Accessing & Safeguarding Administrative Data at the CCPR: The Census Research Data Center (RDC) & German Data Center

Till von Wachter, Director, FSRDC
John Sullivan, Administrator, FSRDC
What is the RDC system?

- **RDC**: secure data lab to access confidential government data

- **Rich Amount of Data:**
  1. Demographic (individual) data (Census)
  2. Business data (Census)
  3. Health data (NCHS)
  4. Labor data (BLS)

- **How to use RDC:**
  project proposal; agency approval; research in lab; disclosure process

- **3 Goals of Talk:**
  1) make you aware of data; 2) tell you how to access data; 3) financing
Some Key Take Aways

1. Some Key Points Regarding Data
   - There are some low hanging fruits
   - Health data big untapped resource

2. RDC Network and Available Data Likely to Grow
   - Increasing number of data sets. Increasing number of branches and projects.

3. Getting project approval not that hard
   - Plenty of researchers can clearly do it
   - We are there to help – hope to do even more in future

4. Funding situation at UCLA
   - Students get in for free. Others have to pay.
   - UCLA last RDC that is mainly fee based. We are taking steps to change that
Part 1: Data Available in the RDC
Background on RDC Network

Census Bureau administers a network of RDCs Across U.S.

- There are currently 24 RDCs (and branches)
- Large number of active research projects (200+)

Goal of RDC is to make data accessible while safeguarding confidentiality

- Stringent rules of access and disclosure (more on this below)

Data in principle not available for research per se

- The law allows Census to give access to data to improve Census data products (more on this below)

An Increasing Number of Agencies is Using RDC Network

- Census is joined by NCHS, AHRQ, BLS
- Changed name to Federal Statistical Research Data Centers (FSRDC)
Overview of Types of Data Sources

1. Demographic Data (Census Bureau)
   - Ex: Decennial Census, ACS, CPS, NLMS, etc.

2. Economic Data (Census Bureau)
   - Ex: Economic Census, Annual Survey of Manufacturing, Longitudinal Business Database (LBD), etc.

3. Health Data (National Center for Health Statistics)
   - Ex: finer geographic detail; finer detail on race/ind/occ; added information

4. Labor Data (Bureau of Labor Statistics)
   - Ex: NLSY with geocodes; occupation injury statistics

5. Merged and Administrative Data Sets
   - Ex: Longitudinal Employer Household Dynamics (LEHD)
1. Demographic Data – Why Use the RDC?

- Micro-data with detailed geography
  - Tract level for most (block-level for Decennial and ACS)
  - Not available in public micro-data
- Less severe top coding
- Some datasets have additional variables
- Opportunities for individual level linkages (PIKs)
- Potential for “unswapped” data
- Not suitable venue for a “special tabulation”
1. Demographic - Available Data (1)

- Decennial Surveys
  - 1950 - 2010

- American Community Survey (ACS)
  - Annual microdata, 1996-2014

- Current Population Survey (CPS)
  - Various Supplements (including March ASEC)
1. Demographic - Available Data (2)

- Survey of Income and Program Participation (SIPP)
- American Housing Survey (AHS)
- National Crime Victimization Survey (NCVS)
- National Longitudinal Mortality Study (NLMS)
- National Longitudinal Surveys (NLS)
  - Young/Mature Men/Women
- Administrative data
  - SSA, HUD
Example: Research from the UCLA RDC

- Ellis, et. al.
- Decennial Census 2000 long-form data
  - Tract-level location of mixed race households
  - Mixed-race household formation reduces metropolitan level racial segregation
2. Economic – Why Use the RDC?

