Conducting Research in the Penn State Census Research Data Center

Chris Galvan, PhD
Administrator, PSUCRDC
Center for Economic Studies
U.S. Census Bureau

Sept. 20, 2014

Disclaimer

Any opinions and conclusions expressed herein are those of the author and do not necessarily represent the views of the U.S. Census Bureau.

What are Census Research Data Centers (RDCs)?

- Secure computing labs where qualified researchers conduct approved statistical analysis on non-public data.
- Non-public data are collected by Census Bureau and other government agencies.
- Established through an agreement between Census Bureau and a local research community.

Census Research Data Center Locations

The Penn State Census Research Data Center

- Located in the Paterno Library
  - 206 E. Paterno

Types of Restricted Data Available

- Economic Data
  - Microdata on firms and establishments
  - Business Register data
  - Transactions data
- Demographic Data
  - Survey data on individuals and households
  - Administrative data on individuals (SSA)
- Employer-Employee Jobs Data (LEHD)
  - Data on employees linked with data on employers
- Health Data
  - National Center for Health Statistics
  - Agency for Healthcare Research & Quality

Advantages of Restricted Data

- Vast number of economic datasets that are not publicly available at the micro level
- Census datasets can be linked together
- Census datasets can be linked to external data
- More detailed level of geographic identifiers
- Very little top or bottom-coding

Economic Datasets

- Annual Survey of Manufactures
- Census of Construction
- Census of Finance and Insurance
- Census of Manufactures
- Census of Mining
- Census of Real Estate
- Census of Retail
- Census of Services
- Census of Transportation
- Census of Wholesale
- Survey of Business Owners
- Commodity Flow Survey
- Import and Export Transactions
- Annual Capital Expenditures Survey
- Business Register
- Longitudinal Business Database
- Manufacturing Energy Consumption Survey
- Medical Expenditure Panel Survey, Insurance Component
- National Employer Survey
- Pollution Abatement Costs and Expenditures
- Quarterly Financial Reports
- Research and Development Survey
- Survey of Manufacturing Technology
- Annual Retail/Wholesale Trade Surveys
- Kauffman Firm Survey

Economic Example #1

"Identifying Agglomeration Spillovers: ...Winners and Losers of Large Plant Openings" Greenstone et al. *Journal of Political Economy* (2010)

- The authors find that five years after a new plant opens, TFP of incumbent plants in winning counties is 12% higher than TFP of incumbent plants in losing counties suggesting that there are substantial spillovers to agglomeration.
- The authors linked external data on new plant openings to the **Business Register**. They then linked this data to the **ASM** and **CM** and calculated TFP for each plant located in the "winning" counties and "losing" counties.

Economic Example #2

"Demand Fluctuations in the Ready-Mix Concrete Industry" Collard-Wexler *Econometrica* (2013)

- The author investigates the role of demand shocks in the ready-made concrete industry under a model of oligopolistic markets. He finds that smoothing local short-term demand (by smoothing public spending over 5 years) expands the market, increasing the number of establishments by 39%.
- The author primarily used the **LBD** to examine establishment entry and exit. He also linked in data on sales and assets from the **ASM** and **CM.**

Demographic Datasets

- 1970, 1980, 1990, 2000, 2010 Decennial Surveys
- American Community Survey (annual microdata, 1996-2011)
- Current Population Survey (selected supplements)
- Survey of Income and Program Participation
- American Housing Survey
- National Crime Victimization Survey
- Administrative data from SSA
- National Survey of College Graduates
- National Longitudinal Mortality Study

Demographic Example #1

"Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes" Bayer, Ross, and Topa *Journal of Political Economy* (2008)

- The authors find a significant effect of social networks on hiring, especially among those with similar socio-demographic characteristics.
- Use **Decennial** survey with census block of residence and census block of work to look for social hiring networks.

Demographic Example #2

"Recent Trends in Top Income Shares in the USA: Reconciling Estimates from the March CPS and IRS Tax Return Data." Burkhauser et al. *Review of Economics and Statistics* (2012)

- The authors find that recent changes in income inequality are driven by changes in the top one percent of households.
- They used the internal, non-top coded **CPS** data instead of public-use data and matched income definitions to those studies using IRS data.

Penn State Research

- Monnat: “Hispanic Health Care Access and Utilization in Different Geographic Locations”
  - Restricted-use SIPP files contain county IDs, enabling a more detailed examination of health care access and use across Hispanic destination types.

LEHD

- Links individuals to place of employment
- Based on unemployment insurance administrative records
- Available on a state-by-state basis
- Quarterly data starting in 1990
- "tracks" a person based on their place of employment
  - Establishment (i.e. the place of work) is exact for single plant companies
  - Establishment is assigned for all others (using geography and industry to improve matches)

LEHD Example #1

"Human Capital Loss in Corporate Bankruptcy" Graham et al., CES Working Paper (2013)

- This paper quantifies the "human costs of bankruptcy" by estimating employee wage losses induced by bankruptcy filings and finds annual wages decline by 30% one year after bankruptcy.
- Authors use an external database on bankruptcy filings and merge it to the **Business Register**. They then link these firms to their respective workers in the **LEHD** and examine how their earnings change over time.

LEHD Example #2

"Job-to-Job Flows in the Great Recession" Hyatt and McEntarfer *American Economic Review* (2013)

- The authors develop job-to-job flow measures and find sharp drops in job mobility compared to previous recessions and high earning penalties for transitions that include nonemployment spells.
- The authors use the **LEHD** to develop these job to job measures. They also analyze examine industry transitions during the Great Recession.

Health Data in the RDC

- These data are collected by:
  - National Center for Health Statistics (NCHS)
  - Agency for Healthcare Research and Quality (AHRQ)
- Dual mission: to provide broad access to health data and statistics, while protecting the privacy of respondents

What Types of NCHS Data?

