Restricted Access US Census Data and the Federal Statistical RDC Program

Wayne Gray

Executive Director, Boston RDC

(Professor of Economics, Clark University)

Chinese Economists Society

Data Training Program

August 13, 2020

Agenda

- **Overview of FSRDC program**
- History of RDCs
- RDC datasets and research examples
- RDC proposal process and research environment
- Challenges and Opportunities

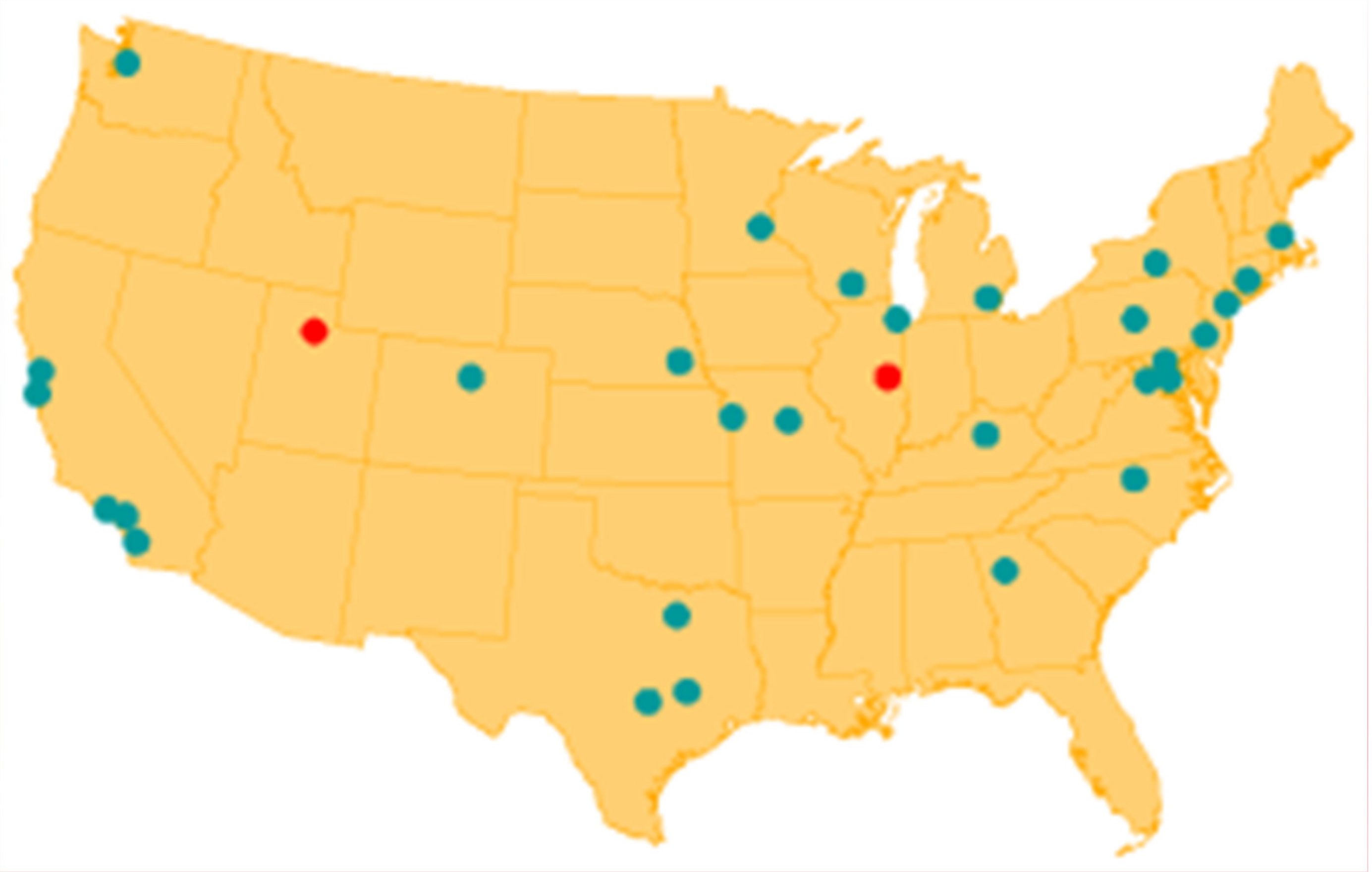
Research Data Center - RDC

- Partnership – Census Bureau and Local Host
  - Host – University or consortium of universities
  - Joint Statistical Project agreement (50/50 cost sharing)
- Census Bureau provides
  - Thin client access to Census linux servers
  - Dataset preparation and RDC system management
- Local host provides
  - Salary of Census Administrator working in RDC
  - RDC space and Executive Director
  - Local support of research activity

