The Kansas City Research Data Center—A Resource for Researchers

For Presentation at:
IPSR
Donna K. Ginther
Professor Department of Economics
Director, Center for Science, Technology & Economic Policy at the
Institute for Policy & Social Research
What is a Census Research Data Center?

- A secure environment where qualified researchers conduct approved statistical research using non-public data collected by Census and other agencies.
- Established through an agreement between Census and a local research community.
What is a Census Research Data Center?

Provided for through:

- **Title 13, sec. 9 & 214 (Census)**
  - use of the data only by employees and sworn officers of Census

- **Title 26, subtitle f, Ch. 61, Sub Ch. B, § 6103(j) (IRS)**
  - allows sharing of information with Commerce for statistical use

- **CIPSEA (Title 44, Ch. 35, Sub Ch. I, § 3501)**
  - use of confidential data must have a statistical purpose
  - provides for severe penalties for disclosing confidential data
Census Research Data Center Locations

- Seattle, WA
- Berkeley, CA
- Stanford, CA
- Los Angeles, CA
- Kansas City, MO
- Chicago, IL
- Minnesota
- Cornell, NY
- Baruch, NY
- Maryland
- Triangle & RTI Int’l
- Boston, MA
- College Station, TX
The Kansas City Research Data Center

- Located in the Federal Reserve Bank of Kansas City
  - 1 Memorial Drive, Kansas City, MO

- Consortium Members:
Advantages of Access to Non-public Data

- Data not available elsewhere
  - Establishment level business data
  - Linked worker-firm data (LEHD)
- Detailed geo-spatial variables
  - Census block group
  - Longitude and latitude
- Virtually no top or bottom coding of variables
  - earnings, income
Advantages, cont.

- Can link many Census datasets with each other
  - Current Population Survey with LEHD
  - Economic Census with Survey of Business Owners
- Can link to other non-Census data
  - NCHS data with ACS
  - Kauffman Firm Survey of new business formation
The Nuts & Bolts of Doing Research in a RDC

- Contact the RDC Administrator
  - Census employee responsible for submitting proposals
  - KC's Administrator: To be hired

- Access and fees (project based)
  - No cost for Consortium Member researchers
  - Non members: $20,000/project "public use review" fee
  - NCHS data: everyone pays a data extraction fee
The Nuts & Bolts of Doing Research in a RDC

- Develop proposal
  - Scientific merit (similar to any other grant proposal)
  - Demonstrate need for restricted data, feasibility
  - Benefit to Census (legal requirement of CIPSEA)
    - statistical purpose for access to restricted data
The Nuts & Bolts of Doing Research in a RDC

- Not for the feint of heart, or the 4th year Ph.D.
  - Census reviews if using Census data (Title 13)
  - IRS reviews if using tax data (Title 26) - lengthiest reviews
  - Health data does not require Census or IRS review
    - ...unless being merged with Census or tax data
  - Background check
    - required by Title 13, Special Sworn Status
The Nuts & Bolts of Doing Research in a RDC

- Conduct research inside the KCRDC lab
  - non-restricted area has workstations and internet access
  - restricted area will have 8 - 10 cubicles
    - thin client terminals connected to data stored in Virginia
    - Linux operating system, SAS, Stata, R, Matlab, word processing
The Nuts & Bolts of Doing Research in a RDC

- Conduct research inside the KCRDC lab
  - No personal computers allowed in lab
  - No paper or output allowed outside of lab (keep in locker)
  - Census public use review required to present results
    - sign and significance allowed
    - penalties for disclosure: $250,000 and/or 5 yrs prison (inadvertent or otherwise)
    - discussion of specific results allowed only inside RDC (even among co-authors on a project)
Why Go the all the Trouble?

- Exclusivity and expanded research questions
  - Only 15 RDC labs across the country
    - however, can easily work with colleagues close to other RDCs
  - Detailed micro data expands research agenda
- The Consortium will fund 1 RA for RDC use
- Funding opportunities through NSF
  - Strong collaboration between NSF and Census
  - NSF-Census Research Network (NCRN)
Types of Data Available

- **Economic Data**
  - Economic Census
  - Business Register data
  - Establishment data vs. Firm data
  - Transactions data

- **Demographic Data**
  - Census data (Decennial Census, ACS, SIPP, etc.)
  - Am. Housing Survey, Nat'I Crime Victimization Survey
Types of Data Available

- LEHD – Longitudinal Employer/Household Data
- Health Data
  - National Center for Health Statistics
  - Agency for Healthcare Research & Quality
- Coming next year: Bureau of Labor Statistics Data
  - Establishment Surveys
  - Price data
Types of Data Available

