Demand and Preferences for Access to Federal Administrative Data: Results of a Survey

Katherine Smith and Luona Lin
Council of Professional Associations on Federal Statistics
COPAFS’ mission is to advocate on behalf of the users of federal data and statistics. AEA is a member as are 14 other professional associations across (mostly) social science disciplines) and 30 Centers, Institutes, Foundations and Firms that rely on federal statistics.

We act as a go-between for users with federal agency producers of federal statistics. We also educate program and policy decision makers about the uses, value, and outcomes of using federal statistics.
Most statistical agencies are operating under severe budget constraints, with no end in sight for budget tightening. As a result, many are (among other actions) exploring the use of administrative data as a complement to or substitute for survey data.
Federal Administrative Data

- Non-survey data used to run federal programs
- Some broad examples:
  - Birth and death records (vital statistics)
  - Tax records
  - Welfare program participation data
  - Unemployment claims
  - Program cost data

Not the direct output of statistical surveys

NOTE: Most administrative data are NOT collected by statistical agencies
Common Challenges in Using Administrative Data

- Statistical agency access – legal interpretations and lack of institutional incentives
- Agency infrastructure – policies, procedures, hardware
- Administrative data quality – fitness for use (timeliness, relevance, accuracy, match rates, etc.)
- Researcher access to data -- Documentation, access modes, access program

OMB M-14-06 is recent guidance issued to help overcome some of these challenges. It asks for reports on administrative data from across all agencies of the U.S. government. It facilitates agency-to-agency transfer of appropriate data. Available for research only via statistical agencies, if then. The climate is good for COPAFS to work with agencies to broaden access to researchers and provide more assistance in providing access Program agencies to Stat agencies to researchers
Council of Professional Associations on Federal Statistics (COPAFS)

- COPAFS mission
- Status of federal statistical agencies
- COPAFS project

Project funded by Alfred P. Sloan foundation, with contributions from AEA and the Economic Research Service. The goal is to expand social scientists’ access to administrative data for research in order to improve the bases for sound and informed public policy design and implementation. The current focus is on economic research.
COPAFS Project Objectives

- Develop an inventory of federal administrative data access processes, procedures and tools
- Determine what administrative data sets are highest priority for economists
- Facilitate dialogues between researchers and the agencies inhibiting researcher access to what we learn are priority data

Develop an inventory of federal administrative data access processes, procedures and tools
Determine what administrative data sets are highest priority for economists
Convene economic researchers and the representatives of agencies collecting high priority administrative data to identify specific actions and begin working jointly on options for expanded access that satisfy the concerns of each; and
Create the basis for an ongoing structure for studying, examining options and taking actions regarding confidential access to administrative data for social scientific research

This presentation focuses on the survey.

Motivation for survey:
- As John and Amy made clear, access to administrative data is essential for policy relevant economic (ala Card, Chetty, Saez)
- With limited resources, it is important to know what administrative data sets are most important to, in this case, economic researchers, in order to design interventions that make a substantive contribution
- We surveyed a sample of AEA members to determine
  - What admin data sets are relevant to their research
  - Of relevant data, which are most important?
  - Of relevant data, which used and, if not used, why not
Survey of AEA Members

- Universe = 6,000
- Usable sample size = 729 completed responses
- 85-percent of sample indicated primary or secondary work activity is research = 614 for data analysis
- 2/3rds of sample from academia
Youngish. Also many respondents who gave us feedback were from small colleges unlikely to have an economic librarian on staff.
Fields that are “overrepresented”: Agriculture, Labor, Education, Health, Urban.

Fields that are “underrepresented”: Financial Economics, Industrial Organization, International Economics, and Mathematical Methods
Many on this list could be valuable to economic research that is not labor economics.

Now, on to results, -- Start with the subset from this list that was most relevant to all who answered any question in this grouping.
Most Relevant (to all) Labor and Demographic Administrative Data

- Quarterly Census of Employment and Wages from BLS
- Longitudinal Employer-Household Dynamics Data from the Census
- Earnings and Employment Data from the Social Security Administration
- OSHA Inspection and Enforcement Data

Reddish font indicates data that are highly relevant and highly important to all respondents who found any data in the category relevant.
Black font indicates data that are relevant to respondents, but do not rank highly in “importance.”
Blue font indicates data that are very important, but to a minority of those responding in the data category.
Largest proportion of total choices that are either relevant, important, or both, to all those who made a selection in the Welfare category.
(For each of Agriculture, Education, and Energy, only one choice was relevant and important)
Most Relevant and Important (to all) Health Administrative Data Sets

• CMS National Health Expenditures Data
• NCHS National Vital Statistics
• AHQR Health Care Utilization Data
• CMS Medicare Claims Data
Most Relevant and Important (to all) International Development Admin. Data