Research Data Center - RDC

- RDC researchers get access to internal Census microdata
  - Security clearance – Special Sworn Employees of Census
  - Providing benefits to Census Bureau (improving data quality)
  - Proposals reviewed by Census
  - Statistical results can be released for publication
  - Output reviewed by Census to avoid disclosing individual data
- Other statistical agencies provide data
  - Pay Census to cover costs of hosting data on system
  - Handle their own proposal and output reviews

Currently 30+ Research Data Centers



Research Data Centers

RDC	State	Open	RDC	State	Open
NBER - Boston	MA	1994	U. Missouri	CT	2015
UC Berkeley/Stanford	CA	1998	U. Wisconsin	WI	2015
UCLA/USC/UC Irvine	CA	1998	Kansas City Fed	KS	2016
Duke	NC	2000	U. Maryland	MD	2016
Chicago Fed	IL	2002	U. Nebraska	NE	2016
U. Michigan	MI	2002	Georgetown	DC	2017
Cornell	NY	2004	U. Kentucky	KY	2017
CUNY Baruch	NY	2006	Philadelphia Fed	PA	2017
U. Minnesota	MN	2010	U. Colorado Boulder	CO	2017
Atlanta Fed	GA	2011	U. Texas Austin	TX	2017
U. Washington	WA	2012	Dallas Fed	TX	2018
Texas A&M	TX	2012	Federal Reserve Board	DC	2019
Penn State	PA	2014	U. Illinois Champaign	IL	2020
Yale	CT	2015	U. Utah	UT	2020

Agenda

- Overview of FSRDC program
- **History of RDCs**
- RDC datasets and research examples
- RDC proposal process and research environment
- Challenges and Opportunities

History of RDCs – Census/CES

- CES – Center for Economic Studies
- Part of Census – Economic Directorate – dealt with business data
- Census business microdata had no public use versions
- Researchers visited Census Headquarters
- Worked with CES employees on research projects

History of RDCs – Census/CES

- 1992 – I was visiting CES as Census Fellow
  - Using business microdata
  - Studying impact of environmental regulation on productivity
- Robert McGuckin – head of CES
  - Wanted to set up remote locations to expand access
  - Thought Boston would be ideal location
    - Prestigious research universities – high local demand for RDC
    - Support from Census regional director – Arthur Dukakis
    - Support from NBER as local partner – Martin Feldstein
  - I agreed to be the coordinator (Executive Director) for BRDC

History of RDCs – Census/CES

- 1993 – Boston RDC proposal to National Science Foundation
  - Approved and funded; RDC opened in January 1994
  - RDC located in Census Regional Office in Boston
- 1998-2004 – 6 RDC locations opened
- 2005-2014 – 8 RDC locations opened
- 2015-2020 – 15 (!) more RDC locations opened
- Continuing high demand for new locations
  - Some are new RDCs with support from NSF
  - Others are branch locations of existing RDCs

History of RDCs – Federal Statistical RDCs

- Originally Census RDCs, hosted only Census data
- Later added health data – from NCHS and AHRQ
- Then came additional federal agencies – BLS, BEA, others
- Re-labeled as Federal Statistical Research Data Centers (FSRDCs)
- Executive Committee – statistical agencies and RDC partners
- Census still manages RDC operations
  - Managed by CED – Center for Enterprise Dissemination
  - Data kept on Census computer network
  - Other agencies pay fees to Census for hosting data

Agenda

- Overview of FSRDC program
- History of RDCs
- **RDC datasets and research examples**
- RDC proposal process and research environment
- Challenges and Opportunities

RDC data – Census business data

- Data on establishments linked over time
- Greenstone, Hornbeck, Moretti (JPE 2010), “Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings”
  - Economic Census and Longitudinal Business Database (LBD)
  - Winning and losing counties have similar trends in incumbents’ TFP prior to a large new plant opening.
  - Five years after the opening, incumbent plants’ TFP is 12 percent higher in winning counties.

