

---

## 5. CASE STUDIES USING ACS DATA

### Case Study #1: American Community Survey: Understanding the Basics

**Skill Level:** Introductory

**Subject:** Understanding the basics

**Type of Analysis:** Working with ACS tables; understanding 5-year estimates and margins of error

**Tool Used:** Data.census.gov

**Author:** Paul Overberg, *Wall Street Journal*

Too often journalists seek a simple fact, like a city's population or poverty rate, and stumble into the endless stacks of the U.S. Census Bureau's data library.

They find variations and cross-tabulations, overlapping data sets, and a geographical menagerie. It's a hall of mirrors for anyone who cannot tell a "family" from a "household," or "earnings" from "income" or how to parse through "Sex by Work Status in the Past 12 Months by Usual Hours Worked Per Week in the Past 12 Months by Weeks Worked in the Past 12 Months for the Population 16 to 64 Years."

But journalists who spend some time to learn about the most important data behind those portals—the American Community Survey, or ACS—find that it pays rewards. This huge annual survey, the "every-year census," is the crown jewel of U.S. social statistics. ACS data shape academic research, public discussion, policy, and law on many issues. The data are especially useful for journalists because they can be used to study many issues—like racial/ethnic diversity and representation—that are central to our pluralistic society and its promise of equality before the law.

The ACS also drives a huge share of federal funding: An estimated \$675 billion in federal funds are distributed to state and local areas each year based on census data. This in turn drives spending of hundreds of billions of dollars each year in state funding, especially for Medicaid and business investment.

Finally, the ACS is relevant to topics that seem unrelated to its content, like disease rates. This is because ACS provides some of the raw material for the population estimates that form the denominators for disease rates reported by the Centers for Disease Control and Prevention and other agencies.

The ACS operates on an industrial scale. Its staff produces hundreds of millions of facts each year from more than 2 million housing units and more than 5 million people. It converts the raw data into 11 billion data points across 80 kinds of geographic areas—more than 650,000 in all.

To get a sense of how this factory works, let's focus on a single ACS table that journalists often use—B19001: Household Income in the Past 12 Months (in 2016 Inflation-Adjusted Dollars) (see Figure 5.1). There is a lot to unpack here. Let's take it in small steps.

Figure 5.1. Sample Table From Data.census.gov

The screenshot shows the data.census.gov interface for Table B19001. The table title is "HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2016 INFLATION-ADJUSTED DOLLARS)". The table is for the United States and includes columns for Estimate and Margin of Error. The data is as follows:

	Estimate	Margin of Error
Total	118,840,065	+/-154,606
Less than \$10,000	7,963,784	+/-44,419
\$10,000 to \$14,999	5,699,549	+/-37,903
\$15,000 to \$19,999	5,815,371	+/-38,838
\$20,000 to \$24,999	5,928,371	+/-43,833
\$25,000 to \$29,999	5,501,899	+/-37,937
\$30,000 to \$34,999	5,777,895	+/-36,728
\$35,000 to \$39,999	5,338,202	+/-35,322
\$40,000 to \$44,999	5,380,237	+/-32,546
\$45,000 to \$49,999	4,734,622	+/-31,909
\$50,000 to \$59,999	9,209,029	+/-49,801
\$60,000 to \$74,999	11,850,363	+/-49,099
\$75,000 to \$99,999	14,672,995	+/-70,478
\$100,000 to \$124,999	10,312,781	+/-48,680
\$125,000 to \$149,999	6,355,018	+/-35,190
\$150,000 to \$199,999	6,924,913	+/-35,488
\$200,000 or more	7,595,040	+/-40,371

Source: U.S. Census Bureau, data.census.gov, <<https://data.census.gov>>.

This table is one of about 1,300 ACS Detailed Tables published each year. Most have names just like that: “B” followed by five digits. If you know that number, you can search for a table directly in data.census.gov. If you do not know the number, you can download a list of Table Shells from the Census Bureau’s Web site.<sup>36</sup> This is a good reference to keep handy because most ACS data products are built from Detailed Tables.

In most cases, each ACS table is published each year for any geographic area with at least 65,000 people. That includes every state, congressional district, metropolitan statistical area, lots of cities, and about a quarter of counties. It takes 5 years to accumulate enough responses so that the Census Bureau can publish the same table for less-populated areas. Those can range from small towns to even neighborhoods. They are also published each year, but each version covers the most recent 5 years.