- Establishment level data
  - No publically available micro-data
- Detailed geography
- Establishment – firm linkage
- Longitudinal linkage
- Linkage across economic and mixed data products
2. Economic - Available Data (1)

- Economic Census
  - Annual Survey of Manufactures, Annual Survey of Retail Trade, Annual Survey of Services, Monthly Wholesale Trade Survey.
- Longitudinal Business Database and ILBD (integrated LBD)
- Quarterly Financial Report (QFR)
- Survey of Business Owners (SBO)
- MEPS-IC
2. Economic - Available Data (2)

- Annual Capital Expenditures Survey (ACES)
- Business Research & Development and Innovation Survey (BRDIS)
- Manufacturers' Shipments, Inventories, and Orders (M3)
- Survey of Pollution Abatement Costs and Expenditures (PACE) and Manufacturing Energy Consumption Survey (MECS)
3. Economic - Available Data (3)

- Commodity Flow Survey (CFS)
- Longitudinal Foreign Trade Transactions Database (LFTTD)
- Kauffman Firm Survey
- LEHD
  - Additional step for ICF, but moving quickly now
Example: Research from the UCLA RDC

- Kemeny, Rigby and Cooke
- LEHD linked to Census of Manufactures, FT – Import/Export, etc.
- Rising import competition from developing economies increases likelihood of job loss among less educated U.S. workers.
3. Health Data - Why use the RDC?

- Lower levels of geography
- NCHS data linkages
  - Mortality
  - Medicare meta-data
  - Social Security Benefits
- Greater detail in variables
  - Race
  - Disease codes
  - Industry and occupation codes
3. Health Data – Available Data

- AHRQ (Agency for Healthcare Research and Quality)
  - Medical Expenditures Panel Survey (MEPS)
- NCHS (National Center for Health Statistics)
  - National Health Interview Survey (NHIS)
  - National Health and Nutrition Examination Survey (NHANES)
  - National Survey of Family Growth (NSFG)
  - National Vital Statistics System (NVSS)
  - National Health Care Surveys – NAMCS, NHAMCS, NHDS, NNHS, NNAS, NSRCF, NSLTCP
Example: Research from the UCLA RDC

- Laura Wherry and Sarah Miller
- National Health Interview Survey with restricted state identifiers
- Effect of ACA state Medicaid expansions on insurance coverage, access and utilization of health care and self-reported health
Example: Research from the UCLA RDC

- Julia Caldwell, et. al.
- MEPS-HC, ACS and Area Resource File (ARF)
- Rural v. Urban exposure to social conditions associated with disparities in access to health care.
4. New Data from BLS

- National Longitudinal Surveys of Youth (NLYS79 and NLSY97)
  - NLSY with regional identifiers
- Survey of Occupational Injuries and Illnesses
- Application process structured similar to NCHS
  - Contact and apply through BLS, but access data through RDC
5. Linkage/Administrative Data

- Possible to link external data to restricted data on the individual level
  - PIKs (Protected Identification Keys)
  - Linkage over time in mandatory collections is restricted
- CPS/SIPP linked to SSA Earnings Records
- CenHRS
- Patience is key
Part 2: How To Access RDC Data
Overview: Data Access

Basic Procedure for Census Proposal:

1. Get idea & check data; talk to RDC administrator & director
2. Write proposal explaining research idea, statistical analysis, and data
3. Submit proposal for review with the RDC administrator
4. Agency reviews proposal, may ask for revisions
5. In the meantime, fill out paperwork for Special Sworn Status (SSS)
6. Once project is approved, work in RDC. Output obtained in disclosure review

Some Differences in Format of Proposal:

1. **Census**: proposal requires a list of statistical “Benefits for the Census Bureau”
2. **NCHS**: proposal does not require benefits, but a specific list of variables
Ease of Access Can Vary Between Data Sets:

“Cookie cutter projects"
- Demographic data (only Census approves)
- Business data (Census & IRS approve)
- Health data (only NCHS approves, but no scientific merit review)
- Easy to add in public data sources (as long as specified in advance)

Higher hanging fruits
- Merge between various data sets
- Merge outside confidential data
- Merge data from various agencies (Ex: LEHD)
Application Process: Census vs. NCHS, AHRQ and BLS