RDC data – Census business data

- Linking establishment data from different surveys
- Gray, Linn and Morganstern (EJ 2019), “The Impacts of Lower Natural Gas Prices on Jobs in the US Manufacturing Sector”
  - LBD and Manufacturing Energy Consumption Survey
  - Connect energy prices (especially natural gas) to employment
  - Detailed data – county * industry – MECS data on NG-intensity
  - Controls for many other factors at county-industry level
  - 50% lower NG prices tied to 0.6% increase in mfg employment
  - Smaller impacts than earlier research, due to greater controls

RDC data – Census business data

- Creating new Census surveys for microdata analysis
- Bloom, et. al. (AER 2019) “What Drives Differences in Management Practices?”
  - Management and Organizational Practices Survey (MOPS)
  - Collected in 2010 and 2015, sent to 35,000+ plants
  - Extensive questions, panel data, links to business outcomes
  - More structured management practices = better performance
    - Productivity, profitability, growth, survival, innovation
  - Enormous variability in practices across plants
  - Variability within firms, more variability in larger firms

RDC data – Census business data

- Detailed location on individual businesses
- Krishnan, Nandy, Puri (RFS 2015), “Does Financing Spur Small Business Productivity? Evidence from a Natural Experiment”
  - Longitudinal Business Database (LBD), Census of Manufacturers, Annual Survey of Manufacturers
  - Interstate banking deregulations -> increased access to bank financing -> increases in firms’ TFP productivity
  - Regression discontinuity around SBA funding eligibility

RDC data – Census business data

- Linking Census data to administrative data
- Links to data on trade flows
- Bernard, Jensen, Schott, (NBER 2005), "Importers, exporters, and multinationals: A portrait of firms in the US that trade goods"
  - Longitudinal Business Database (LBD) – plant characteristics
  - International trade data on imports and exports
  - Identifies arms-length vs. related-party trades
  - Trade is very concentrated (1% of firms = 81% of trade)
  - Job creation concentrated in firms that begin trading

RDC data – Census business data

- Merging individual businesses to external data
- Links to Environmental Protection Agency data
- Gray and Shadbegian, (JRS, 2007) “The Environmental Performance of Polluting Plants: A Spatial Analysis”
  - Longitudinal Business Database (LBD) – plant characteristics
  - EPA data – emissions, compliance and enforcement
  - Enforcement activity improves compliance
  - Effects at inspected plant and nearby plants
  - But not across state boundaries (different jurisdictions)

RDC data – Census demographic data

- Detailed location on residence and workplace for individuals
- Bayer, Ross, and Topa (JPE, 2008) “Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes”
  - Use census block of residence and census block of work to look for social hiring networks.
  - They find a significant effect of social networks on hiring, especially among those with similar socio-demographic characteristics.

RDC data – Census demographic data

- American Housing Survey – detailed location of houses
- Lucas Davis (RESTAT 2011), “The Effect of Power Plants on Local Housing Prices and Rents”
  - Using census block, merged in data on the location of waste incinerators, coal-burning plants, and nuclear power plants.
  - Neighborhoods within two miles of plants experienced a 3-7 percent decrease in housing values and rents.

RDC data – Census LEHD

- LEHD = Linked Employer Household Data
- Based on state unemployment insurance records
- Barth, Bryson, Davis, Freeman (JLE 2016), “It’s Where You Work: Increases in the Dispersion of Earnings across Establishments and Individuals in the United States”
  - Contribution of establishments in the upward trend in earnings dispersion
  - LEHD linked to establishment and decennial data

RDC data – Health data – NCHS/AHRQ

- Detailed codes – medical condition, industry, occupation
- Detailed location – state and county, tract and block group
- Dates of birth, death, exams
- Medical Expenditure Panel Survey (AHRQ)
- National Health and Nutrition Examination Survey (NCHS)
- National Health Interview Survey (NCHS)
- National Vital Statistics System (NCHS)
- Linkages to mortality, air quality, benefit history, Medicare claims
- https://www.cdc.gov/rdc/b1datatype/dt100.htm

