Research Opportunities at the Cornell Research Data Center

Nichole Szembrot
Cornell Research Data Center
Federal Statistical Research Data Centers (FSRDC)
U.S. Census Bureau

Disclaimer: Any views expressed in this presentation are those of the author and not of the U.S. Census Bureau.

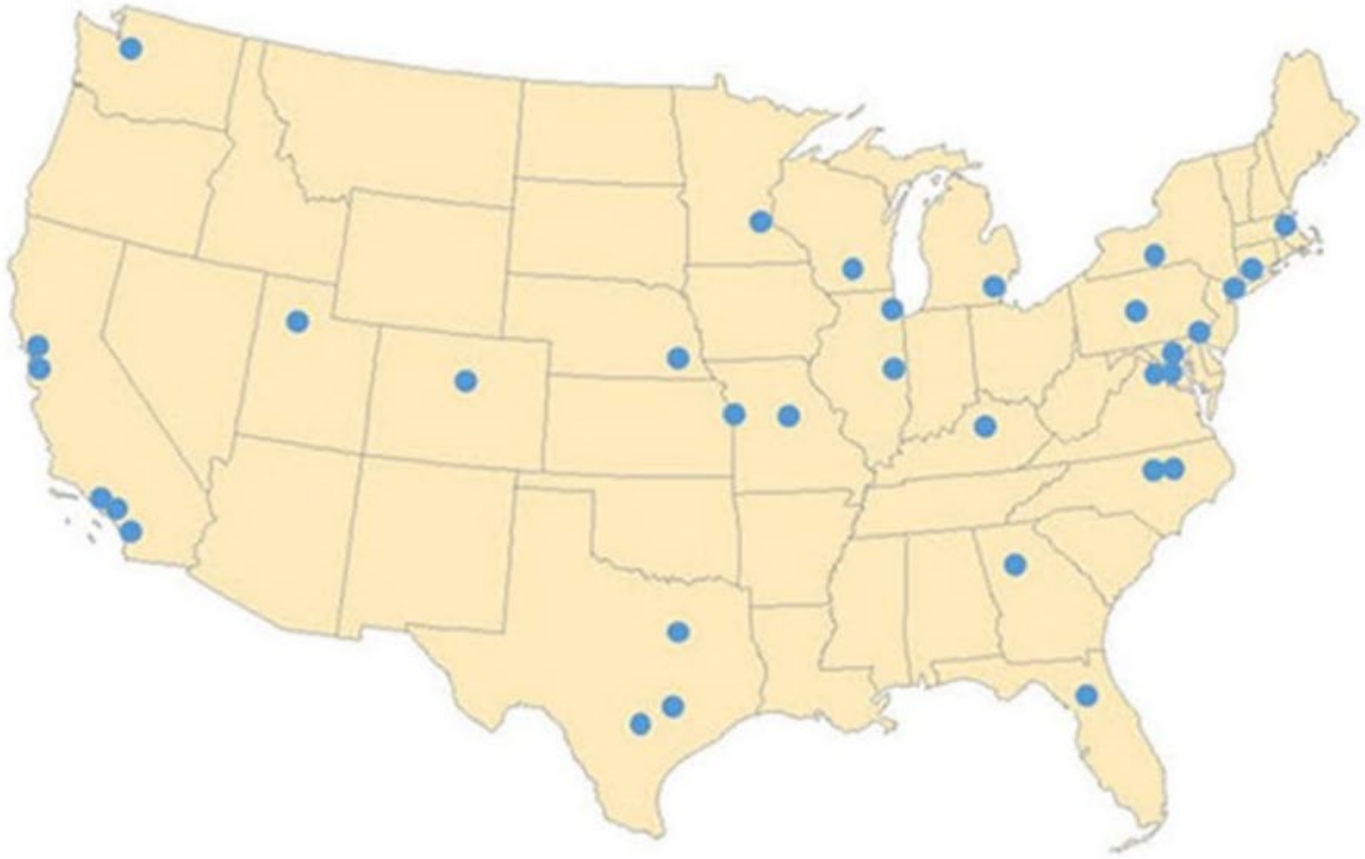
Outline

- FSRDC overview
- Available data
- Proposal process
- Questions

What are Federal Statistical Research Data Centers (RDCs)?

