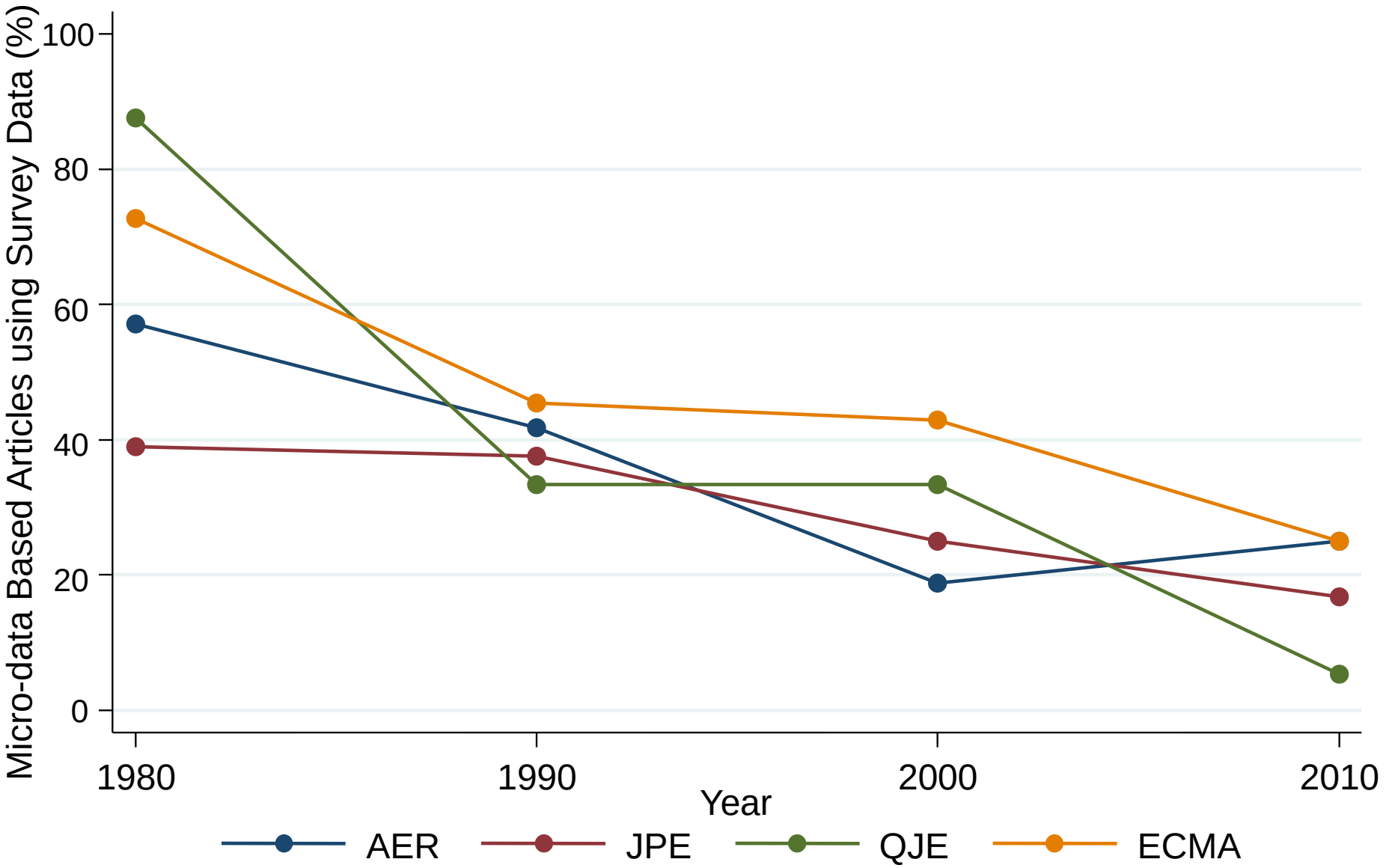
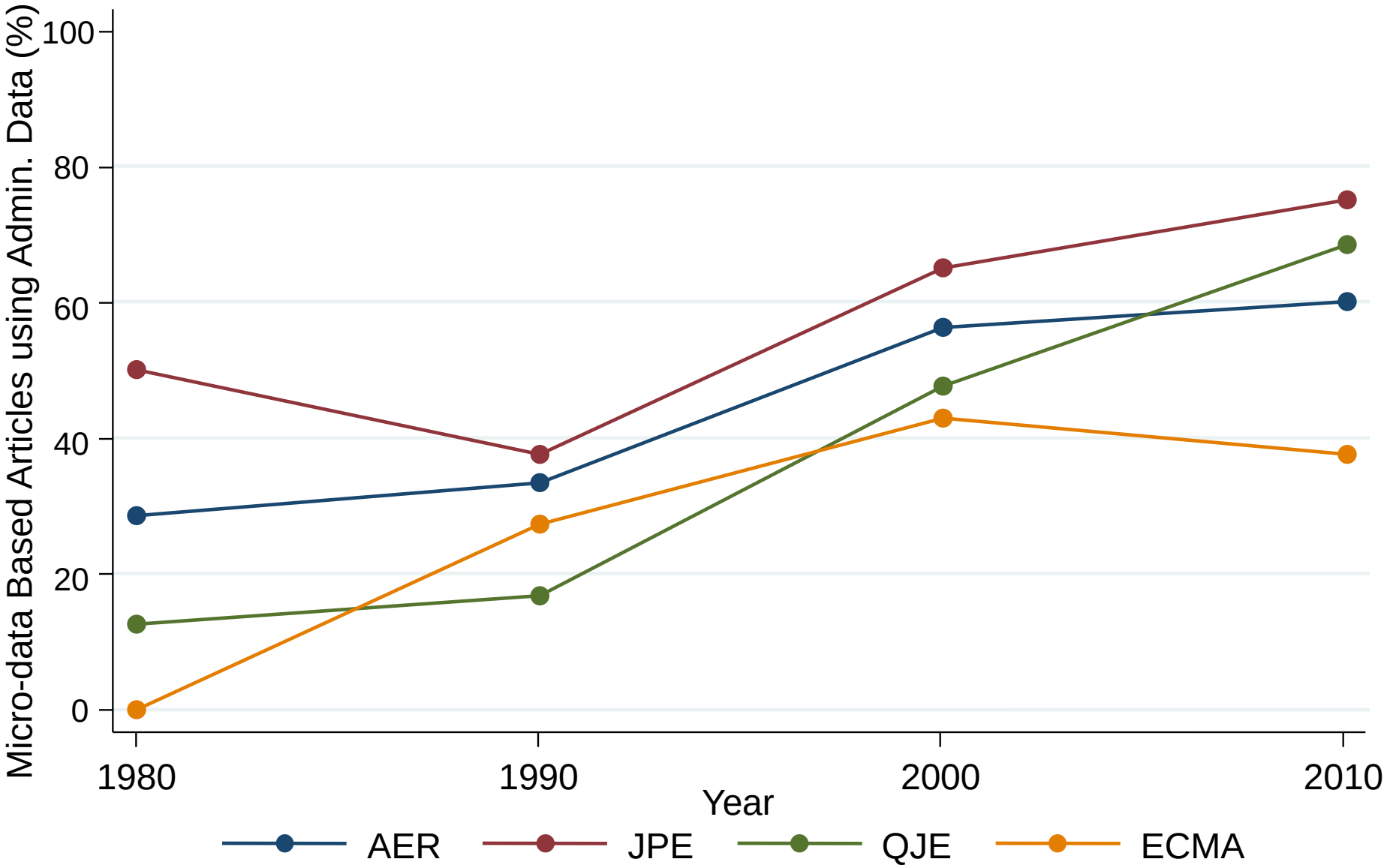
Use of Pre-Existing Survey Data in Publications in Leading Journals, 1980-2010