RDC data – Bureau of Labor Statistics

- Detailed location – depending on dataset
- Some BLS datasets have establishment identifiers
- National Longitudinal Survey (NLS) with detailed geography
- Census of Fatal Occupational Injuries (CFOI)
- Survey of Occupational Injuries and Illnesses (SOII)
- National Compensation Survey, All Benefits Quarterly
- Producer Price Index (PPI)
- https://www.bls.gov/rda/eligibility-and-access-modes.htm

RDC data – BLS - SOII

- Survey of Occupational Injuries and Illnesses
  - Comes with establishment identifiers
  - Done at BLS Headquarters – now would be possible in RDCs
- Gray and Mendeloff (ILRR 2005), “The Declining Effects of OSHA Inspections on Manufacturing Injuries: 1979 to 1998”
  - Link SOII injury data to OSHA workplace inspections
  - Do workplace injuries decline after OSHA inspections?
  - Earlier work for early 1980s had shown large reductions
  - We found much smaller impacts in later 1990s
  - Bigger impacts on smaller plants and non-union plants

RDC data – BLS – CFOI

- Census of Fatal Occupational Injuries data
  - Provides state and some workplace characteristics (not identifiers)
- Gray and Mendeloff – ongoing project
  - Determinants of state-level construction fatality rates
  - State policies include workers compensation, OSHA enforcement
  - Subsectors of industry, small firms, self-employed workers
  - Preliminary findings show some impacts of state policies
  - Higher fatalities in states with less strict WC policies

RDC data – BEA

- BEA does surveys (quarterly and annual) with an international focus
- Foreign direct investment
  - Both inward and outward flows of investment and income
- Activities of multinational enterprises
  - Foreign affiliates of US parent firms
  - US affiliates of foreign parent firms
- International trade in services
  - Transport, financial, insurance, intellectual property
- https://www.bea.gov/research/special-sworn-researcher-program

Agenda

- Overview of FSRDC program
- History of RDCs
- RDC datasets and research examples
- **RDC proposal process and research environment**
- Challenges and Opportunities

Legal Protections on Census Microdata

- U.S. laws regulate use of data by federal statistical agencies
- Census – Title 13 protects confidentiality
  - Only employees can access data; must benefit Census programs
  - Benefits – examine data, identify issues, create new questions
  - Special Sworn Status – temporary unpaid Census employees
  - Original use – statisticians helping design surveys
  - Later expanded to external researchers doing statistical analyses
  - Penalties for violations – up to 5 years in jail, $250,000 in fines

Rules on Data Access – other agencies

- Laws and rules vary across statistical agencies
- Justification for data access?
  - For BLS and NCHS, using data to get research results is a benefit
- Who can get access?
  - Census doesn’t require US citizenship, some other agencies do
  - Only those at US institutions - to give US laws jurisdiction
- Sometimes multiple agencies involved
  - Census business data based on tax returns – IRS approval needed
  - LEHD – requires approval by state agencies that provide data
  - Projects merging datasets across agencies (complex negotiations)

Proposal Process - Overview

- Generate research idea
  - Identify data needed, agency involved, research team
  - Submit initial/preliminary proposal
- Work with agency, get feedback, further proposal refinement
  - Also contact RDC – confirm access, any fees required
- Submit final proposal to agency
  - Agency proposal review – possibly multiple agencies
  - Possible need for revisions to proposal
  - Project Approval
- Apply for Special Sworn Status
- Start research!

Proposal Process - Contents

- Initial/preliminary proposal
  - One page description of goals and data needed
- Final (Census) proposal
  - Typically 15-20 pages (plus list of datasets needed)
  - Need to show feasibility, Census benefits, scientific merit
    - Know the data, understand its advantages and shortcomings
    - Not “competing” on scientific merit – but needs to be sensible
  - Work with RDC Administrator on developing benefits
    - PPS = Predominant Purpose Statement (benefits to Census)
- Final proposal requirements for other agencies vary

Proposal Submission – ResearchDataGov at ICPSR

- Centralized portal for access to restricted federal microdata
  - https://www.icpsr.umich.edu/web/pages/appfed/about.html
- Searchable by agency and dataset
- Portal allows entry of initial information about project
  - CV, list of researchers, one-page description of goals and data
  - Sent to agency for initial review
  - Develop full proposal (interaction with agency personnel)
  - Final review by agency or agencies (if multiple ones involved)
- Note – portal is still under development
  - Eventually including final proposal submission, non-RDC data

