

# Comments on BDS papers

Lars Vilhuber

Cornell University

# Breaking silos

- Papers are a good example of the benefits of breaking down data silos
  - Foreign trade statistics and business dynamics
  - Patents and businesses
  - Improving links by using worker flows for business registers

# A primer for non-Americans

- Isn't the Business Register where every business is listed?
  - Why aren't patents associated with them?
  - Why don't exports simply register with them?
- Why can there be workers who don't have quarterly wage records (=unemployment insurance contributions)?

# A primer for non-Americans

- Business Register is compiled from tax filings, but is NOT a central register of the numbers where such things happen
  - Other registers: IRS (where EINs are assigned), BLS (QCEW), private entities (Dun and Bradstreet), state agencies (including state business registers, which are often true registers)
- Yearly income taxes are filed on earnings (=W2) which is not the same purpose as collecting quarterly UI taxes (=LEHD), nor is it the same purpose as assessing Medicare Taxes (also reported on W2...)

# Moving to 21<sup>st</sup> Century Statistics

- Leveraging data that is already collected
  - But not necessarily with the current linkage in mind!
  - Not just an American problem
- Rethinking how statistics are collected
  - The administrative record collection as a survey
  - Holistic view where possible (within the federal government), coordination elsewhere (between states and federal government)
  - ... thus breaking down silos

# Specifics on papers

BDS XF

# BDS-XF

- BDS by export status → why not importing status?
- Complete description

	<i>Exporter</i>	<i>Non-exporter</i>
<i>Importer</i>	JC, JD	JC, JD
<i>Non-importer</i>	JC, JD	JC, JD



# BDS-XF Industry classification

- Impact of NAICS reclassification based on switching import/export status
  - LBD uses a constant (modal) NAICS (here as well?)
  - What is the likelihood that a new exporter also changes primary activity?

# BDS-XF

- Reporting requirements: why do firms report export transactions?
  - What does it mean that the EIN does not match to an employing firm?
  - What does it mean that the EIN does not conform to legal EIN values? (<9, >9 digits)
    - Field is actually the USPPI
    - Legal values are (DUNS, EIN, or Foreign Entity).
  - note: DUNS is a legal value, and maps into different identifier set! Is also 9-digit
  - non-match may mean not in EIN universe!

# BDS-XF matching technology

- Separation of exact vs. fuzzy matching
  - Names available on BDS-XF
  - If yes, also do matching exercise based on name, validate using EIN
  - Noted earlier: non-match may mean that number is not a EIN → match based on EIN lookup is invalid!
  - Consider treating “EIN” field as a text field that can be miscoded, and to which fuzzy matching techniques applied
- Uncertainty
  - Deterministic  $t/t+1/t-1$  algorithm yields multiple matches (how many?)
  - Treat as multiply imputed, with normalized weight equivalent to the distance to  $t$
  - Extend to  $k>1$  years, assess robustness of results

# BDS-XF: fuzzy matching

- Used for Canadian data only
- Standardization
- Fuzzy matching using SAS DQ
  - SAS DQ uses sensitivity levels and resultant hashes
  - No good measure of quality of match
- Would like to see
  - Robustness to matches based on the minimum “2-word criterion” (2 out of 5 seems low)
  - Robustness to variations to the sensitivity level to simulate availability of match score
  - Possible robustness to use of string functions (Jaro-Winkler) and possible Fellegi-Sunter match scores

# BDS-XF: fuzzy matching

- What has changed relative to the original LFTTD?

- More generous conditions for “EIN” match?
- Different input data?
- Better?

	Non-CDN/CDN		Non-CDN/CDN (old)	
1993	73	69	72	66
1994	73	70	73	66

- Statistics

- Match rates (count) by arm's length status
- Match rates by t/t+1/t-1
- Rates for 1992 are very low – is the t-1 contribution that strong (absent in 1992)

# BDS-XF statistics

- Ever-exporting status
  - Assess robustness by looking at sliding two-year windows: knowing the universe for 2001-2008, how much would the number of exporting firms change if you were to use 1999-2010 as the basis? Repeat for 2002-2009.
  - How does this result to firms becoming exporters?
  - Use a definition similar to identifier-based birth/death classification: a firm is a non-exporter in  $t$  if it is not observed exporting in  $[t-k, t]$

# BDS Patents

# BDS Patents

- Match rates
  - Would like to hear more about matching technology
  - Sensitivity analysis to matching (assign match probability, include in subsequent tables)
  - 
  - Distribution of patents
    - Encourage adoption of a moving window – patent holder if filed a patent in the past 5 years, but may need longer data (does not need match)



# BDS Patents

- Match rates
  - Would like to hear more about matching technology
  - Sensitivity analysis to matching
    - assign match probability
    - include in subsequent tables
    - Obtain some error bounds
  -

# BDS Patents

- Because of the use of subsidiaries, is there a benefit to leveraging the COS?
- What information is there that “filing entities” are “employers”? If they never employ anybody and just serve as shell companies, QCEW, BR, COS may not be of much use

# BDS Patents in the future

- Collaboration between Patent Office and Census Bureau
- What ability does the Patent Office have to
  - Collect more information about firms
  - Survey patent submitters at time of submission

# BDS Patents

- Distribution of patents
  - Encourage adoption of a moving window – patent holder if filed a patent in the past 5 years, but may need longer data (does not need match)

BDS-HC

# BDS-HC

- Next step: why are there demographic differences in working for dying firms?
  - Location may be important, but not included here

# BDS-demographics

- Obvious comparison, already noted
  - quarterly QWI with yearly BDS-HC
- Question:
  - What does it mean to incorporate quarterly earnings information into yearly BDS-HC?
- More importantly:
  - highlight the value-added to doing these kinds of analyses at the yearly level
  - Linkage is easy
  - Statistics need more work

Thank you