Linking ACS Data to Data from Other Sources: Issues to Consider

by Doug Hillmer

Independent consultant

Linking ACS Data to Data from Other Sources: Issues to Consider

Topics covered in this talk:

- **Why linking to other data sources is so important:**
- **Description of my current project – “What is rural”?**
- **Choosing the other data sources as an iterative process**
- **Questions about data sources to link – definitions; data collection methods; etc**
- **Vintage issues:**
- **Tools and techniques**
- **Questions I still have for my project**

Linking ACS Data to Data from Other Sources: Issues to Consider

Why linking to other data sources is so important:

- Every major work of social and economic research draws on multiple sources
- No single dataset has “all the answers”
- Datasets can represent different perspectives and “angles of attack” on a problem

Linking ACS Data to Data from Other Sources: Issues to Consider

Why linking to other data sources is so important:

An example: list of data sources for the Consumer Price Index (CPI) from BLS:

- data from the Billion Prices Project (Cavallo and Rigobon, 2016),
- retail scanner data,
- information on used cars from J.D. Power and Associates,
- stock exchange bid and ask prices and trading volume data,
- data on hospitals from the American Hospital Association,
- diagnosis codes from the Agency for Healthcare Research and Quality,
- administrative data on crude petroleum from the Energy Information Administration,
- administrative data on baggage fees from the U.S. Department of Transportation,
- SABRE data on airline pricing, and
- Medicare Part B reimbursement information.

Source: M. Horrigan, 2017 as quoted in the National Academy of Science report **"***Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps (2017) "*, p.19

Linking ACS Data to Data from Other Sources: Issues to Consider

- **Description of my current project – "What is rural"?**
- "Rural" is a term used often without being defined
  - This headline was for an article appearing in Politico in November, 2020
    - A few basic facts about Dunn County, Wisconsin:
      - Total pop is roughly 45,000
      - It is part of the Menomonie, WI micropolitan area
      - It is about a 30-minute drive from Eau Claire, Wisconsin which has a population of about 105,000
      - Farming plays virtually no role in the Dunn County economy

POLITICO

MAGAZINE

OPINION | PRIMARY SOURCE

Why Democrats Keep Losing Rural Counties Like Mine

I'm the chair of the local Democratic Party in a Wisconsin county that Donald Trump won. It wasn't for a lack of progressive organizing. It was because national Democrats have failed communities like mine.