- Economic Data
  - Economic Census
    - Establishments surveyed every 5 years
    - Manufactures: 1967-2012
      - Construction: 1972-2012
    - Retail & Wholesale Trade: 1977-2012
    - Mining: 2002-2012
    - Services: 1977-2012
    - Trans, Comm, Utilities: 1997-2012
Types of Data Available

- Economic Data
  - Economic Census
  - Business Register data

  Standard Establishment List
  COMPUSTAT-SSEL Bridge
  Form 5500 Bridge
  Longitudinal Bus. Database
  Integrated Bus. Database
  Ownership Change Database
Types of Data Available

- Economic Data
  - Economic Census
  - Business Register data
  - Establishment data vs. Firm data

  Annual Survey of Manufactures
  Current industrial Reports
  Survey of Manufacturing technology
  Manuf. Energy Consumption
  Survey of Pollution Abatement costs and expenditures
  National Employer Survey
  Survey of Plant Capacity Utilization
Types of Data Available

- **Economic Data**
  - Economic Census
  - Business Register data
  - Establishment data vs. Firm data
    - Capital Expenditure Survey
    - Annual Retail Trade & Service Surveys
    - Business Expenditure Survey
    - Enterprise Summary Report
    - Quarterly Financial Report
    - Survey of Business Owners
    - Business Research Development & Innovation
    - Kauffman Firm Survey
Types of Data Available

- **Economic Data**
  - Economic Census
  - Business Register data
  - Establishment data vs. Firm data
  - **Transactions data**
    - Foreign Exports & Imports Trade Data
    - Commodity Flow Survey
Sample Research Questions: Economic Data

  Using the *Survey of Business Owners*, the research underscores the crucial role of human and market capital and the limits of social capital in facilitating entrepreneurs' economic success.

- "Identifying Agglomeration Spillovers: ...Winners and Losers of Large Plant Openings" Greenstone et al. *JPE* (2010)
  Matching SSEL with the ASM, and CM, 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. This effect is larger for incumbent plants that share similar labor and technology pools with the new plant.
Types of Data Available

- **Demographic Data**
  - **Census data**
    - Decennial Census
    - American Community Survey
    - March Current Population Surveys
    - NLS Original 4 Cohorts
    - Survey of Income & Program Participation
  - **American Housing Survey**
  - **National Crime Victimization Survey**
Sample Research: Demographic Data

Using the 2000 Decennial Census Long Form, this paper finds that different socioeconomic characteristics at the census tract level are significantly related to the occupational segregation of Chinese men and women.

Using the American Housing Survey, this study compares the relationship between official crime rates in census tracts and resident perceptions of crime, and finds that tract violent crime is the strongest predictor of residents’ perception of crime.
Types of Data Available

- LEHD – Longitudinal Employer/Household Data
- Health Data
  - National Center for Health Statistics
  - Agency for Healthcare Research & Quality
LEHD Public Tool: On The Map
Sample Research: LEHD

- "Increased Accountability, Teachers' Effort, and Moonlighting" Champion (Stanford), Nov. 2010
  This paper uses the **LEHD** to investigate moonlighting behavior of teachers. This analysis reveals a substantial reduction in public school teacher moonlighting for teachers in schools exposed to performance pressure from a state accountability system. A reasonable interpretation of this evidence is that moonlighting is *a good indicator of (lack of) effort and that accountability systems do* actually create incentives for teachers to make effort in their school.

- "Comparing Measures of Earnings Instability Based on Survey and Administrative Reports" Juhn & McCue
  *CES WP#10-15 (2010)*
  The authors compare earnings levels and jobs in the **SIPP** and administrative data (**LEHD**) to identify differences in measured earnings instability. Much of the difference in earnings instability across the two data sets comes from the lower tail of the earnings distribution.
Types of Data Available

- **Health Data**
  - **National Center for Health Statistics**
    - National Health Status Surveys
    - Linked Data across NCHS, SSA, EPA
    - Vital Statistics Data
    - National Healthcare Provision Surveys
  - **Agency for Healthcare Research & Quality**
    - Medical Expenditure Panel Survey (MEPS)
Sample Research: Health Data

"Employer Health Benefit Costs and Demand for Part-time Labour"
The link between rising employer costs for health insurance benefits and demand for part-time workers is investigated using non-public data from the Medical Expenditure Panel Survey- Insurance Component (MEPS-IC). Conditional on establishments offering health insurance to full-time employees, a 1 percent increase in employer health insurance contributions results in a 3.7 percent increase in part-time employees working at establishments in the U.S.

