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](https://digitalcommons.ilr.cornell.edu/ldi)  
Thank you for downloading an article from DigitalCommons@ILR.  
Support this valuable resource today!

This Article is brought to you for free and open access by the Centers, Institutes, Programs at DigitalCommons@ILR. It has been accepted for inclusion in Labor Dynamics Institute by an authorized administrator of DigitalCommons@ILR. For more information, please contact catherwood-dig@cornell.edu.

If you have a disability and are having trouble accessing information on this website or need materials in an alternate format, contact web-accessibility@cornell.edu for assistance.
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.
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 vs. 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
Public-use data very successful

IDEAS Search

Search page with instructions

Search for: current population survey


Sort by: relevancy title

   Downloadable! This paper examines the 1997 additions to the Current Population Survey education question. These new questions allow researchers to come closer to the "highest grade completed" measure available before 1992. Using the new information, the average imputed "highest grade completed" is one-tenth grade greater and the estimated return to education is 8 to 1 percentage points lower than when using only the post-1991 "highest degree received" question.

2. Discussion and Presentation of the Disability Test Results from the Current Population Survey [58.981%]
   Terence M. McNemar & Stephen M. Miller & Anne E. Polivka (2006)
   Downloadable! In accordance with Executive Order 13078, the Bureau of Labor Statistics, in cooperation with the Employment Rate Measurement Methodology Interagency Working Group, identified the goal of placing a small set of questions within the Current Population Survey (CPS) to measure disability. A set of potential questions was drawn from existing surveys, cognitively tested, and placed in the National Comorbidity Survey (NCS) for testing. Based on an analysis of the NCS data, a set of seven questions was identified. These seven questions were then cognitively tested to ensure that they would work within the CPS context. This question set was placed in the February 2006 CPS for field testing. The two primary goals of the test were to compare the CPS disability rate to that obtained from the NCS, and to evaluate the effect on CPS response rates in the following month. Analysis of the test data revealed a lower overall disability rate as measured in the CPS than in the NCS, with lower positive response rates following.

3. On hourly wages and weekly earnings in the current population survey [58.365%]
   Downloadable (with restrictions)! This paper examines hourly wages and weekly earnings in the Outgoing Rotation Group (ORG) and the March Supplement of Current Population Survey (CPS) from 1998 to 2004. The findings suggest that the ORG contains less errors than the March CPS, and that weekly earnings contain less errors than hourly wages. The paper further finds that earnings differ systematically in the ORG and in the March CPS by gender and education levels.

   Debnath, Bhaskar (2001)
Restricted-access data less so
1. Forecasting Incidence Of Work Limitations, Disability Insurance Receipt, And Mortality In Dynamic Simulation Models Using Social

*Melissa Farreau* (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

Los Angeles County (Los Angeles), Calif.

- Population (2010): 9,818,668
- Population (2000): 9,706,375
- Inbound income per cap. (2010): $23,900
- Outbound income per cap. (2010): $24,400
- Non-migrant income per cap. (2010): $24,600

Graph showing inbound and outbound migration with selected data for different years.
Sometimes opportunities next door

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

• The bane of integrated data

Mr. Data-truncation
Current residence
Current workplace
Historical workplaces
Higher education
Not just me.
Sibling locations
Current colleague locations
Past colleague locations
Sibling locations
Current colleague locations
Past 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

![XSEDE User Portal](image_url)

<table>
<thead>
<tr>
<th>NAME</th>
<th>INSTITUTION</th>
<th>SYSTEM</th>
<th>PEAK TFLOPS</th>
<th>MEMORY GB</th>
<th>STATUS</th>
<th>LOAD</th>
<th>RUNNING JOBS</th>
<th>QUEUED JOBS</th>
<th>OTHER JOBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kraken</td>
<td>NICS</td>
<td>Cray XT5</td>
<td>1174.00</td>
<td>147.00</td>
<td>Up</td>
<td></td>
<td>215</td>
<td>674</td>
<td>334</td>
</tr>
<tr>
<td>Ranger</td>
<td>TACC</td>
<td>Sun Constellation Cluster</td>
<td>579.40</td>
<td>123.00</td>
<td>Up</td>
<td></td>
<td>445</td>
<td>222</td>
<td>368</td>
</tr>
<tr>
<td>Gordon</td>
<td>SDSC</td>
<td>Gordon Cluster</td>
<td>341.00</td>
<td>84.00</td>
<td>Up</td>
<td></td>
<td>7</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Lonestar</td>
<td>TACC</td>
<td>Dell Linux Cluster</td>
<td>302.00</td>
<td>45.00</td>
<td>Up</td>
<td></td>
<td>181</td>
<td>444</td>
<td>50</td>
</tr>
<tr>
<td>Forge</td>
<td>NCSA</td>
<td>PowerEdge C6145 with NVIDIA Fermi M2070</td>
<td>150.00</td>
<td>64</td>
<td>Up</td>
<td>13</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Trevis</td>
<td>SDSC</td>
<td>Appro AMD Magny-Cours Cluster</td>
<td>100.00</td>
<td>20.25</td>
<td>Up</td>
<td></td>
<td>1487</td>
<td>3216</td>
<td>32</td>
</tr>
<tr>
<td>Steele</td>
<td>Purdue</td>
<td>Dell Intel 64 Linux Cluster</td>
<td>60.00</td>
<td>12.40</td>
<td>Up</td>
<td></td>
<td>790</td>
<td>640</td>
<td>81</td>
</tr>
<tr>
<td>Queen Bee</td>
<td>LONI</td>
<td>Dell Intel 64 Linux Cluster</td>
<td>50.70</td>
<td>5.31</td>
<td>Up</td>
<td></td>
<td>28</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Blacklight</td>
<td>PSC</td>
<td>SGI UV</td>
<td>37.20</td>
<td>32.00</td>
<td>Up</td>
<td></td>
<td>67</td>
<td>156</td>
<td>37</td>
</tr>
</tbody>
</table>

*Indicates failure of one or more status tests
Hover mouse pointer over Resource Name, Resource Status, and headings to see additional information.
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