There is more to the 5-year data than that. The Census Bureau restates dollars to reflect their value in the most recent year. It applies the most recent poverty thresholds. It retallies data to reflect changes in geographic boundaries, which is critical when city or metro boundaries change.

<sup>36</sup> U.S. Census Bureau, American Community Survey (ACS), Table Shells and Table List, <[www.census.gov/programs-surveys/acs/technical-documentation/table-shells.html](http://www.census.gov/programs-surveys/acs/technical-documentation/table-shells.html)>.

---

Right under its name, this table declares: “Universe: Households.” This is key. The universe is the total that is being counted. In this case, it is “households,” which has a very specific meaning in the ACS. Most table universes are people, households, housing units, or subsets of those three. Some examples:

- Mortgage status: “Owner-occupied housing units.”
- People enrolled in school: “Population 3 years and over.”
- Language spoken at home: “Population 5 years and over.”
- Fertility rate: “Women 15 to 50 years.”
- Highest level of education: “Population 25 years and over.”

Checking a table’s universe helps you sharpen the questions you are asking, narrow your search, and shape how you will describe the data.<sup>37</sup>

For example, reporters often get confused between “household income” and “family income.” If you want to write about everyone in households, you would pick “household income.” Every occupied housing unit contains a household. That includes homes where just one person lives—28 percent of all households—as well as the homes of unrelated people, such as unmarried partners or roommates. “Family income” only covers families, which are two or more people related by blood, marriage, or adoption who live together. One measure of the difference can be seen in their median incomes. In 2016, the U.S. median household income was \$57,617; the median family income, \$71,062.

What period does this table cover? It says income “in the past 12 months,” but do not call it “2016 income.” Confusingly, the 2016 1-year estimates cover the 23 months that end with November 2016. Why? Since the ACS interviews respondents monthly and asks for income received during the “past 12 months,” it ends up with 12 different reference periods for income data collection in a single calendar year. The data are then put in terms of constant 2016 inflation-adjusted dollars by using the CPI-U-RS inflation factors from the Bureau of Labor Statistics.<sup>38</sup> For example, in 2016 a respondent filling out the survey in February would have a reference period of February 2015 through January 2016. A respondent filling out the survey in October would have a different reference period, October 2015 through September 2016. The inflation-adjustments put these different reference periods in constant dollars; in this example, “2016 inflation-adjusted dollars.”

Do people really tell the Census Bureau their incomes? Mostly. But you can look it up and couch your words accordingly. The ACS comes with 100 question-specific, quality-checking tables whose numbers begin with “B99.” Table B99192 tells us that the Census Bureau did not have to allocate any income responses for 65 percent of households in 2016.<sup>39</sup> Another 16 percent of households had 100 percent of their income allocated. The remaining 19 percent had either partial answers that had to be completed or a dollar value of zero allocated. How does the Census Bureau know how to impute missing values? It uses a set of rules to pull information from other parts of the household’s form. Then, if necessary, it uses statistical methods to pull data from a similar household nearby. Not surprisingly, people leave income questions blank more often than almost any other type. Just 1.7 percent of respondents leave the age and birthdate questions blank.

Almost all ACS tables carry a margin of error for each number of interest. Journalists tense up about the word “error,” so let’s work through what it means here. It’s a term to describe variation inherent in random samples. The ACS is a very big, complex version of the scientific opinion polls that some news organizations sponsor. (Guess where the demographic benchmarks for those polls come from?) If you had enough money to poll 1,000 different random samples of the U.S. population, each would have a slightly different makeup. One might overrepresent women, another homeowners. If you kept polling and averaged the results, you would see a bell curve of results form for, say, the percentage of women. Most often, you would get 51 percent. A little less often, you would get 50 percent or 52 percent. Most pollsters work at the 95 percent confidence level, which produces the familiar “ $\pm 3$  percentage points.” That means that 19 times out of 20 polls (90 percent of the time), you’d get 51 percent of women,  $\pm 3$  percentage points. And one time out of 20, you’d get a sample outside that margin—women would represent less than 48 percent or more than 54 percent.