Note: “Pre-existing survey” datasets refer to micro surveys such as the CPS or SIPP and do not include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.

Use of Administrative Data in Publications in Leading Journals, 1980-2010



Note: "Administrative" datasets refer to any dataset that was collected without directly surveying individuals (e.g., scanner data, stock prices, school district records, social security records). Sample excludes studies whose primary data source is from developing countries.

What are the Benefits of Administrative Data?

1. Higher quality information: virtually no missing data or attrition
   - CPS non-response rate now 31% for income
   - SIPP attrition rate exceeds 30% within three years

2. Longitudinal tracking over long periods
   - Match rates of 95% over 20+ years in studies of long-term impacts of early childhood education [Chetty et al. 2011, Chetty, Friedman, Rockoff 2012]

3. Very large sample sizes: 2,000 times the size of the CPS
   - Can develop new non-parametric, quasi-experimental research designs

Administrative Data in the U.S.

- Demand for admin. data has led to shift to studying European countries
  - Population registers in Sweden, Austria, Germany, Norway, Denmark
  - Similar data infrastructure starting to be developed in the U.S.
  - This talk summarizes recent work on developing U.S. tax return data for research

The IRS Databank: A Population Panel Dataset for Tax Policy Research

Raj Chetty, Harvard and NBER
John N. Friedman, Harvard and NBER
Nathaniel Hilger, Harvard
Emmanuel Saez, UC Berkeley and NBER
Danny Yagan, Harvard