- Census proposals are submitted to the RDC administrator
- Separate proposal process for NCHS/AHRQ/BLS, submit direct to agency – does not go through the RDC administrator
- Generally, easier to apply and applications are processed more quickly than projects using Census, IRS or other agency data
- Fees paid to NCHS for data extracts
Discuss Some Important Practical Issues

1. Practical considerations for writing proposals

2. What constitutes a “Benefit for the Census Bureau”

3. Questions About Special Sworn Status

4. Information About Confidentiality and the Disclosure Process
Suggestions for Census Proposals

- Plan ahead
  - At least 6 months (health) to a year (business) to get access
- Work with the RDC Administrator
- Written for a data expert rather than a content expert
  - IRS may also be a reviewer and should be considered
- Description limit is 15 pages single spaced (30 pages double)
- Benefits often emerge as proposal is developed
Proposal Outline

- Intro (1-3 pgs.)
  - Overview of benefits; describe research question; brief lit. rev.; overview of research plan and data
- Methodology (8-9 pgs.)
  - Detailed model specification, key variables, how data will be used in estimation; methods to complete benefits
- Data (1-3 pgs.)
  - Bureau-provided data; External Data
  - Linkage
- Output and Disclosure Risk (1-3 pgs.)
  - Model-based output (emphasized)
  - Tabular output
  - Technical memos
  - Disclosure risk and mitigation
- Duration and Funding (<1 pg.)
The **predominant purpose** of projects approved under Title 13 is to provide benefits to the Census Bureau.

13 benefit criteria (IRS only recognizes 9 of the 13).

#11 - Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5;

- All projects claim #11 and usually only one additional benefit.

- E.g. estimating non-response; develop weighting strategy; improve imputation; understand/improve data quality; construct/verify/improve sampling frames; evaluate concepts and practices of data collection
Opportunities and Challenges for Graduate Students

- Proposal development and review take a significant amount of time
  - Hierarchy of review time – SSA>IRS>Census>Health
- Work environment
  - Using a server cluster; data cleaning and documentation; disclosure review; advisor access
- Often only health projects are feasible for a grad student PI
- Work on a faculty member’s proposal or existing project
Background on Confidentiality

- Balancing the benefits of making restricted data available to the research community and the legal requirement to ensure respondent confidentiality.

- Disclosure of confidential material is prohibited by law:
  - Title 13 U.S.C. section 9 prohibits the disclosure of confidential information.
  - Disclosure is punishable by a fine of up to $250,000 or a prison term of up to five years (or both).

- Federal Tax Information (FTI):
  - Many economic datasets are “comingled” with IRS data
  - Title 26, U.S.C. Sections 7213, 7213A, and 7431 provide civil and criminal penalties for unauthorized use or disclosure of FTI.
  - Punishable by a fine of up to $250,000 or a prison term of up to five years (or both).
Special Sworn Status

- Special Sworn Status (SSS) with the Census Bureau is required to work from an RDC – regardless of which data you use.
  - SSS is granted to experts who can help the Census Bureau fulfill its mission. SSS holders are sworn for life to protect confidentiality.
- Application includes risk assessment and background check.
  - Separate from proposal review
  - 2-3 months to process (slightly longer for foreign nationals)
  - No fee
  - SSS is maintained through mandatory annual trainings
Disclosure Prevention

- A number of steps are taken to limit the risk of disclosure
  - Add to the cost of using the data and are often unfamiliar to researchers accustomed to using public data.
- Physical security limits access to the lab
  - Badged access - alarm protected lab
  - Thin Clients – no data onsite
Releasing Results

- Disclosure Avoidance Review
  - Process to review output to ensure no risk of disclosure
- Performed by RDCA and/or agency disclosure officer
  - Review process worked out in proposal stage
    - Catalog all samples, report cell sizes, detailed memos describing all releases
    - Turn around generally 1-2 weeks but plan for 3-4 weeks
    - Descriptive data can be problematic
    - Limit Intermediate output
Part 3: Some Stats About the RDCs
FSRDC Active Projects by Type
Projects by RDC and by Type