National Health Status Surveys

- National Health and Nutrition Examination Survey (NHANES) I, II, and III
- National Health Interview Survey (NHIS)
- Longitudinal Study on Aging I and II (LSOA)
- National Survey of Family Growth
- National Survey of Children's Health
- National Survey of Early Childhood Health
- National Survey of Children with Special Health Care Needs
- National Asthma Survey

National Health Care Surveys

- National Ambulatory Medical Care Survey
- National Hospital Ambulatory Medical Care Survey
- National Survey of Ambulatory Surgery
- National Hospital Discharge Survey
- National Nursing Home Survey (NNHS)
- National Home and Hospice Care Survey
- National Employer Health Insurance Survey
- National Health Provider Inventory
- National Immunization Survey

Vital Statistics

- Mortality and Multiple Mortality
- Birth
- Fetal Death
- National Death Index
- Marriage and Divorce

What Types of NCHS Data?

Linked Data Sets

- Linked mortality data: NHIS, NHANES LSOA II, NNHS
- Linked Medicare Enrollment and Claims data: NHIS, NHANES, LSOA II
- Linked Social Security Administration Data: NHIS, NHANES, LSOA II, NNHS
- Linked EPA data

What's Restricted?

- Geographic variables (state, county, or metropolitan area)
- Most dates (date of interview, date of death, date of birth)
- Income and employment data (industry codes)
- Specific diagnoses (ICD-9 codes are generally coarsened)
- Details about facilities (accreditation, payments, number of employees)
- Some information about children and adolescents (e.g. height and weight, depression, behavior problems, and drug use)
- Some information about race, ethnicity, and country of origin
- Contextual data (nearest hospital, % of population with diploma)
- Sample design variables (necessary for estimating variances)

NCHS Examples

- NHANES-Mortality: Individuals in the lowest quartile of bone mineral density had a 1.53 times greater risk of death than those in top quartile (Mussolino and Gillum 2008)
- NHIS: Military service increased the probability of smoking, but in later adulthood, the effect attenuated and did not lead to large negative health effects (Eisenberg and Rowe 2009)

Penn State Research

- Van Hook et al.: "Neighborhood Context, Weight, and Weight-Related Behaviors among Mexican American Children"
  - Restricted-use NHANES files contain county and tract IDs, as well as detailed place of birth, food security, and occupation.
  - Preliminary findings suggest that Mexican-origin children who live in neighborhoods with higher concentrations of foreign-born Mexicans had lower levels of dietary acculturation, but dietary acculturation was higher among children living in neighborhoods with higher percentages of non-Hispanic whites and persons with low educational attainment.

What types of AHRQ Data?

- Medical Expenditure Panel Survey (MEPS) files include:
  - Household Component
  - Provider Component
  - Insurance/Employer Component
  - Nursing Home Component (1996 only)
  - Area Resource File
  - Two-year two panel file
  - MEPS-NHIS linked data
- Only Household Component and portions of Provider Component are publicly available

MEPS – Household Component

- Large, longitudinal study of civilian non-institutionalized population
- Collects information from all members of household
  - Demographics
  - General health status and health problems
  - Health events within survey period
  - Healthcare utilization and expenditure
  - Insurance coverage

MEPS Insurance & Provider Components

- Follows up with medical providers and pharmacists to confirm household-reported diagnoses and procedures
- Collects data on total charges, payments, and sources
- Follows up with employers and collects data on insurance options
- Collects data on characteristics of firms

AHRQ Examples

- MEPS-IC and Economic Census: Employers who offer health insurance have 25% greater productivity and 32% higher pay, all other variables held constant (McCue and Zawacki 2006)
- MEPS-IC: Higher employee contributions for insurance are associated with lower enrollment (Cooper and Vistnes 2006)

How to Access the RDC

- Develop proposal
  - Different guidelines for Census data vs. NCHS/AHRQ guidelines
  - For Census data, begin by contacting Chris with a brief project summary
- Submit proposal for agency review
  - Census (and agency sponsors)
  - NCHS/AHRQ
- Obtain Special Sworn Status (SSS)
- Pay one-time fee for NCHS/AHRQ data

Guidelines for Census proposal

- Proposal must meet Basic Requirements
  - Need for *Non-Public* data
  - Maintains Confidentiality
  - Must emphasize statistical models vs. tabular output
  - Feasibility
  - Describes Census Benefits (LEGAL REQUIREMENT)
  - Scientific Merit
- Work with Census Administrator to Craft Final Proposal

Working in the RDC lab

- All analysis conducted in the RDC lab:
  - Data located on server in Maryland
  - Access data via thin client terminals located in cubicles
- No internet access or personal computers allowed in lab
- Statistical software available: SAS, Stata, R, Matlab, GIS, Sudaan, etc.
- Agency reviews output before releasing
  - Penalty for disclosure is $250,000 and/or 5 yrs in prison (inadvertent or otherwise)

Timeframe – "Patience is a Virtue"

- Census Data
  - Plan on 9 to 12 months
  - Title 13 (Census approval only) vs. Title 26 (Census & IRS approval)
- NCHS/AHRQ Data
  - Timeline dependent on agency approval process
  - Census approval NOT required
- Special Sworn Status
  - 3 additional months for your "security clearance"

Contact Information

- People:
  - Chris Galvan, PSUCRDC Administrator
    chris.a.galvan@census.gov, 301-763-0342
  - Mark Roberts, PSUCRDC Executive Director
    mroberts@psu.edu, 814-863-1535
- Resources:
  - PSUCRDC website:
    http://www.psurdc.psu.edu/
  - CES Working Paper Series:
    https://ideas.repec.org/s/cen/wpaper.html