3-27-2012

Presentation: Data access in North America: Current state and future consequences

Lars Vilhuber  
Cornell University, lv39@cornell.edu

William C. Block  
Cornell University, block@cornell.edu

Follow this and additional works at: https://digitalcommons.ilr.cornell.edu/ldi

Thank you for downloading an article from DigitalCommons@ILR.
Support this valuable resource today!
Presentation: Data access in North America: Current state and future consequences

Abstract
Presentation at the First DwB European Data Access Forum.

Keywords
Data access, synthetic data, mobility

Comments
Suggested Citation

Required Publisher's Statement
Copyright held by authors.

This article is available at DigitalCommons@ILR: https://digitalcommons.ilr.cornell.edu/ldi/4
Data access in North America

Current state and future consequences

William C. Block and Lars Vilhuber
Disclaimer:

The opinions expressed in this presentation are those of the authors and not the National Science Foundation, the U.S. Census Bureau, or any other government agency.

No confidential, restricted-access data was used to prepare this presentation.
Caveats

• Economist
• Labor Economist
• Micro-data preferred
• US bias
Classifying North American data

- Access-type
  - Public-use data
  - Contractual access
  - Restricted-access data

- Data source
  - Survey data
  - Administrative data

- Strength of SDL
Degree of detail

Ease of access
Access: Public-use (micro-) data

- Very successful from a usage perspective
- Examples:
  - Current Population Survey (CPS)
  - PSID
  - NLSY
  - Kauffman Firm Survey
- Used in teaching, research, deep scientific corpus
RA: Contractual restriction

- Examples:
  - NLSY (detailed geo)
  - HRS (additional data)
- Some restrictions on usage in exchange for details
- Few constraints in combining with other data
RA: Remote controlled access from anywhere

• Examples:
  – CRADC @ Cornell
  – Data enclave @ NORC
  – Synthetic data server @ Cornell

• Typically still cross-dataset access restrictions even within the same environment

• Reduced ability to combine with other data
RA: Remote execution

• from anywhere
• Examples:
  – NCHS micro data ($)
  – Statistics Canada
  – (implicit in Synthetic Data Server)
• May be limited in complexity of models that can be estimated
Remote access from controlled location
Remote access from controlled location

- Examples:
  - Census, BLS, Canadian RDC
  - Even IAB data (from Cornell)
- Limited access (few locations)
- Long application process
- Limited ability to add additional data
Detail and access

- As detail increases, access restrictions also increase
- What other methods are used?
Trade-off: geographic detail vs. timeliness

• Decennial Census
  – Tract level
  – Limited characteristics

• American Community Survey
  – More person/household characteristics
  – Precision increases with multi-year estimates
Trade-off: geographic detail vs. timeliness

- Current Population Survey
  - Monthly estimates
  - No sub-state estimates (exception: 12 large MSAs)
Data without Boundaries

• Increased access to restricted access data
• Access to data from multiple jurisdictions
• Access to data from multiple “access domains”
• Increasingly detailed public-use data
Increased access to restricted access data

• Expansion of RDC network
  – USA
  – Canada

• Expansion of data accessible in RDC network
  – Agency for Health Care Research (AHRQ)
  – National Center for Healthcare Statistics (NCHS)
Access to data from multiple jurisdictions

• Long-standing access
  – IRS, SSA data in Census RDC, can be combined with Census data sources

• New
  – Multi-state access (education-oriented longitudinal data warehouses)
Not everything is advancement

- BLS, Census, other agencies remain distinct and separate (despite CIPSEA)
- No cross-border access (Canadian data in US or vice-versa)
- Multi-jurisdiction access may be reduced, not increased (state employment agencies at Census Bureau) for research purposes
Access to data from multiple “access domains”

- How to get MUCH public-use data into
  - Census RDC
  - CRADC?
- No data curation other than own data
  - > CCBMR (see our presentation at WDA)
- Synthetic data, more detailed geo data
  - Increased ease of combining data
Other methods

- Increasingly detailed public-use statistics
  - Use of
    - synthetic data
    - new methods of SDL
  - Quarterly Workforce Indicators
  - Business Dynamics Statistics
  - Synthetic SIPP
  - Synthetic LBD
Example: Abowd and Vilhuber (2012)

- “Did the Housing Price Bubble Clobber Local Labor Market Job and Worker Flows When It Burst?” (AEA, PP, 2012)

- Data sources:
  - FHFA's Housing Price Index
  - BLS' National and Local Unemployment Statistics
  - Census Bureau's Quarterly Workforce Indicators
  - Our own national aggregation of those
Why do we do this?
Research lifecycle

Modelling

Critique
Why?