July 2012

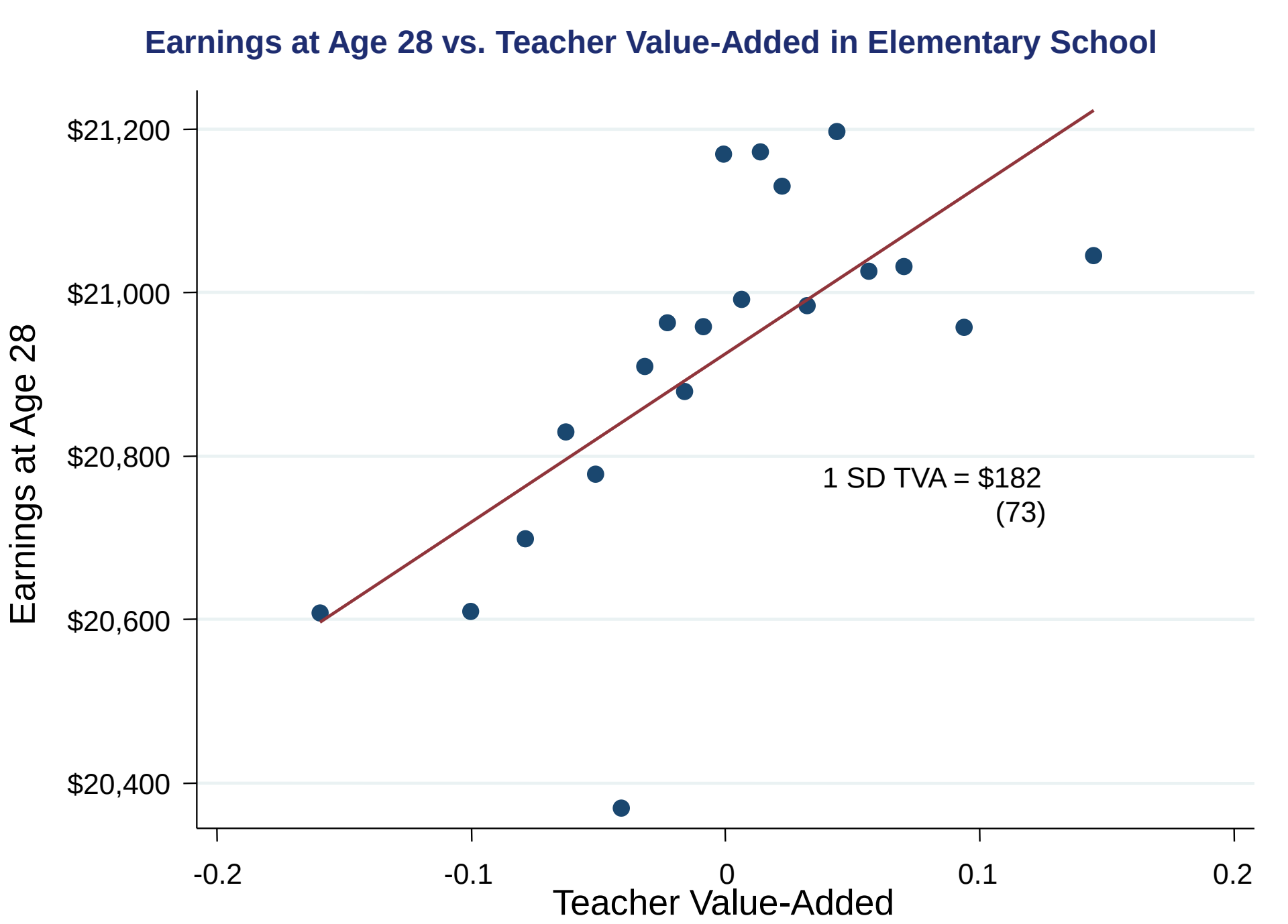
Prepared as part of IRS contract TIRNO-09-R-00007