"The Effects of Smoking in Young Adulthood on Smoking and Health Later in Life:..." Eisenberg & Rowe, Forum for Health Economics and Policy (2009) This paper investigates whether cigarette smoking in young adulthood has significant lasting effects into later adulthood. Using the restricted NHIS and exploiting the randomness of the Vietnam era draft lottery we find that military service increased the probability of smoking.
Important Web Sites

- Center for Economic Studies, U.S. Census
  - http://www.census.gov/ces/

- NCHS Research Data Center
  - http://www.cdc.gov/rdc/

- MEPS: Restricted Data Files Available at Data Centers
  - http://www.meps.ahrq.gov/mepsweb/data_stats/onsite_datacenter.jsp
NSF Funding Opportunities for RDC Data

- **Science of Organizations (SoO)**
  - SoO funds research that advances our fundamental understanding of how organizations develop, form and operate.

- **Decision, Risk and Management Sciences (DRMS)**
  - Disciplinary and interdisciplinary research, doctoral dissertation research, and workshops are funded in the areas of judgment and decision making; decision analysis and decision aids; risk analysis, perception, and communication; societal and public policy decision making; management science and organizational design.

- **Smart Health and Wellbeing (SHB)**
  - Through the Smart Health and Wellbeing (SHB) Program, NSF seeks to address fundamental technical and scientific issues that would support much needed transformation of healthcare from reactive and hospital-centered to preventive, proactive, evidence-based, person-centered and focused on wellbeing rather than disease.
NSF Funding Opportunities for RDC Data

  - Welcomes efforts by the research community to use NCSES data for research on the science and technology enterprise, to develop improved survey methodologies for NCSES surveys, to create and improve indicators of S&T activities and resources, and strengthen methodologies to analyze and disseminate S&T statistical data.

- NSF Graduate Research Fellowship Program (GRFP)
  - The purpose of the NSF Graduate Research Fellowship Program (GRFP) is to help ensure the vitality and diversity of the scientific and engineering workforce in the United States. The program recognizes and supports outstanding graduate students who are pursuing research-based master's and doctoral degrees in fields within NSF's mission.

- Science of Science and Innovation Policy (SciSIP)
  - The Science of Science & Innovation Policy (SciSIP) program supports research designed to advance the scientific basis of science and innovation policy. Research funded by the program thus develops, improves and expands models, analytical tools, data and metrics that can be applied in the science policy decision-making process.
NSF Funding Opportunities for RDC Data

- Methodology, Measurement, and Statistics (MMS)
  - The Methodology, Measurement, and Statistics (MMS) Program seeks proposals that are methodologically innovative, grounded in theory, and have potential utility for multiple fields within the social and behavioral sciences.

- NSF-Census Research Network (NCRN)
  - The activities will be expected to advance both fundamental and applied knowledge as well as further the training of current and future generations of researchers in research skills of relevance to the measurement of economic units, households, and persons.
The future of social science research entails combining Administrative Data with multiple data sources resulting in—BD2K

**Big Data to Knowledge**
BD2K: The Future is Now

Percent of applied microeconomics Articles in top Economics Journals using existing survey data (e.g. CPS, SIPP, etc.)

Source: Raj Chetty
BD2K: The Future is Now

Percent of applied microeconomics articles in top Economics Journals using administrative data e.g. scanner data, social security records etc.

Source: Raj Chetty
1) McKinsey Global Institute: Big Data refers to data that is beyond the ability of the typical database software tools to capture, store, manage, and analyze.

2) IBM: Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content. . .to answer questions that were previously considered beyond your reach.

3) BLS: Big Data is nonsampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference.
Big Data to Knowledge: BD2K

• The future of economic and social science research will be built on big data from several sources, especially administrative data.

• Strengths of Administrative data
  • No missing data or attrition
  • Very large sample sizes
  • Less expensive to collect than survey data
BD2K: Economic Growth & Development

• McKinsey Global Institute (2011) estimates 100s of billions of dollars in economic opportunity based on Big Data for US economy

• Big Data Creates value:
  • Used by firms to enhance productivity and customize products
  • Used by Obama campaign to win the 2012 election
  • Replaces human decisions with algorithms
  • Locally, significant opportunities will exist in healthcare (e.g. Cerner) & Google Fiber
BD2K: Data Science

McKinsey (2011) estimates the US Economy will need

• 140,000 – 190,000 Data Scientists
• 1.5 million data-literate managers to take advantage of big data opportunities
• Data Science—emerging field that combines skills of:
  • Data Management
  • Computer Programming
  • Statistical Knowledge
• North Carolina State University MS in Analytics Starting Salary ~$89,000 in 2012
• KU should develop its capacity to train students in statistics and data science.
Conclusions

• Access to data in the KCRDC will yield new insights and inform policy.

• *There is and will be a strong demand for students with data science skills*

• KCRDC:
  It’s all about Data
Conclusions

Thank You

Mahalo

Kiitos

Toda

Thanks

Merci

Grazie

Obrigado

Takk

Gracias

Donna K. Ginther, Ph.D.