- Secure computing labs where qualified researchers conduct approved statistical analysis on non-public data.
- These data are collected by various government agencies (Census Bureau, NCHS, BEA, BLS, SSA, etc.).
- Established through an agreement between federal statistical agencies and a local research community.
- Managed by the Census Bureau.

FSRDC Locations



Data Availability

Census Bureau Data

- Economic Data
- Demographic Data
- Longitudinal Employer-Household Dynamics (LEHD)
- UMETRICs Data
- Criminal Justice Administrative Records System (CJARS)

Other Agency Data

- Agency for Healthcare Research and Quality (AHRQ)
- National Center for Health Statistics (NCHS)
- Bureau of Labor Statistics
- Bureau of Economic Analysis
- National Center for Science and Engineering Statistics (NCSES)
- Substance Abuse and Mental Health Services (SAMHSA)
- Equal Employment Opportunity Commission (EEOC)
- Federal Reserve Board-Microeconomic Surveys Unit

Restricted Data Advantages

- No publicly-available microdata
  - Internal data at the establishment and firm level
  - Universal scope
  - Detailed industry and geography
- Linking data
  - Consistent identifiers
  - Business register
  - External data
- Less top-coding and more detail for demographic data

Types of Restricted Census Data Available

- Economic Data
  - Microdata on firms and establishments
- Demographic Data
  - Microdata on individuals and households
- Administrative records
  - Data originally collected for administrative use, repurposed for research
- Linked Data
  - Data on employees linked with data on employers (LEHD)
  - Data on university grant awardees linked with data on employees and vendor firms (UMETRICS)

Most Popular Economic Microdata

Data Set	Unit of Enumeration
Economic Censuses (Manufactures, Retail Trade, Services, etc.)	Establishment
Annual Surveys (Manufactures, Services, etc.)	Establishment/Firm
County Business Patterns Business Register	Establishment
Business Research & Development and Innovation Survey (BRDIS)	Firm
Longitudinal Business Database	Establishment
Longitudinal Firm Trade Transactions Database (LFTTD)	Firm
Survey of Business Owners/Annual Survey of Businesses	Firm
Many more economic datasets!	Establishment/Firm

Economic Example

"Firms' Internal Networks and Local Economic Shocks," Giroud and Mueller, *American Economic Review* (2019)

- The authors examine how local shocks spread across US regions through firms' internal networks of establishments.
- They construct a spatial network of an entire firm's establishments using the **LBD** and categorize establishments using zip codes and 4-digit NAICS codes. Shocks to consumer demand are measured using zip code level changes in the housing prices from 2006-2009.
- They find that establishment-level employment is sensitive to shocks in distant regions in which the parent firm is operating, and that aggregate county level employment is sensitive to shocks in distant counties linked by firms' internal networks.

Demographic Data

Data Set
Decennial Census
American Community Survey
American Housing Survey
Current Population Survey (and Supplements)
National Crime Victimization Survey (and Supplements)
National Survey of College Graduates
Rental Housing Finance Survey
Survey of Income and Program Participation
National Longitudinal Mortality Study

Linking - the Power of the PIK

- PIK – a unique Census identifier assigned to individuals using name, birthdate, address information, and social security number.
- Many restricted Census demographic datasets contain PIKs
- PIK crosswalks permit linking restricted demographic datasets together:
  - 1940, 2000, and 2010 Decennial Census
    - Coming in the next couple of years: 1950, then 1960-1990 Decennial Census
  - American Community Survey
  - Survey of Income and Program Participation
  - Several administrative datasets
  - …and more