United States Tax Data

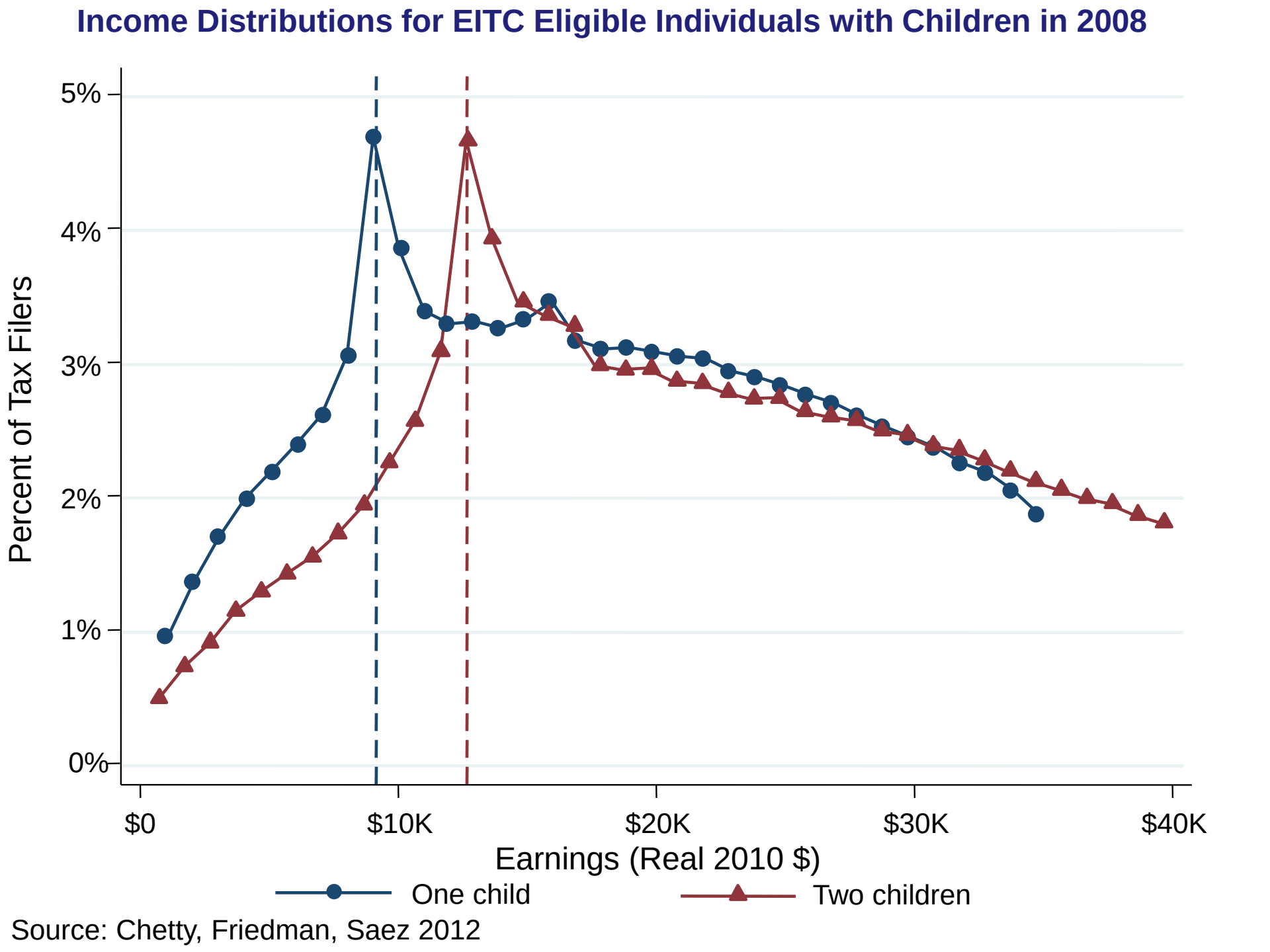
- In collaboration with Statistics of Income (SOI), we have built a panel dataset for research called the " IRS Databank" over the past two years
  - Combines raw variables spread across billion-observation databases
    - Optimized for audits rather than statistical research
  - Adding new variables and creating extracts is now "fast"
- Databank covers everyone in the U.S. on any tax form from 1996-2010
  - Approximately 6.7 billion rows
- This taxpayer data may only be accessed by those with statutory authority
  - Must be used for purposes related to tax administration as defined in Internal Revenue Code 6103

