

# Challenges and Bottlenecks of Working with Administrative Data

Evidence Based Policy for Agricultural and Applied Economists  
Workshop

October 8, 2021

Nichole Szembrot, Cornell FSRDC Administrator, U.S. Census Bureau

# Outline

- Federal Statistical Research Data Center (FSRDC) overview
- How Census links person records
- Administrative data in the FSRDC (not exhaustive)
- Caveats and challenges

# 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.
- Access requires detailed proposal, background check
- Projects often link survey and administrative data

# Person Identification Validation System (PVS)

- All records are linked probabilistically to a common “Reference File”
  - Contains personally identifiable information (PII) from the Social Security Administration Numident
  - Addresses are appended from IRS, Medicare, HUD, etc.
- Assigns Protected Identification Keys (PIKs)
- PIKs are used to unduplicate records and link across files
- Details at [The Person Identification Validation System \(PVS\) \(census.gov\)](https://www.census.gov/pvs)

# PVS Challenges

- Systematic differences in PIK coverage across groups (e.g. high coverage for workers and those receiving government benefits, low for immigrants without SSNs or ITINs)
- Some datasets do not contain necessary PII
- Cost for adding PIKs to researcher-provided data can be a barrier

# Administrative Data in the FSRDC

- 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): Complete list of addresses with lat/long, geography variables, and address ID (MAFID)
  - MAF Auxiliary Reference File (ARF): Links MAFID to individuals' PIKs on annual basis
- SNAP / TANF / WIC data from states

# Administrative Data in the FSRDC

- Longitudinal Employer-Household Dynamics (LEHD)
  - Link individuals to place of employment
  - Based on state unemployment insurance administrative records
  - Contains quarterly earnings for every company employee
- Universities: Measuring the Impacts of Research on Innovation, Competitiveness, and Science (UMETRICS)
  - Information on university grants (32 universities currently), including people and businesses paid by grants

# Administrative Data Challenges

- Securing approvals to share data and use for research
  - Example: LEHD research projects can use data from only a subset of states
- Data may not match desired concepts or variables
  - Example: LEHD does not include hours or occupation
  - Example: Work required to match place of birth in the Numident to standard geographies such as counties
- How to interpret zeros or missing data
  - Example: payroll=0