• BEA data on Foreign Direct Investment
• BEA International Accounts data
• Foreign Exchange Rates Data from the Federal Reserve.
Most Relevant and Important (to all) Natural Resources Administrative Data

• Cropland data by National Agricultural Statistics Service
• USGS Land Cover and Land Use Data
• USGS Water Resources Data
• National Marine Fisheries Service: Commercial and Recreational Fisheries Statistics
Most Relevant and Important (to all) Urban, Regional and Transportation Administrative Data

• County and Zip Code Business Patterns Data from Census
• BTS Air Carrier Statistics,
• EPA Superfund Database
• HUD Fair Market Rents Data
Most Relevant and Important (to all) Macroeconomic Admin. Data

- BEA National Income and Product Accounts
- IRS corporate and individual tax statistics
- Department of Treasury Interest Rate Statistics
- Government Expenditure and Receipts data from the Federal Reserve
Most Relevant and Important (to all) Business Administrative Data

- Census Business Register Data and Longitudinal Business Database
- Consumer Credit Data from the Federal Reserve
- SEC Electronic Records and Filings Data
Top 10 relatively stable across different weighting schemes (no. respondents; identification of respondent’s concentration; different relative weights for relevance vs importance)

All of the data sets are relevant across multiple areas of concentration. We ran a little logit analysis to determine the extent to which an individual respondent’s choice of a data set as relevant is related to the individual’s area of concentration.

Data Sets that are relatively more “cross-border” than others:
- Department of Treasury: Interest Rate Statistics
- Bureau of Economic Analysis: National Income and Product Accounts (NIPA) Data
- Internal Revenue Service (IRS): Corporate Tax Statistics and Individual Tax Statistics

### Ten Most Relevant and Important Data Sets

<table>
<thead>
<tr>
<th>Rank</th>
<th>Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BLS Quarterly Census of Employment and Wages</td>
</tr>
<tr>
<td>2</td>
<td>BEA National Income and Product Accounts</td>
</tr>
<tr>
<td>3</td>
<td>Census: Longitudinal Employer-Household Dynamics Data</td>
</tr>
<tr>
<td>4</td>
<td>Census: County and Zip Code Business Patterns</td>
</tr>
<tr>
<td>5</td>
<td>Social Security Admin. Earnings and Employment</td>
</tr>
<tr>
<td>6</td>
<td>OSHA Enforcement Data (Inspection Data)</td>
</tr>
<tr>
<td>7</td>
<td>IRS: Corporate and Individual Tax Statistics</td>
</tr>
<tr>
<td>8</td>
<td>Census Business Register and Longitudinal Business Data</td>
</tr>
<tr>
<td>9</td>
<td>Department of Treasury: Interest Rate Statistics</td>
</tr>
<tr>
<td>10</td>
<td>OSHA: Work-related Injury and Illness Data and Worker Fatalities/Catastrophes Report (FAT/CAT)</td>
</tr>
<tr>
<td>Rank</td>
<td>Data Sets</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>Bureau of Labor Statistics: Quarterly Census of Employment and Wages</td>
</tr>
<tr>
<td>3</td>
<td>Census: Longitudinal Employer-Household Dynamics Data</td>
</tr>
<tr>
<td>4</td>
<td>Census: County and Zip Code Business Patterns Data</td>
</tr>
<tr>
<td>5</td>
<td>Social Security Administration: Earnings and Employment Data</td>
</tr>
<tr>
<td>6</td>
<td>Occupational Safety and Health Administration (OSHA): Enforcement Data (Inspection Data)</td>
</tr>
<tr>
<td>7</td>
<td>Internal Revenue Service (IRS): Corporate Tax Statistics and Individual Tax Statistics</td>
</tr>
<tr>
<td>8</td>
<td>Census Bureau: Business Register Data and Longitudinal Business Database</td>
</tr>
<tr>
<td>9</td>
<td>Department of Treasury: Interest Rate Statistics</td>
</tr>
<tr>
<td>10</td>
<td>OSHA: Work-related Injury and Illness Data and Worker Fatalities/Catastrophes Report (FAT/CAT)</td>
</tr>
</tbody>
</table>

A check mark indicates statistical significance of relation between respondent’s area of concentration and the choice of the data set as relevant.
## Most Relevant and Important Data Sets With Access Issues

<table>
<thead>
<tr>
<th>Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLS Quarterly Census of Employment and Wages</td>
</tr>
<tr>
<td>Census: Longitudinal Employer-Household Dynamics Data</td>
</tr>
<tr>
<td>Social Security Admin. Earnings and Employment</td>
</tr>
<tr>
<td>IRS: Corporate and Individual Tax Statistics</td>
</tr>
<tr>
<td>Census Business Register and Longitudinal Business Data</td>
</tr>
</tbody>
</table>
QCEW

• 34 percent who thought it relevant had not used it, and of those, some indicated that non-use was due to the restricted nature of the data