United States Tax Data

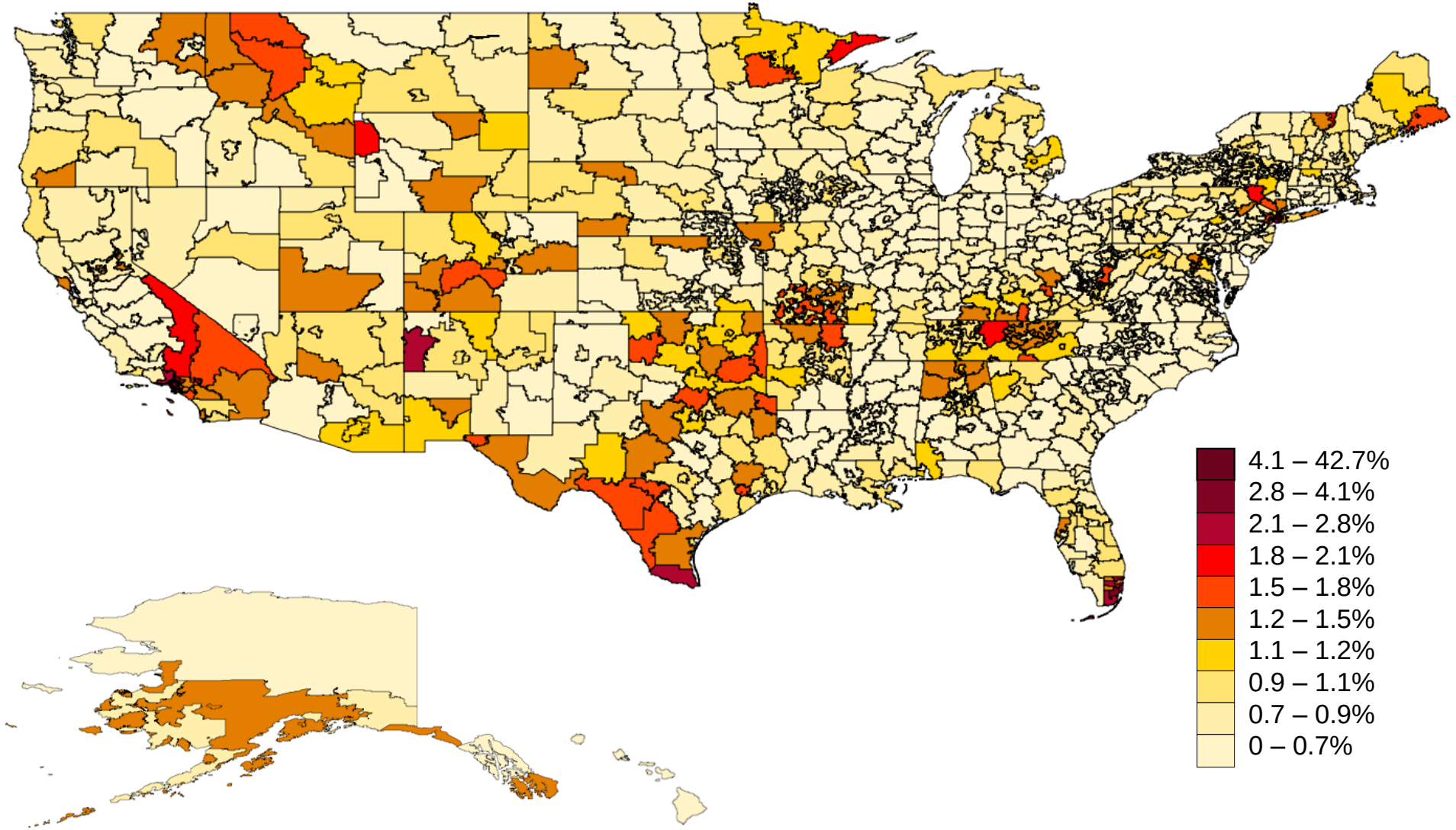
- Databank includes a rich set of information on individuals
  - Earnings from W-2's (covers non-filers)
  - Employer ID
  - College attendance
  - Retirement savings, charitable contributions
  - Housing and mortgage interest
  - Geographical location
  - Birth, death, marriage, children, family structure
- Analogous corporate databank contains income statement and some balance sheet information for 5 million firms per year, linked to workers