RDC Access – Special Sworn Status

- Anyone using FSRDC needs to become Census “employee”
  - Because RDC is “Census” location
  - Applies to all projects, not just those using Census data
- Special Sworn Status – security review, like Census employee
  - Requires affiliation with US institution (usually university)
  - Requires US residence in 3 of past 5 years
  - Includes individual interview
  - Process can take 3+ months

RDC Research Environment

- RDC = secure Census-controlled environment
  - Census badge needed for access; security camera monitoring
  - Thin client terminals connected to Census server
  - Standard statistical packages (SAS, Stata, R, etc.)
  - Census datasets on server, access controlled based on project
  - External datasets submitted to agency, uploaded by Census staff
- Need to maintain confidentiality of data
  - Results submitted to agency review to avoid disclosures
  - Can’t discuss results outside of RDC before they’re reviewed
  - Can’t identify respondents (“fact of filing” is confidential)

RDC Research – Census Results

- Results of statistical analysis
  - Coefficients from model estimation
  - Limitations based on sample “supporting” the number
  - Can’t report median (might report average of 45%-55%)
  - Graph distribution with bin averages
  - Model can include firm/plant dummies – but can’t report them
  - Sample sizes reported with rounding
- Noise infusion can also be used (“blurring” data points)
  - Required for analysis in small geographic areas
  - May help with releasing some graphical representations

Agenda

- Overview of FSRDC program
- History of RDCs
- RDC datasets and research examples
- RDC proposal process and research environment
- **Challenges and Opportunities**

Costs of working in FSRDCs

- Substantial investment in time
  - Proposal development and review (6-12+ months)
    - Variation across agencies – maybe quicker for health projects
    - But can be longer if complex project involving multiple agencies
  - Special Sworn Status process (3+ months)
  - Disclosure review of results (takes weeks or months)
- Can be substantial financial investment
  - RDC access – if not partner institution, “list price” ~ $20,000/year
  - Agency project fees can be $3,000 or more
  - Non-Census projects may also require SSS processing fee
- Investment – best for multiple papers, substantial research agenda

Benefits of working in FSRDCs

- All about the Data!
  - Microdata enables better research design and analysis
  - Linking microdata datasets within Census
  - Linking microdata across agencies (e.g. SSA linked to CPS)
  - Linking microdata to external research datasets (e.g. Compustat)
- Possible future restrictions on publicly available datasets
  - Concerns with disclosure risk from public use microdata
  - 2020 US Census – less publicly available data
  - Canada – no public use files -> use RDCs for demographic research

Other Noteworthy Data Developments

- Differential Privacy
  - Based on formal statistical analysis of disclosure risks
  - Any release based on underlying microdata conveys information
  - Tradeoff – releasing more results vs. protecting privacy
  - Privacy budget – measure of tradeoff
  - Adding (well-defined) noise to microdata before analysis can help
  - https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html
  - https://www.brown.edu/Departments/Economics/Faculty/John_Friedman/dp_aea.pdf

Other Noteworthy Data Developments

- Synthetic Data
  - Simulated microdata, based on distribution of actual microdata
    - Including joint distribution of some variables
  - Allows public-use datasets in difficult situations
    - Survey of Income and Program Participation Synthetic Beta
    - Synthetic Longitudinal Business Database
  - Allows public provision of data-querying tools
    - On The Map – workforce statistics with geographic detail
  - https://www.census.gov/newsroom/blogs/research-matters/2014/10/synthetic-data-public-use-micro-data-for-a-big-data-world.html

Other Noteworthy Data Developments

- Administrative data
  - Linking in individual (SSN) and business (EIN) data
  - Reduce respondent burden, e.g. Economic Census from tax data
- Data partnerships
  - UMetrics data – University of Michigan – data collected externally
  - Collects data on researchers (students+faculty), research grants
  - Linked to Census data (individuals + businesses)
  - https://www.census.gov/programs-surveys/ces/data/restricted-use-data/umetrics-data.html

FSRDC References

- Wayne Gray, Clark University, wgray@clarku.edu
  - (please feel free to send me any questions you might have)
- FSRDC website, https://www.census.gov/fsrdc
  - List of FSRDC locations
  - List of Federal Partner Agencies
- ResearchDataGov at ICPSR – initial proposal submission
  - https://www.icpsr.umich.edu/web/pages/appfed/index.html

- Thank you for your interest in the data!