The Census Bureau reports ACS margins of error at the 90 percent confidence level. Looking back to our table, 7,595,040 households ( $\pm 40,371$ ) had incomes of \$200,000 or more. That means we can be pretty comfortable

---

<sup>37</sup> The Census Bureau produces a series of Appendix Tables for data users who want more information about ACS Detailed Tables, including table universes. The 1-year and 5-year appendixes can be downloaded as Excel files from the Summary File Documentation page at [www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html](http://www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html).

<sup>38</sup> U.S. Bureau of Labor Statistics, CPI Research Series Using Current Methods (CPI-U-RS), [www.bls.gov/cpi/research-series/home.htm](http://www.bls.gov/cpi/research-series/home.htm).

<sup>39</sup> A commonly used approach to imputation (a statistical procedure to fill in missing responses) is known as hot-deck allocation, which uses a statistical method to supply responses for missing or inconsistent data from responding housing units or people in the sample who are similar.

---

writing “7.6 million” and saying they outnumber the “6.9 million” (6,924,913, ± 35,488) who made \$150,000 to \$200,000. See how handy rounding and error margins can be?

Instead of “margin of error,” think of this as a way to quantify the squishiness of the data. This is something that journalists deal with all the time—the ambiguity of the real world. Error margins just tell us how to precisely state the numbers and tune the words around them.

It’s also important to remember that surveys, even massive ones like ACS, are better tools for expressing quantitative relationships than totals. You can report rounded ACS totals, but the main strength of the ACS lies in its ability to tell you a group’s share of the population, or the ratio of two groups in the population, or how a group’s share has shifted across time.

Inevitably, you will want to add or subtract ACS totals. Let’s say we want to know the number and share of households with more than \$100,000 in annual income. It’s easy to just add the table cells, but to get the margin of error, you cannot just sum the same cells’ margins of error. The relevant formulas are available in the section on “Calculating Measures of Error for Derived Estimates” in the Census Bureau’s handbook on *Understanding and Using American Community Survey Data: What All Data Users Need to Know*.<sup>40</sup>

Finally, journalists often wonder “How big an error margin is too big?” Statistics teachers say, “It depends.” They are right, but that is less than helpful. Consider a simpler measure of data squishiness called the coefficient of variation (CV). Without getting too deep into statistics: Divide the margin of error for the ACS number you care about by 1.645. The result is what is called its standard error. Divide that by the ACS number itself and multiply by 100. A CV of 10 percent does not seem too wobbly, but one of 50 percent probably is for most purposes.

Finally, journalists who do learn a bit about the ACS quickly appreciate that it offers honest answers. Too often, reporters get fed dubious numbers by people who are not keen to explain where they came from or how they were massaged. By comparison, ACS documentation spells out:

- Exactly how each question was asked, including a facsimile of the questionnaire.
- Why each question was asked.
- The share of surveyed households each year that actually responded to each question in any given area. For instance, in 2016, the Census Bureau managed to get 98 percent of people to provide their race but just 84 percent of households to provide their income.
- How often each question gets left blank, and how the Census Bureau fills in blanks.

Like most people, the ACS becomes a better source as you learn more about what it knows and how to interrogate it. For example, you can ask questions that can’t be answered by any of its hundreds of tables if you know how to use its microdata to create a custom table.

So, the ACS can seem complicated, but it repays the time you invest in learning it with powerful ways to find and tell stories on many subjects and at many scales.

---

<sup>40</sup> U.S. Census Bureau, American Community Survey (ACS), *Understanding and Using American Community Survey Data: What All Data Users Need to Know*, <[www.census.gov/programs-surveys/acs/guidance/handbooks/general.html](http://www.census.gov/programs-surveys/acs/guidance/handbooks/general.html)>.