Demographic Administrative Datasets

- Numident (Social Security Administration)
  - Birth place, Birth date, Death Date (Social Security records)
- Census Household Composition Key
  - Links the PIKs of children born beginning in 1996 to the PIKs of their parents (from SSN registrations)
- Master Address File
  - MAF Extract (MAFX): Yearly snapshot of MAF for research purposes. Contains complete list of addresses with lat/long, geography variables, and address ID (MAFID)
  - MAF Auxiliary Reference File (ARF): Links MAFID to individuals' PIKs so can track annual residence beginning in 2000

Demographic Administrative Datasets

- SSA data for CPS, SIPP, and ACS respondents
  - Supplemental Security Record (SSR)
  - 831 Disability File
  - Master Beneficiary Record (MBR)
  - Summary Earnings Record (SER)
  - Detailed Earnings Record (DER)

Demographic Administrative Datasets

- SNAP / TANF / WIC
  - Data availability varies by state
  - Must provide benefits to the Food and Nutrition Service
    - Evaluate and improve surveys
    - Provide estimates to improve understanding of poverty and food insecurity
    - Contribute to evidence-based policy and program evaluation
- Moving to Opportunity Survey (MTO)
  - Designed to help low-income families in public housing move to “opportunity” neighborhoods with vouchers and counseling.
  - Ran in five large cities between September 1994 and August 1998
  - Data collected on participants multiple times from 1994 through 2010
  - Individuals in RDC files have been assigned PIKs
- Criminal Justice Administrative Records System (CJARS)
  - Microdata from participating agencies following individuals through the criminal justice system
  - Includes arrestees, criminal defendants, inmates, and probationers/parolees

Demographic Administrative Datasets

- IRS Form 1040 External Commingled File
  - 1969, 1974, 1979, 1984, 1989, 1994
  - MAFID assigned to mailing address
  - 1969: PIK for filer only
  - 1974-1989: PIKs for filer and spouse
  - 1994: PIKs for filer, spouse, and up to 4 dependents
  - Data include income amounts, exemption counts, and schedule filing indicators
  - Newly Available Individual-Level U.S. Tax Data from 1969-1994

Demographic Example

"The Effects of Gentrification on the Well-Being and Opportunity of Original Resident Adults and Children," Brummett and Reed, *FRB of Philadelphia Working Paper* (2019)

- The authors link individuals appearing in both the 2000 **Decennial** and the 2010-2014 **ACS** and observe their neighborhood (census tract) of residence at both points in time, detailed demographic and housing characteristics, and a variety of outcomes**.**
- They find that gentrification modestly increases out-migration, though movers are not made observably worse off. Many original residents stay and benefit from declining poverty exposure and rising house values. Also, there is some evidence gentrification increases the probability that children of less-educated homeowners attend and complete college.

Linked Data - LEHD

- Link individuals to place of employment
- Based on unemployment insurance administrative records
- Contains quarterly earnings for every company employee
- “Tracks” a person based on their place of employment
- Contains some demographic information on individuals
- Can link LEHD to Census business data via firm identifier
- Can link LEHD to Census household data via PIK
- Available on a state-by-state basis

LEHD Example

"Employee Costs of Corporate Bankruptcy" Graham et al., *NBER Working Paper* (2019)

- 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 following bankruptcy.
- This paper quantifies the earnings losses to employees induced by bankruptcy filings. The researchers finds wages decline by 10% in the year of the bankruptcy filing and decline by a cumulative present value of 67% over 7 years.

Linked Data - UMETRICS

- **Universities: Measuring the Impacts of Research on Innovation, Competitiveness, and Science**
- Database containing:
  - Information on university grants (32 universities currently)
  - People paid on the grants
  - Businesses paid by the grants
- Can link grant recipients to internal Census Bureau data on individuals and firms

UMETRICS Example

"Research Funding and Regional Economies," Goldschlag et al, *Center for Economic Studies Working Paper* (2016)

- Authors link **UMETRICS** vendors to **LBD** establishments. They link in information on R&D at the firm level from **SIRD** and **BRDIS**.
- Grant money is more likely to be spent at businesses located closer to universities, providing evidence on the effect of university research on local economies. Vendors who previously supplied a grant are more likely to supply grants in future years and more likely to open an establishment near the university.