- Atlanta
- Boston
- California, Berkeley
- California, Stanford
- California, Los Angeles
- California, USC
- California, Irvine
- Central Plains
- Chicago
- Kansas City
- Missouri
- Michigan
- Minnesota
- Wisconsin
- New York, Baruch
- New York, Cornell
- New York, Yale
- Northwest
- Pennsylvania
- Suitland
- Texas
- Triangle

CENSUS
HEALTH
Projects by RDC and by Type

- Atlanta
- Boston
- California, Berkeley
- California, Stanford
- California, Los Angeles
- California, USC
- California, Irvine
- Central Plains
- Chicago
- Kansas City
- Missouri
- Michigan
- Minnesota
- Wisconsin
- New York, Baruch
- New York, Cornell
- New York, Yale
- Northwest
- Pennsylvania
- Suitland
- Texas
- Triangle
Graphical Representation – Census Projects

Notes: Node (FSRDC) size and color are proportional to number of sponsored external projects. Edge thickness and color are proportional to number of external projects each pair of FSRDCs collaborate.
Approval Rates

<table>
<thead>
<tr>
<th>Year</th>
<th>Submissions</th>
<th>Approval Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CY2013</td>
<td>40</td>
<td>67.5%</td>
</tr>
<tr>
<td>CY2014</td>
<td>54</td>
<td>70.4%</td>
</tr>
<tr>
<td>CY2015</td>
<td>56</td>
<td>64.3%</td>
</tr>
<tr>
<td>CY2016</td>
<td>38</td>
<td>70.3%</td>
</tr>
</tbody>
</table>
Census Review Duration

- **Mean**
  - CY2013: 89
  - CY2014: 68
  - CY2015: 64
  - CY2016: 54

- **Minimum**
  - CY2013: 25
  - CY2014: 7
  - CY2015: 4
  - CY2016: 14

- **Max**
  - CY2013: 244
  - CY2014: 202
  - CY2015: 309
  - CY2016: 116
Duration of SSS Processing
Summary and Last Questions

RDC is a growing resource for researchers to access confidential data

1. Demographic Data
2. Economic Data
3. Health Data
4. BLS Data
5. Merged Data

There is some up-front cost, but it is worthwhile to plan ahead

- Many research projects take longer anyways!

We are here to help

- RDC will have increasing resources to help in proposal writing!
*** 2017 RDC Conference held at UCLA! ***

Stay tuned for dates in Spring 2018

To get idea of projects check out previous conferences:
Synthetic Data Alternatives

- “Synthetic” versions of some popular micro-data are available
  - Data are simulated from statistical models and designed to mimic the distributions of the underlying real data
  - Results can be verified against real data
  - Easy access; preparation for full RDC proposal
- Access through Cornell University
  - SynLBD
  - SIPP Synthetic Beta (SIPP – SSA linked)
Universe or Sample?

- **Universe**
  - Establishments
    - LBD, SSEL, BR, Economic Census
  - Persons
    - Census Numident
  - Workers
    - LEHD (within participating states)
  - Transactions
    - LFTTD

- **Sample**
  - Establishments
    - Annual economic surveys held in intercensal years
    - BRDIS/SIRD, SBO, MEPS-IC, ACES, PACE, MECS
  - Domestic Shipments
    - CFS
  - Persons
    - ACS, SIPP, CPS, NCVS, NLMS, etc.
Linking External Data to Internal Data

- External data aggregated above individual level
  - Contextualize person or establishment records with external data at tract, zip code or county level
  - Describe in proposal
- Linking on Individual level (persons)
  - Protected Identification Keys (PIKs)
  - Not all internal micro-data are PIKd
  - For external data to be PIKd:
    - SSN, name, place of birth, address, etc.
    - MOU between Census and data owner
    - Additional fee paid to Census to PIK external records
- Clearinghouse, CARRA, CLIP
Pathways for Graduate Student Access

- Work on existing project
  - Contact Administrator or Executive Director to see if existing project fits your interest
  - Work will need to fall within the scope of existing project
- Make own application
  - Start early
  - Consult with advisor