• Although there is a very detailed Public Use Data set for QCEW, the microdata require that the researcher make a proposal to BLS and, if approved, may use only at BLS in Wash., DC
  – [http://www.bls.gov/bls/blsresda.htm#eligibility](http://www.bls.gov/bls/blsresda.htm#eligibility)
LEHD Data

Longitudinal Employer–Household Dynamics (LEHD) data are the result of a partnership between the Census Bureau and U.S. states to provide high quality local labor market information and to improve the Census Bureau's economic and demographic data programs. LEHD data are based on different administrative sources, primarily Unemployment Insurance (UI) earnings data and the Quarterly Census of Employment and Wages (QCEW), and censuses and surveys. Firm and worker information are combined to create job level quarterly earnings history data, data on where workers live and work, and data on firm characteristics, such as industry. Access to these data will only be granted to qualified researchers on approved projects with authorization to use specific data sets. All researcher access to restricted—use data occurs at one of the secure Census Research Data Centers (RDCs). The table below lists person, job and establishment based data available at the RDCs.

All LEHD data files except the Business Register Bridge are by state. A subset of states have data available at the RDCs. The list of states can be found on page 16 of LEHD Infrastructure Files in the Census RDC - Overview (2.5 MB). This document provides detailed information about the LEHD data. In general, LEHD data are available from 2000 onwards. The availability of historical data prior to 2000 varies by state and data set. In the Years column of the table below, the start year is the year in which the state(s) with the earliest data has (have) data available for that data set. The latest year of data available at the RDCs is 2011. Some LEHD data contain Federal Tax
Information (FTI). Use of LEHD data containing FTI requires approval by the Internal Revenue Service (IRS).

In addition to the restricted-use data available at the RDCs, LEHD creates public-use data sets and online tools. Quarterly Workforce Indicators (QWI) data and the online tools QWI Explorer and the LED Extraction Tool contain workforce statistics by demography, geography, and industry for each state. LEHD Origin-Destination Employment Statistics (LODES) and the OnTheMap web application have partially synthetic data on where workers live and work. These data and online tools have statistics for quarters up to about one year ago and include data for all states that have joined the LED Partnership.
## LEHD

LEHD Restricted-Use Microdata

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Data Type</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Register Bridge (BRB)</td>
<td>Establishment</td>
<td>1990–2011</td>
</tr>
<tr>
<td>Employer Characteristics Files (ECF)</td>
<td>Establishment – Quarter</td>
<td>1989–2011</td>
</tr>
<tr>
<td>Employment History Files (EHF)</td>
<td>Job (Person–Firm)</td>
<td>1985–2011</td>
</tr>
<tr>
<td>Geocoded Address List (GAL)</td>
<td>Establishment</td>
<td>1990–2011</td>
</tr>
<tr>
<td>Individual Characteristics Files (ICF)</td>
<td>Person</td>
<td>1985–2011</td>
</tr>
<tr>
<td>Quarterly Workforce Indicators (QWI)*</td>
<td>Establishment – Quarter</td>
<td>1990–2011</td>
</tr>
<tr>
<td>Unit-to-Worker (U2W)</td>
<td>Job (Person–Establishment)</td>
<td>1990–2011</td>
</tr>
</tbody>
</table>
Social Security Earnings and Employment Data

- Public Use files are available
- Access to microdata: A research plan, confidentiality pledges and data protection activities are required, as per most federal microdata access procedures
- But, the SSA’s unique relation with several research consortia offer unique pathways to the use of microdata or synthetic data for research
IRS Tax Statistics

• Public use data are very broad and general
• Microdata access is possible, but highly limited
Census Business Register and Longitudinal Business Data

- Restricted use, with standard access procedures and access limited to Research Data Centers
So, based on AEA respondents, we would get the biggest bang for the buck in working with BLS, Census, SSA and IRS – each of which is or has a statistical agency,

Lots of creative options

But,,,,,for policy analysis...
“Important” Data with Limited Demand

- USDA, Food and Nutrition Services: Commodity Supplemental Food Program Data
- USDA, Farm Services Agency: administrative data on program participants
- USDA, Food Safety and Inspection Service: Inspection and Enforcement Activity Data
- USDA: Web Based Supply Chain Management Reports Data
- National Marine Fisheries Service: Commercial and Recreational Fisheries statistics
- Bureau of Transportation Statistics (BTS): Air Carrier Statistics and International Air Travel Statistics (I-92 Form)
- EPA: Superfund Sites (CERCLIS database)
- Department of Housing and Urban Development: Fair Market Rents Data
- Department of Veteran’s Affairs: Veterans Benefits Administration Reports
- Securities and Exchange Commission (SEC): Electronic Records and Filings Data
Conclusions

• There is a general lack of awareness among AEA members of the breadth of administrative data sets available for research
• A few restricted data sets are both relevant and important across numerous economist areas of concentration
• Data for welfare program evaluation and linkage remains a challenge