Data from Partnering Agencies Accessible in the RDC

- Agency for Healthcare Research and Quality (AHRQ)
- National Center for Health Statistics (NCHS)
- Bureau of Labor Statistics (BLS)
- Bureau of Economic Analysis (BEA)
- National Center for Science and Engineering Statistics (NCSES)
- Substance Abuse and Mental Health Services (SAMHSA)
- Equal Employment Opportunity Commission (EEOC)
- Federal Reserve Board-Microeconomic Surveys Unit

- Aside from jointly approved projects, the proposal process is with individual agencies and not Census, though the background check is done by Census.

Health Data

National Center for Health Statistics (NCHS)

- RDC - On-site at a FSRDC (cdc.gov)
- Email: rdca@cdc.gov

Substance Abuse and Mental Health Services Administration (SAMHSA)

- Recently includes the National Survey on Drug Use and Health (NSDUH)
- RDC - On-site at a FSRDC (cdc.gov)
- Email: rdca@cdc.gov

Agency for Healthcare Research and Quality (AHRQ)

- Medical Expenditure Panel Survey (ahrq.gov)
- Email: CFACTDC@AHRQ.HHS.GOV

National Center for Health Statistics (NCHS)

Types of Restricted Variables

- Geographic Variables
- Linked Data Products
- Genetic Variables (NHANES phenotype data)
- Temporal Variables
  - e.g. dates of birth, death, exams
- Detailed Race/Ethnicity Variables
- Sensitive Variables
  - e.g. youth sexual behavior and mental health
- Long Term Health Care Survey Merging Variables

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

NCHS Example

"Local Employment Conditions and Unintended Pregnancy," Su, *Journal of Marriage and Family* (2018)

- Used the **National Survey of Family Growth** with the restricted county variable linked to publicly available data from the Census Bureau on local employment conditions
- In areas with higher unemployment rates, women were less likely to experience unintended pregnancy. Effects were larger for the least-educated women.

AHRQ Data

Restricted MEPS Data Available
Household Component-Insurance Component Linked File
Nursing Home Component
Medical Provider Component (except directly identifiable data)
Two-Year, Two-Panel Files
Area Health Resources Files
MEPS Link Files to NHIS

Agency for Healthcare Research and Quality (AHRQ)

- Fully specified ICD-9 medical condition codes
- Fully specified industry and occupation codes
- Lower levels of Geography
  - State and county FIPS codes
  - Census tract and block-group codes
- Non-public use data elements
  - Asset information
  - Imputed NDC codes
- Federal and state marginal tax rates

Bureau of Economic Analysis Data

- Allow researchers to analyze economic behavior of
  - Multinational Enterprises (MNEs)
  - Firms that trade in services
- Can be linked to Census establishment level data
- Require researchers to be U.S. citizens to access data
- https://www.bea.gov/research/special-sworn-researcher-program or email SpecialSwornResearch@bea.gov

BEA Data

Data Set
Annual and Benchmark Surveys of U.S. Direct Investment Abroad (BE-10/11)
Quarterly Survey of U.S. Direct Investment Abroad (BE-577)
Annual and Benchmark Surveys of Foreign Direct Investment in the United States (BE-12/15)
Quarterly Survey of Foreign Direct Investment in the United States (BE-605)
Benchmark and Quarterly Survey of Transactions in Selected Services and Intellectual Property With Foreign Persons (BE-120/125), Annualized Data
Benchmark and Quarterly Survey of Financial Services Transactions between U.S. Financial Services Providers and Foreign Persons (BE-180/185), Annualized Data
Benchmark and Quarterly Surveys of Insurance Transactions by U.S. Insurance Companies with Foreign Persons (BE-45/140), Annualized Data
Annual Survey of New Foreign Direct Investment in the United States (BE-13)

BEA Example

"Innovation in the Global Firm," Bilir and Morales, *Journal of Political Economy* (2020)

- Authors use the **Survey of U.S. Direct Investment Abroad** to create an annual panel of U.S. parent companies and each foreign affiliate.
- This paper studies how technological gains developed at one plant are shared across a firm's plants in other countries. For the median multinational, 20% of the return to R&D investment in the U.S. is realized outside of the U.S.