- Accelerate the research cycle
- Increase the body of research for any given data source
- Improve economic/social/demographic/etc. models through more detailed data
Restricted-access data less so
Forecasting Incidence Of Work Limitations, Disability Insurance Receipt, And Mortality In Dynamic Simulation Models Using Social

Melissa Farray (2002)

Downloadable! In examining a number of important research questions related to the reform of the Social Security program, it is helpful to understand patterns of participation in the Disability Insurance (DI) program. DI beneficiaries comprise a large fraction, approximately 15 percent, of the pool of workers who receive Social Security benefits (Social Security Administration, 2001, Table 5 A16). They are a particularly vulnerable group in later life, with poverty rates more than twice as high as those for recipients of retirement or survivor benefits from Social Security (Thompson and Smith, 2002, Table A9-L3c). Those who receive DI also have very different mortality experiences than those who do not (Zayatz, 1999), so careful modeling of the overlap between mortality and disability is essential when trying to determine the lifetime distributional consequences of Social Security reform. In addition, the larger disabled population, consisting of those who report work limitations but do not necessarily receive...
Richness of data is an incredible asset

• Macro economic CGE models rely on a multitude of parameters – dozens, maybe hundreds
• Micro economic (partial equilibrium) models rely on feasible estimation
• New modeling strategies: networking, micro-simulation
Goal of research

• Understanding of economic and social phenomena
  – Better model-based predictions
  – Better experimental analysis
Modelling
Weather modelling
Behind this:

- A set of models
- Computed using observed data, simulations
- National Centers for Environmental Prediction has two 156-node compute clusters running 24/7
- Precision of predictions?
Experiments

- Experiments provide useful data under controlled circumstances
- They are sometimes frowned upon...
Nuclear experiments nowadays
ASC computing environment

• Sequoia next-generation BlueGene/P compute cluster:
  – 98,304 compute nodes
  – 1.6 million processor cores
  – 1.6 PB memory
Bad policy and “experiments” have bad outcomes

Berlin 1923

Zimbabwe
The logical next step?

• If we can simulate...
  – atomic bombs
  – Weather

• Given the right input data (integrated DwB!)

• Can we provide (better) simulations of economic phenomena and policy?
Let's consider labor market mobility...
Sometimes only very little mobility
Sometimes a lot of mobility
Sometimes opportunities next door

May not be included in the data!
... almost certainly for immigrants
Presenting

• The bane of integrated data

Mr. Data-truncation
Not just me.
Sibling locations
Current colleague locations
Sibling locations
Current colleague locations
Past colleague locations
Sibling locations
Current colleague locations
Past colleague locations
It gets worse...

• Siblings in Montana (works in Silicon Valley) and Grenoble (used to live in Egypt)
• Parents somewhere in Europe (long live retirement), with retirement income from two state retirement systems (US and Germany)
Historical data offers some insights

- We can link Tor Janson from Oslo (1880) to his records in the United States
- But we cannot link 21st century Lars Vilhuber
Hourly data available...
And I didn't even mention...

• F...b..k
• G....l.
• Tw.....
This is not the end

- Suppose we solve most of the data access issues
- What kind of data usage models will we see?
Example mobility

- Kennan and Walker (2003, 2011)
- Model determinants of individual location and employment choices along a mobility path
- Computational limitations:
  - 500 HS dropouts
  - State-level choices
  - Only two at any time
  - > 1 day @ 50CPUs to estimate
Models are always a simplification

• But:
  – 5.6 million Americans moved to a different state (IRS SOI, 2008-2009)
  – 7.4 million moved to a different county in the same state
  – 300,000 entered the US, 198,000 left the US
Resources are still limited in RA
but resources exist where the data is not
Some attempts get close

- “Exploring New Methods for Protecting and Distributing Confidential Research Data” at Michigan (Felicia LeClere) is already working in the cloud
- Census Bureau working with network of researchers, working group on next-generation flexible compute architecture within restricted-access environment
Outlook
Consequences of successful DwB

• If you create it (the integrated data environment), they will come
• … but they may wish for more than you can provide
• Successful data integration must also provide the tools for new (pent-up) modelling strategies
The next frontier

- Tera-scale compute resources for the social sciences, using integrated confidential data