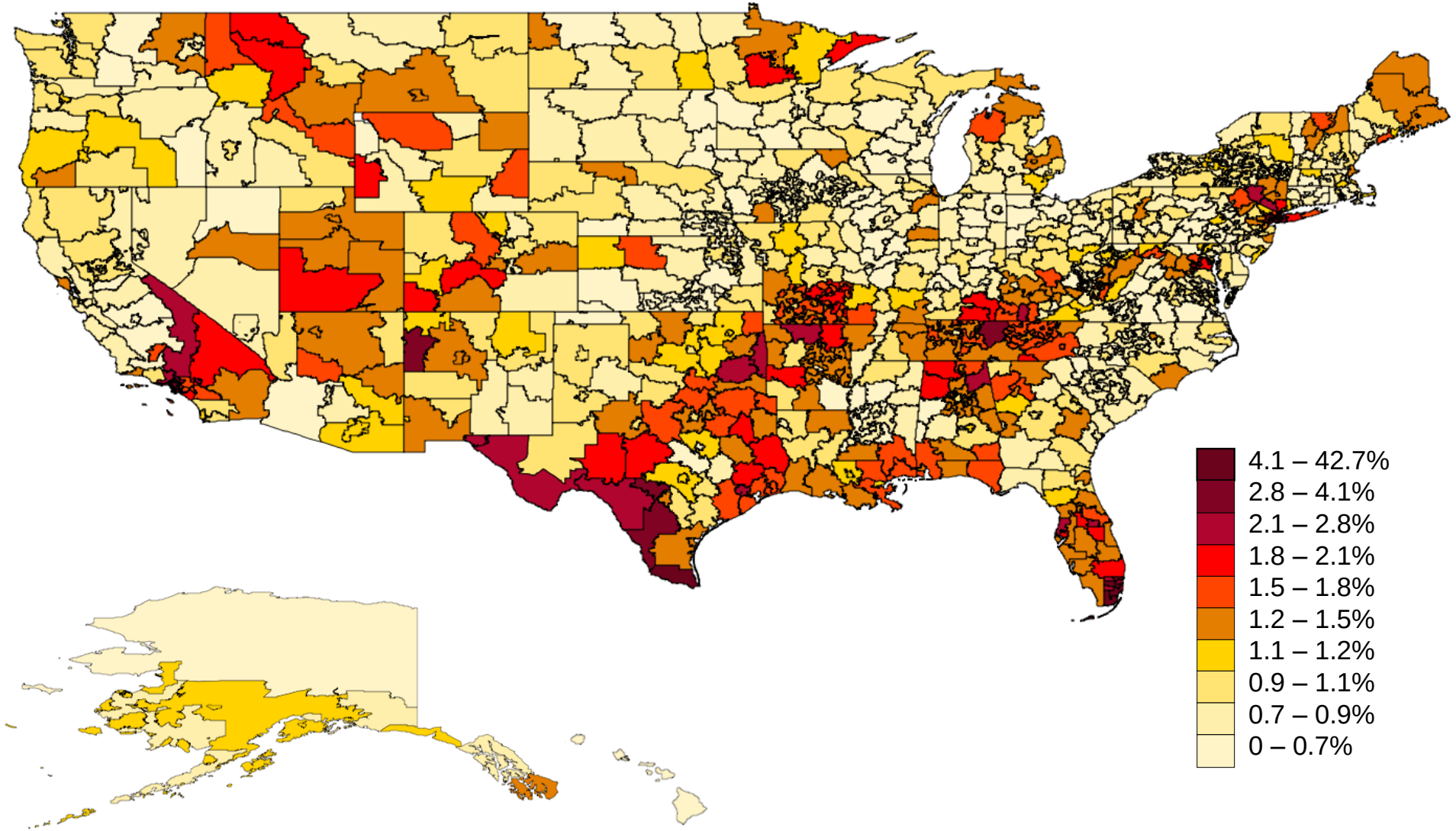
Source: Chetty, Friedman, Rockoff 2012


Income Distributions for EITC Eligible Individuals with Children in 2008
5%
4%
3%
2%
1%
0%
Percent of Tax Filers
$0
$10K
$20K
$30K
$40K
Earnings (Real 2010 $)
One child
Two children
Source: Chetty, Friedman, Saez 2012

