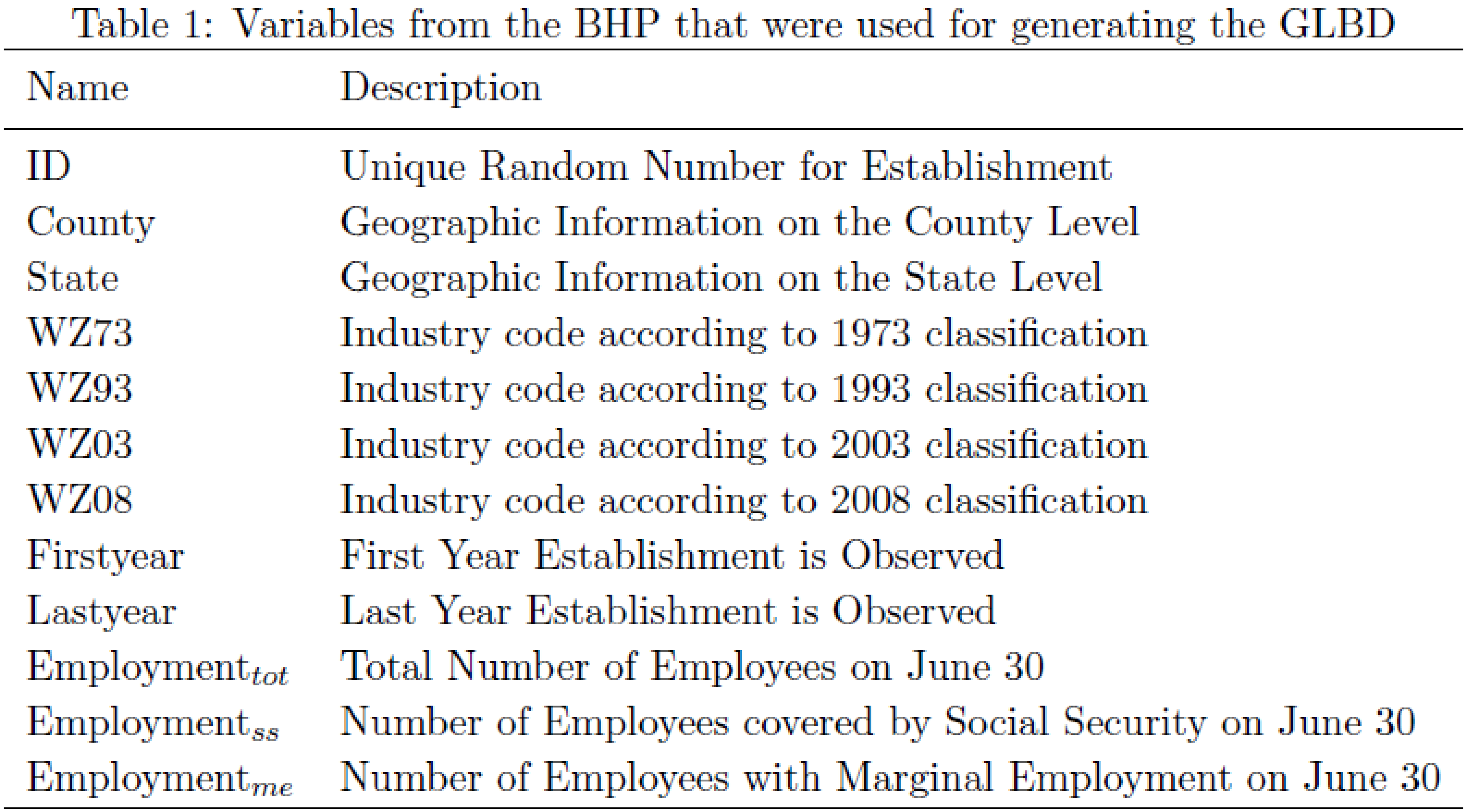

Table 1: Variables from the BHP that were used for generating the GLBD

Name	Description
ID	Unique Random Number for Establishment
County	Geographic Information on the County Level
State	Geographic Information on the State Level
WZ73	Industry code according to 1973 classification
WZ93	Industry code according to 1993 classification
WZ03	Industry code according to 2003 classification
WZ08	Industry code according to 2008 classification
Firstyear	First Year Establishment is Observed
Lastyear	Last Year Establishment is Observed
Employment_{tot}	Total Number of Employees on June 30
Employment_{ss}	Number of Employees covered by Social Security on June 30
Employment_{me}	Number of Employees with Marginal Employment on June 30

Ensuring Consistent Establishment Size

- until 1999 employers only had to report all their employees covered by social security
- since 1999 all employees need to be reported
- significant changes in the data between 1999/2000

  - many establishments report more employees although they didn't grow
  - increase in the number of establishments since establishments with only marginally employed are also included
- to ensure consistency, we

  - subtract the number of marginally employed from total number of employees
  - set all establishment sizes = 0 to missing
  - drop all establishments that never report their establishment size after the adjustments
- final dataset contains 6,916,183 establishments

Generating a unique geographic location and industry code

- geographic location and industry code are constant in the LBD
- this is not true for the BHP
- select the mode of both variables over the lifespan of the establishment
- if two modes exist, the first one is selected
- might be improved