## Case Study #2: Reporting Poverty Trends in Michigan

**Skill Level:** Introductory

**Subject:** Poverty trends

**Type of Analysis:** Analyses of economic trends within and across large communities

**Tools Used:** Data.census.gov and spreadsheet

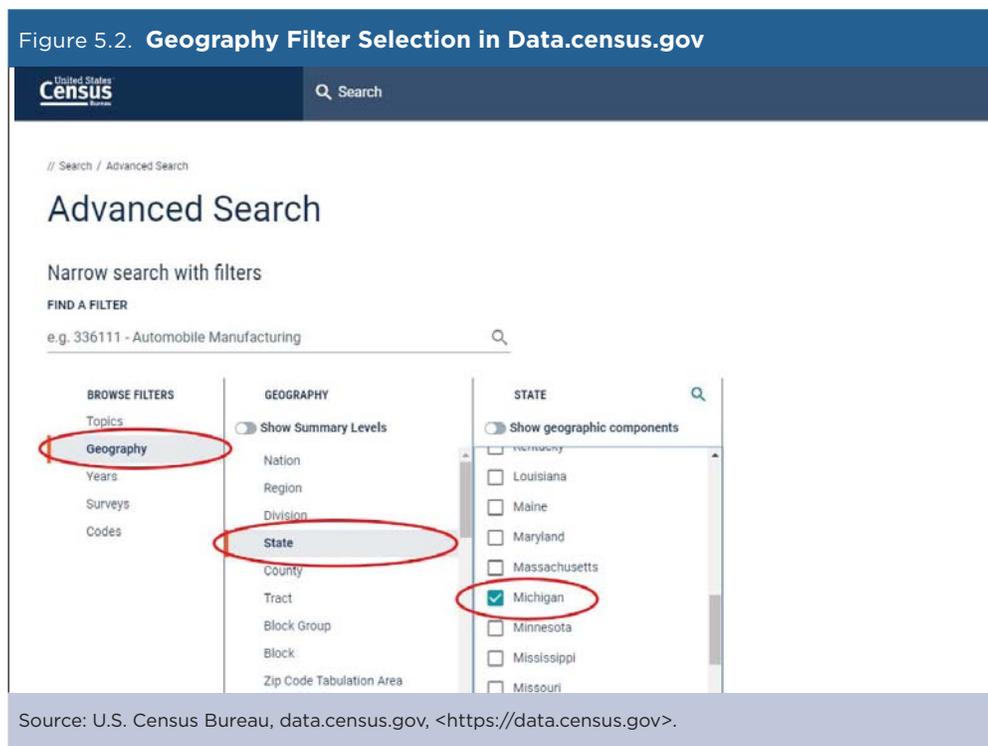
**Author:** Kristi Tanner, *Detroit Free Press*

Population and demographic estimates from the U.S. Census Bureau are a great resource for building data sets and writing about local demographic trends. American Community Survey (ACS) 1-year estimates are the most current data available for large geographic areas—65,000 residents or more—including cities, counties, metropolitan areas, and states. The ACS achieved full, nationwide implementation in 2005 for the household population and was expanded to cover the full population (including group quarters—such as college dormitories) in 2006.

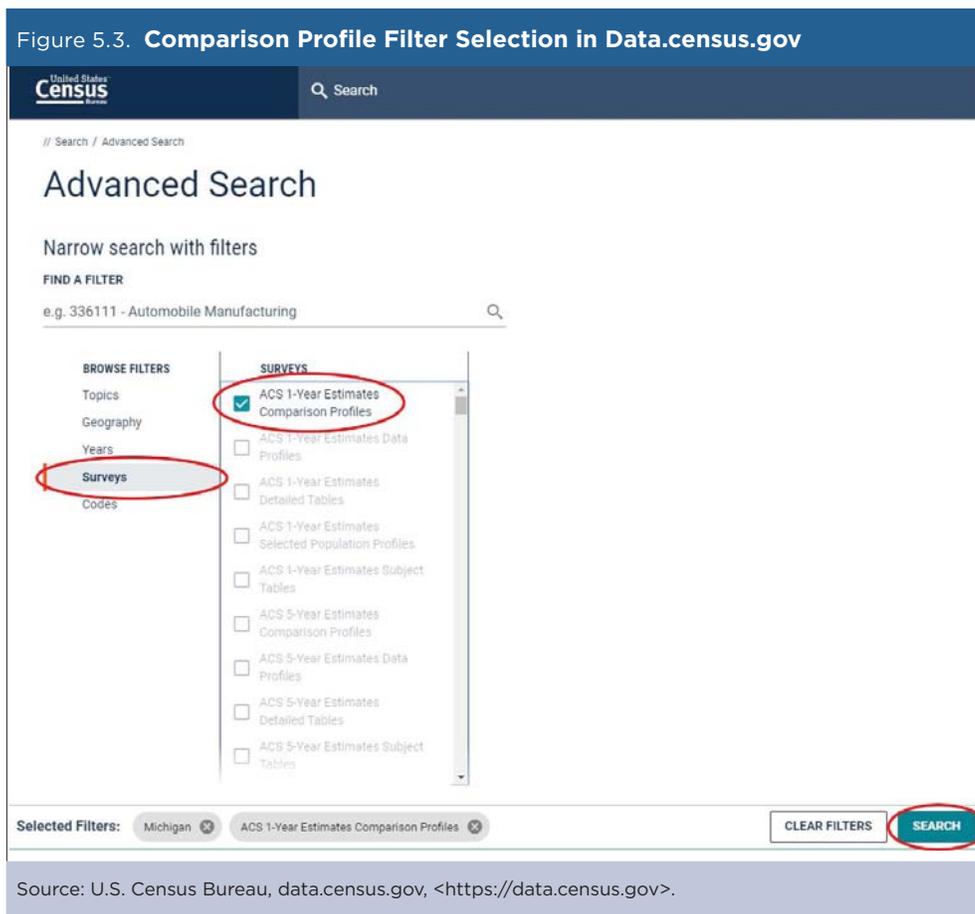
Included in the annual ACS release are Comparison Profiles. These reports cover hundreds of variables and compare a current statistic to each year’s values up to 4 years prior; any changes that are statistically significant are highlighted in a separate column in the report.