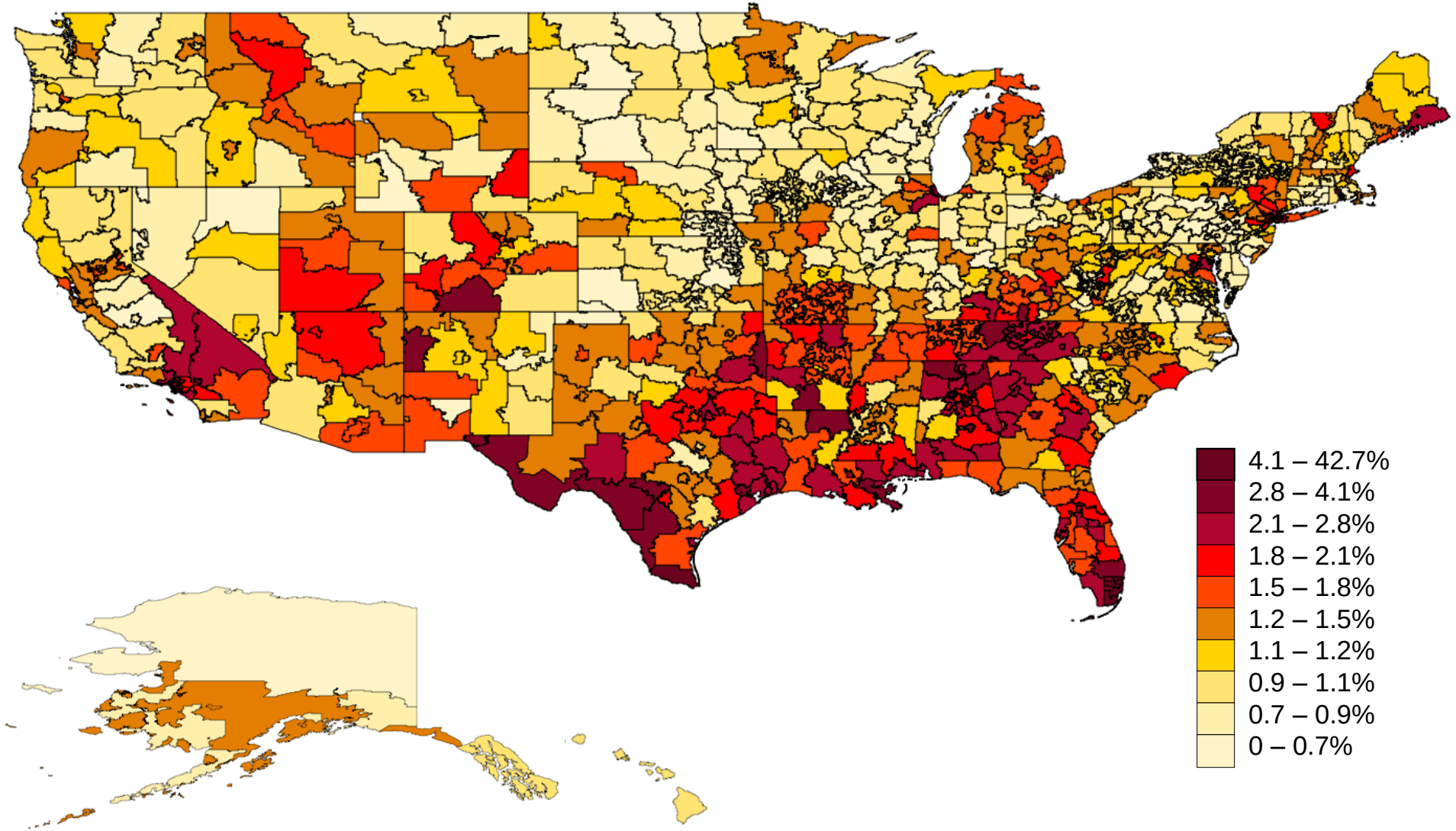
Self-Employed Sharp Bunching in 1996



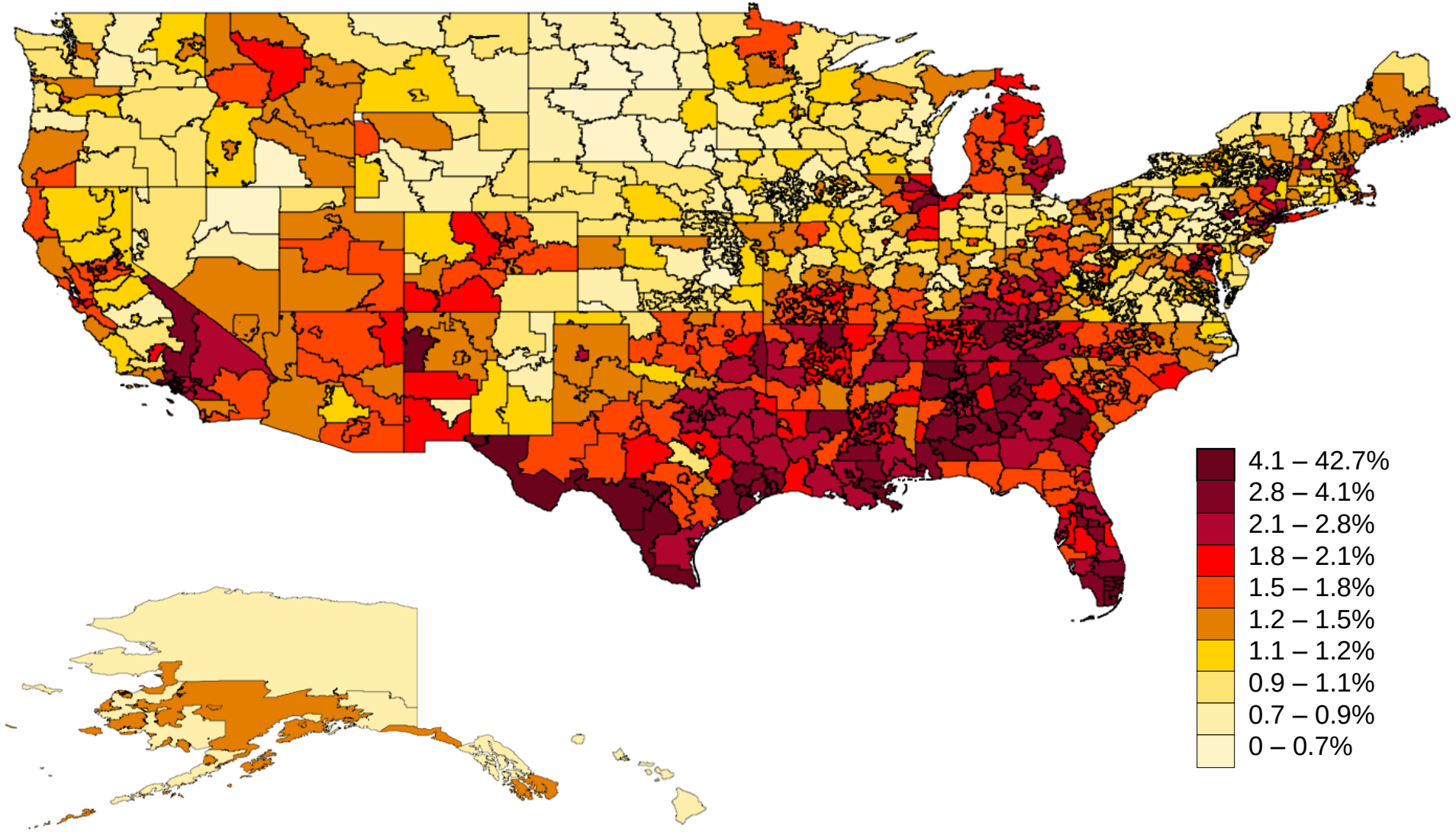
Self-Employed Sharp Bunching in 1999



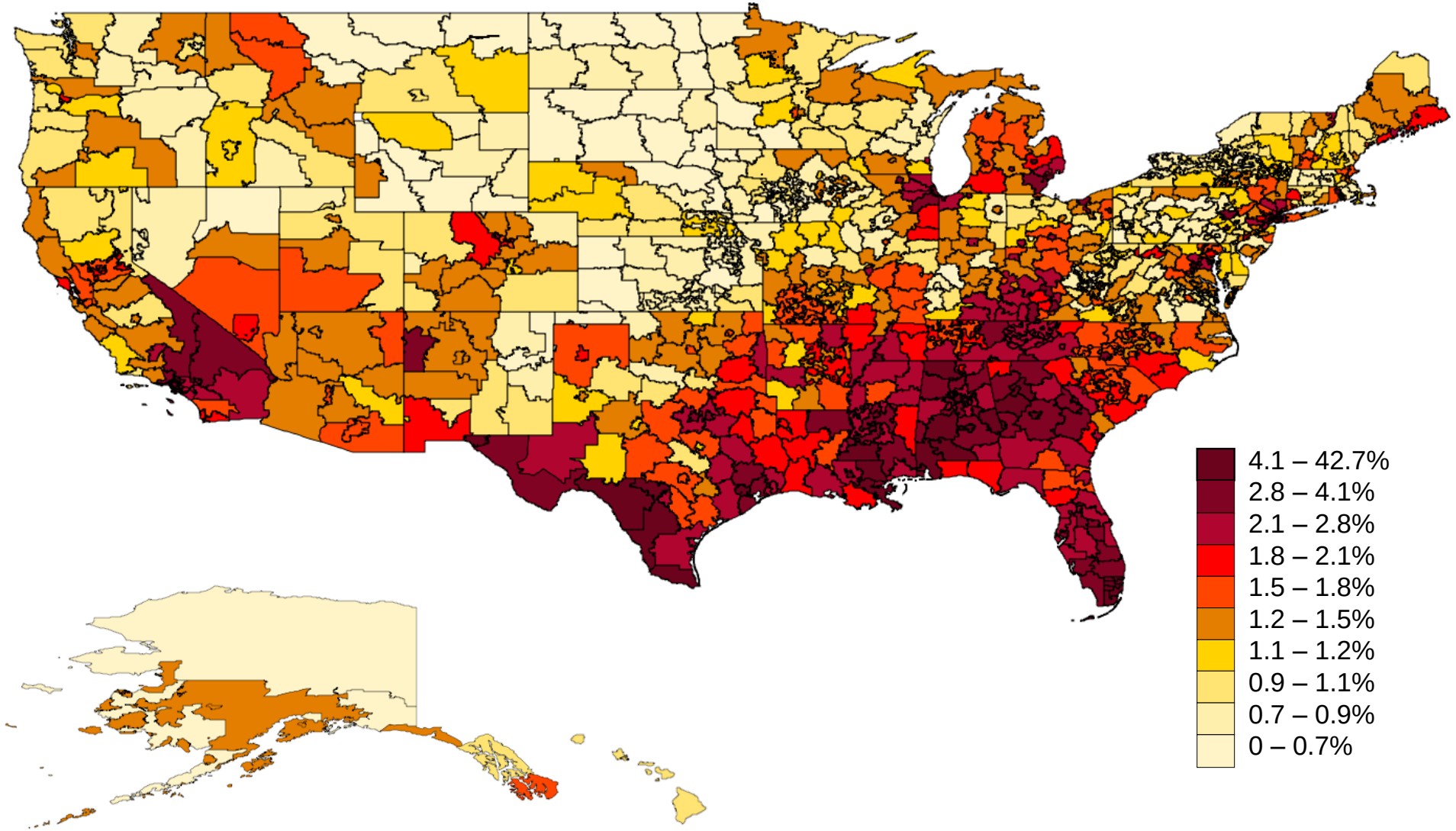
Self-Employed Sharp Bunching in 2002



Self-Employed Sharp Bunching in 2005



Self-Employed Sharp Bunching in 2008



Access to U.S. Tax Data

- SOI issued a call for proposals to work with tax data in Fall 2011
  - 50 submissions received and 20 research projects approved
- Researchers currently using three models for conducting research:
  1. Collaborate with Treasury/SOI staff to work with data on-site in Washington, D.C.
  2. Send programs to Treasury/SOI staff and access tabulations indirectly
  3. Work with data directly at IRS offices
- SOI planning to provide an update on external research program at Fall NBER PE program meeting