  - select mode randomly
  - weight the years by establishment size

Generating a unique geographic location and industry code

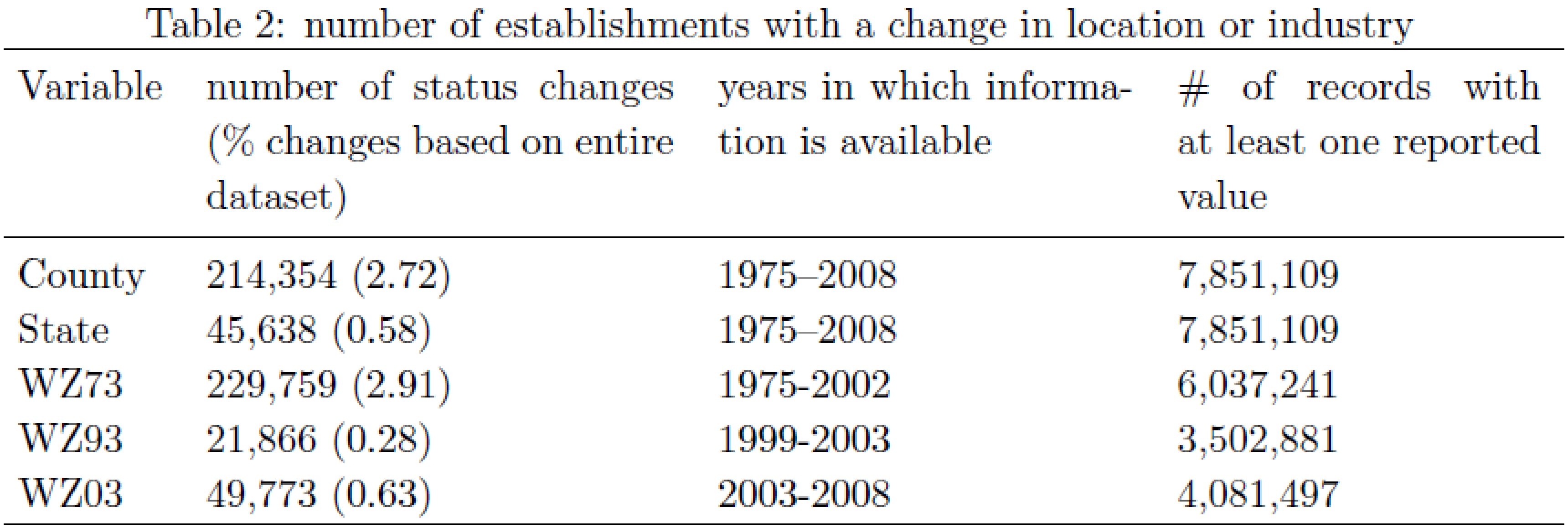

Table 2: number of establishments with a change in location or industry

Variable	number of status changes (% changes based on entire dataset)	years in which information is available	# of records with at least one reported value
County	214,354 (2.72)	1975–2008	7,851,109
State	45,638 (0.58)	1975–2008	7,851,109
WZ73	229,759 (2.91)	1975-2002	6,037,241
WZ93	21,866 (0.28)	1999-2003	3,502,881
WZ03	49,773 (0.63)	2003-2008	4,081,497

- only a small number of establishments report a change
- even fewer have two or more modes
- we stick with the simple approach

Updating the information on establishment births and deaths

- information on the first/last year establishment is observed is not necessarily equivalent with the birth/death of the establishment

  - data are left and right censored
  - new establishment appears whenever a new establishment id is generated

- new ids are not necessarily equivalent to a new establishment

- several other reasons possible, e.g.

  - change of ownership
  - change in declared industry classification

- these new ids should not be treated as establishment births

- similarly disappearance of ids should not always be treated as establishment deaths

Updating the information on establishment births and deaths

- use the employee flows to identify real births and deaths based on a similar approach by Benedetto et al (2007)
- flow-files generated by Hethey and Schmieder (2010)
- basic idea
  - if (almost) all employees of an exiting establishment work in the same new establishment in the following year this is most likely an id change
  - if (almost) all employees of a new establishment worked in the same establishment in the year before but this establishment still exists in the current year, the new establishment is most likely a spin-off.
- the files also contain suggestions how the observed births and deaths should be categorized
- we use the suggested classification

Updating the information on establishment births and deaths

- all birth and death categories are treated as births and deaths
- establishments with unknown status are treated according to the information in the BHP
- spin-offs are left unchanged
- id changers are merged and employment and payroll information is aggregated
- industry and geographic information are based on the mode of the observed variables in the linked record
- Some establishments identified as dead reappear in the data later
- Since number is small (3,941 establishments) we ignored this

Adding payroll information

- payroll information only available at a reference date in the BHP
- possible to derive yearly payroll by aggregating the information from each employee that was ever employed in a specific establishment for a given year
- aggregated yearly payroll information also available from another project at the IAB
- only includes the payroll of all full time employees
- for almost 230,000 (3.3%) records no payroll information is available
- payroll information for all establishments in the BHP based on all employees from the underlying administrative data could be incorporated in the future

Plans for the imputation of the industry classification

- we plan to impute the industry classification whenever it is missing due to the changes in the reporting system
- we will use a simple probabilistic crosswalk based on the methodology used for the LEHD ECF (Abowd et al., 2009)
- relies on doublecoding for at least some periods
- uses for example P(WZ08|WZ03= wz03) to impute WZ08 whenever it is missing
- subsequent to the imputation, the modal industry across all years for each establishment is computed

Next steps

- once the core GLBD is created, the SynLBD data synthesizing algorithms will be applied
- remaining disclosure risks will be evaluated
- data will be made available at the IAB and on Cornell University's Synthetic Data Server
- will allow comparative studies between the US and Germany
- similar data products from other countries should be added in the future

Thank you for your attention