To see if Michigan’s poverty rate changed last year, look at the latest Comparison Profile report:

- Go to the data.census.gov Web site at <<https://data.census.gov>>.
- Click on “Advanced Search” under the search bar. This will bring you to the Advanced Search page.
- Begin with the Geography filter. Select “Geography” in the navigation pane on the left side of the screen to display a list of available geographies.
- Select “State” and then select “Michigan” from the “State” filter (see Figure 5.2).



- Next, choose the “Surveys” filter and select “ACS 1-Year Estimates Comparison Profiles.”
- Both filters should appear in the “Selected Filters” at the bottom of the page.
- Next, click on “Search” in the lower right corner of the page (see Figure 5.3).



Comparison Profiles are organized by social, economic, housing, and demographic subject areas. Make sure you select the most recent data set available. When this analysis was initially conducted, the 2016 ACS 1-year estimates were the most current.<sup>41</sup> To find data on poverty and income, click on the link “Comparative Economic Characteristics” (see Figure 5.4).

<sup>41</sup> ACS 5-year estimates are useful for small geographic areas and are available down to the block group level.

Figure 5.4. Choosing a Comparison Profile

The screenshot shows the U.S. Census Bureau website interface. At the top, there is a search bar and navigation tabs for 'ALL', 'TABLES', 'MAPS', and 'PAGES'. Below the navigation, it indicates 'About 1,178 results | Filter'. The main content area is titled 'Tables' and lists several comparison profiles. The profile 'COMPARATIVE ECONOMIC CHARACTERISTICS' is circled in red. Below this profile, a table is displayed with columns for '2018 Estimate', '2017 Estimate', '2018 - 2017 Statistical Significance', '2016 Estimate', and '2018 - 2016 Statistical Significance'. The table includes rows for 'EMPLOYMENT STATUS', 'Population 16 years a...', 'In labor force', 'Civilian labor force', 'Employed', and 'Unemployed'.

	2018 Estimate	2017 Estimate	2018 - 2017 Statistical Significance	2016 Estimate	2018 - 2016 Statistical Significance
EMPLOYMENT STATUS					
Population 16 years a...	8,092,018	8,052,502	*	8,002,285	
In labor force	61.5%	61.5%		61.2%	
Civilian labor force	61.5%	61.4%		61.1%	
Employed	58.2%	57.8%	*	57.3%	
Unemployed	3.3%	3.6%	*	3.8%	

Source: U.S. Census Bureau, data.census.gov, <<https://data.census.gov>>.

Next, Select “Customize Table” (see Figure 5.5).