BLS Data

https://www.bls.gov/rda/what-datasets-are-available.htm
RDA_Admin@bls.gov

Data Set
Census of Fatal Occupational Injuries
International Price Program (IPP)
National Compensation Survey (NCS)
National Longitudinal Surveys of Youth 1979 (NLSY79)
National Longitudinal Surveys of Youth 1997 (NLSY97)
NLSY97 School Surveys
Occupational Requirements Survey (ORS)
Producer Price Index
Survey of Occupational Injuries and Illnesses (SOII)

National Center for Science and Engineering Statistics

- Survey of Earned Doctorates
- Survey of Doctoral Recipients
- PIK crosswalks allow for links with other Census data

Federal Reserve Board-Microeconomic Surveys Unit

- Survey of Consumer Finances
  - Survey of household balance sheets, income, employment, attitudes, and demographics
  - Restricted version includes geographic identifiers and more detailed responses

Equal Employment Opportunity Commission

- EEO-1: Job Patterns for Minorities and Women in Private Industry
  - Number of individuals employed by job category, sex, and race or ethnicity
- EEO-3: Job Patterns for Minorities and Women in Referral Local Unions
  - Membership, applicant, and referral information by sex and race/ethnicity
- EEO-4: Job Patterns for Minorities and Women in State and Local Government
  - Data by race/ethnicity, sex, job category, and salary band
- EEO-5: Job Patterns for Minorities and Women in Elementary-Secondary Schools
  - Data by sex, race/ethnicity, and activity assignment classification
- EEOC Employment Discrimination Charge Statistics

How to Access the FSRDC

- Develop proposal
  - Different guidelines for the various agencies
  - Submit proposal for agency review through Standard Application Process (SAP) portal
  - Review time varies by agency
- Obtain Special Sworn Status (SSS)
  - Residency requirement (3 years of last 5)
- Pay fees for NCHS, BEA, BLS data (some include fees for Special Sworn Status)

Timeframe

- Census data
  - Plan on 3 to 7 months for review process
  - Economic data requires IRS approval in addition to Census
- Other agency data
  - Timeline dependent on agency approval process
  - Census approval NOT required
- Special Sworn Status
  - 3 - 4 additional months for your “security clearance”
  - Longer for foreign nationals

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 RDC Administrator to craft final proposal. Enter and submit proposal in SAP once finalized.

Working in the FSRDC 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, Python etc.
- Agency reviews output before releasing
  - Penalty for disclosure is $250,000 and/or 5 yrs in prison (inadvertent or otherwise)

Virtual RDC

- Pilot virtual access program expanded during the COVID-19 pandemic
- Projects that use only Census Bureau data or BEA data are eligible
- New researchers are trained in the RDC, then are approved for virtual access from a secure workspace within their home
- Same rules apply as the RDC: no visitors, no phone use or internet browsing while working on RDC project. Output goes through usual disclosure avoidance review

Contact

- Nichole Szembrot, Cornell RDC Administrator
  - nichole.e.szembrot@census.gov
  - 607-645-3091
- Zhuan Pei, Cornell RDC Executive Director
  - zhuan.pei@cornell.edu
- For more information: Federal Statistical Research Data Centers (census.gov)
- Standard Application Process portal and metadata: Research Data Gov

Fall 2024 Workshop Evaluation: Research Opportunities at the Cornell FSRDC