Linking ACS Data to Data from Other Sources: Issues to Consider

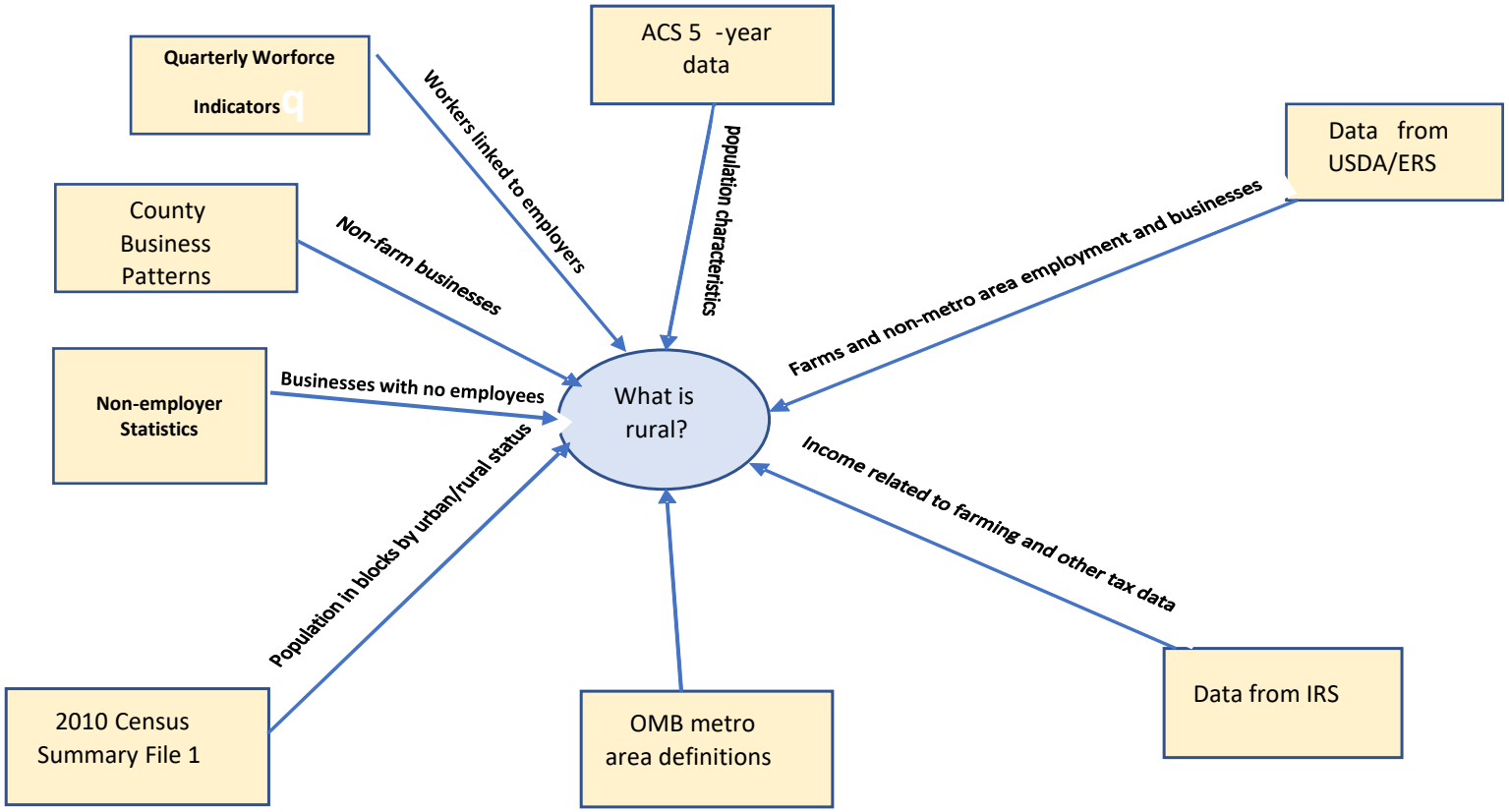
- **Description of my current project – "What is rural"?**
- Counties outside of all metropolitan and micropolitan areas are often used to represent "rural".
  - Ex. A research paper from the American Enterprise Institute, Feb., 2020, *title "Rural America's Stagnant Economic Performance"*. This paper basically equates nonmetro with rural.
- Census Bureau defines "rural" once every 10 years and only at the block level.
- OMB's process for defining metropolitan areas
- Comparing percent rural for counties inside metro areas to counties outside metro areas – ex. Total population in rural blocks in Colorado, metro vs nonmetro counties.

Linking ACS Data to Data from Other Sources: Issues to Consider

From the USDA Economic Research Service --		
Urban influence code descriptions	Average of Population_2010	Number of counties
Large-in a metro area with at least 1 million residents or more	362,021	472
Micropolitan adjacent to a large metro area	55,013	132
Micropolitan adjacent to a small metro area	45,876	245
Micropolitan not adjacent to a metro area	32,653	269
Noncore adjacent to a large metro area	21,770	149
Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents	9,709	164
Noncore adjacent to a small metro with town of at least 2,500 residents	21,193	344
Noncore adjacent to micro area and contains a town of 2,500-19,999 residents	15,212	184
Noncore adjacent to micro area and does not contain a town of at least 2,500 residents	7,129	189
Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents	15,674	125
Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents	5,078	184
Small-in a metro area with fewer than 1 million residents	124,536	764
	Total counties	3221

Linking ACS Data to Data from Other Sources: Issues to Consider

Current plan for data sources to use on the "What is rural?" project



Under consideration:

- County-level election data from 2012, 2016, 2020
- CDC Behavioral Risk Factor Surveillance Survey
- BLS Quarterly Census of Employment and Wages
- Hospitalization data from the American Hospital Association (fee-based data on 5,000+ hospitals)

Linking ACS Data to Data from Other Sources: Issues to Consider

Choosing the other data sources as an iterative process:

- Difficult to formulate an idea or hypothesis without looking at data
- Initial attempt to formulate and analyze a hypothesis can lead to problems and discoveries
- Data sources not seen at first may be discovered
- Initial approach may have to be re-considered and possibly tossed out

Linking ACS Data to Data from Other Sources: Issues to Consider

The problems associated with linking ACS and CBP:

- Goal: to discover types of non-farm businesses that may be emerging in less populated counties
- Problem: new disclosure avoidance rule for CBP effective since 2017
- How to assess the effects of this change? “Missing data” problem?

Linking ACS Data to Data from Other Sources: Issues to Consider

The problems associated with linking ACS and ERS typology data:

- Goal: to discover types of types of economic activities most prevalent in each county
- Definitional issues: use of “employment” not further specified (minimum number weeks worked?)
- Vintage issue : the ERS typology has not been updated since 2015. Is it still usable?

Linking ACS Data to Data from Other Sources: Issues to Consider

Some examples of problems associated with linking ACS and IRS data:

- Goal: to learn more about economic activity in each county
- Definitional issues: what are “qualified business income deductions”?
- Vintage issue: Most recent IRS data at the county level is from 2018. How do I “harmonize” these data with 5-year ACS data, 2015-2019?

Linking ACS Data to Data from Other Sources: Issues to Consider

How can we associate concepts from one data source to relevant concepts from another data source?

Example: ACS and USDA Agricultural Census data

- “farm laborers” vs people working in “crop and animal production”

Linking ACS Data to Data from Other Sources: Issues to Consider

Vintage issues:

- Continuous measurement means the ACS implies a “vintage issue” for almost any other data source
- Data that is collected and released on a regular basis (equal time intervals) – how often?
- How do I analyze the impact of associating measures from two datasets of different vintages?
- Data collected on an irregular basis – when was the last time; when does the provider expect the next release (will there be one)?

Linking ACS Data to Data from Other Sources: Issues to Consider

Tools and techniques:

- Techniques for determining if two estimates are (statistically) the same

Example. QWI measure for the "average number of workers" in New York state in 2019: 9,452,370
ACS estimate in 2019: 9,434,909
The QWI estimate is inside the 90% confidence interval for the ACS estimate
So, it appears that "workers" is being defined the same way in both datasets

Linking ACS Data to Data from Other Sources: Issues to Consider

Questions I still have for my project:

- Is “rural” partly a state of mind or attitude? Do I need to supplement my quantitative research with some qualitative research?
- Do I know people or organizations to contact when I have questions about data sources?
- Is there a role for realtime, transaction data (ex. point-of-purchase data, data on visits to doctors and healthcare providers)
- How do I address the “missing data” problem?
- Does the data collection method for any data source imply a systematic bias? How can we evaluate this?
- Are there other data sources I should include in my research?

Linking ACS Data to Data from Other Sources: Issues to Consider

There are many articles and books addressing issues that arise when using data from multiple sources.
One that I found particularly useful is
Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps (2017)
Published by the National Academies of Science. It can be downloaded for free at
https://www.nap.edu/catalog/24893/federal-statistics-multiple-data-sources-and-privacy-protection-next-steps
Another good reference is this talk:
https://www.nidcr.nih.gov/grants-funding/grant-programs/behavioral-social-sciences-research-program/comparing-and-linking-survey-data

There are websites that can be useful in searching for relevant data sources, especially mre recently published sources. Here are two which I have found helpful:

- Data.gov
- Amerigeo

And, there's always Google (using the Advanced Search to find more recent data sources helps)

Linking ACS Data to Data from Other Sources: Issues to Consider

Thank you.