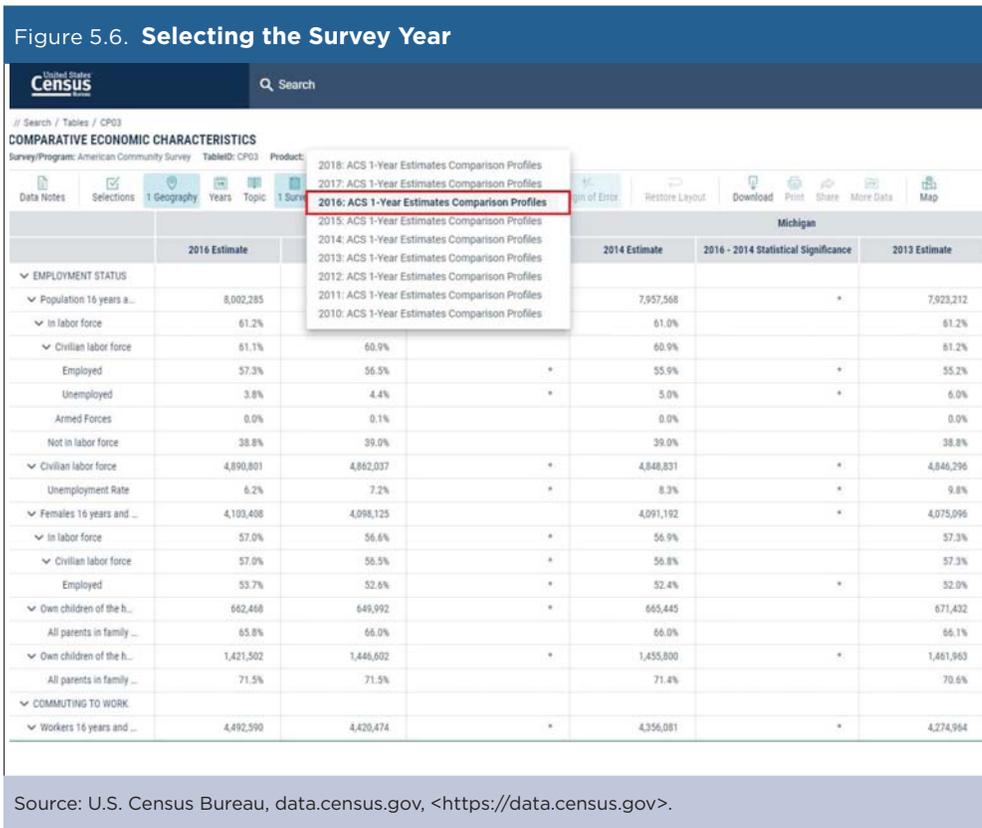
Figure 5.5. Customize Table in Data.census.gov

The screenshot shows the U.S. Census Bureau website interface with the 'COMPARATIVE ECONOMIC CHARACTERISTICS' table selected. The 'CUSTOMIZE TABLE' button is circled in red. The table displays data for '2018 Estimate', '2017 Estimate', and '2018 - 2017 Statistical Significance'. The table includes rows for 'EMPLOYMENT STATUS', 'Population 16 years a...', 'In labor force', 'Civilian labor force', 'Employed', 'Unemployed', 'Armed Forces', 'Not in labor force', 'Civilian labor force', 'Unemployment Rate', 'Females 16 years and ...', 'In labor force', 'Civilian labor force', 'Employed', 'Own children of the h...', 'All parents in family ...', 'Own children of the h...', and 'All parents in family ...'.

	2018 Estimate	2017 Estimate	2018 - 2017 Statistical Significance
EMPLOYMENT STATUS			
Population 16 years a...	8,092,018	8,052,502	*
In labor force	61.5%	61.5%	
Civilian labor force	61.5%	61.4%	
Employed	58.2%	57.8%	*
Unemployed	3.3%	3.6%	*
Armed Forces	0.0%	0.1%	
Not in labor force	38.5%	38.5%	
Civilian labor force	4,973,988	4,945,632	*
Unemployment Rate	5.3%	5.9%	*
Females 16 years and ...	4,142,957	4,121,837	*
In labor force	57.0%	57.1%	
Civilian labor force	57.0%	57.1%	
Employed	54.2%	54.0%	
Own children of the h...	663,450	656,743	
All parents in family ...	67.5%	66.3%	
Own children of the h...	1,397,836	1,407,481	
All parents in family ...	71.7%	71.9%	

Source: U.S. Census Bureau, data.census.gov, <<https://data.census.gov>>.

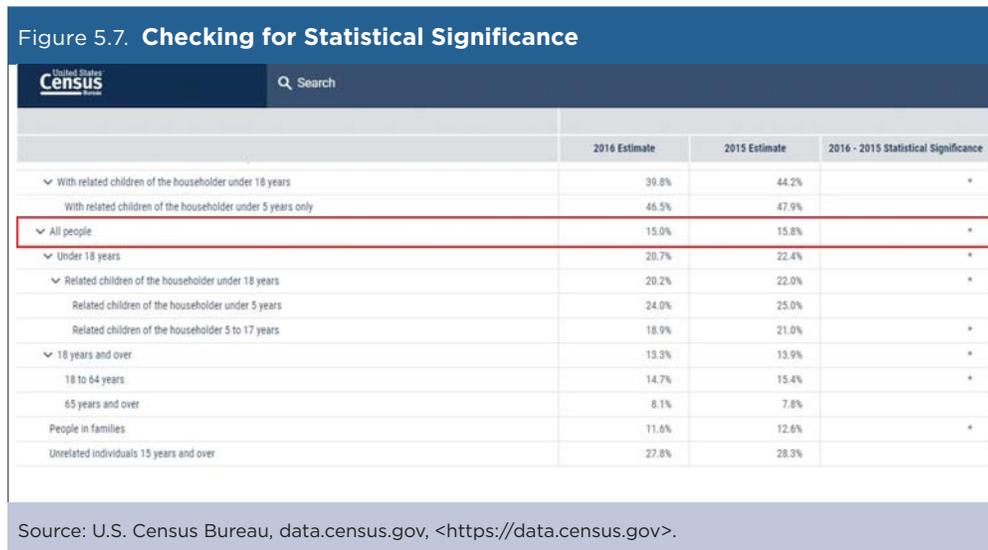
Check the data set, year, and geography on the report. Select the desired survey year by clicking on the current “Product” selection. Again, for the purposes of this case study, we are using 2016 ACS 1-year estimates. The header should read “2016 American Community Survey 1-Year Estimates” and the geography—“Michigan” (see Figure 5.6).



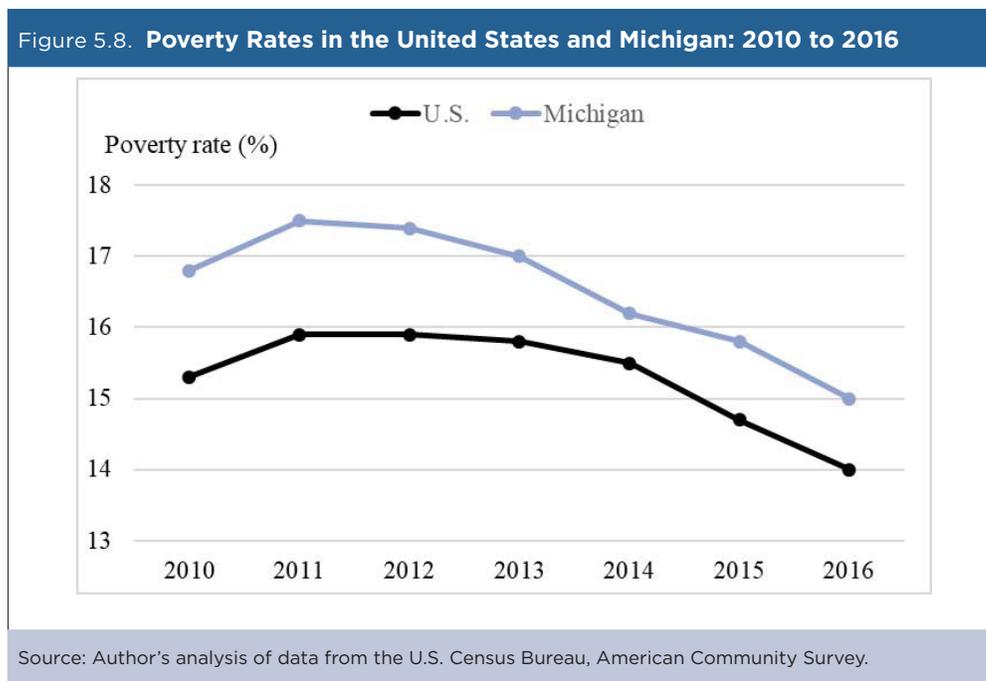
The poverty data are near the bottom of the table. If you want to expand the width of a column, hover over a column border in the shaded section at the top of the table and click and drag the border to the desired width. Poverty percentages are found below the “Percentage of Families and People Whose Income in the Past 12 Months Is Below the Poverty Level” row.

In 2016, Michigan’s poverty rate for all residents was 15 percent, a drop of nearly 1 percentage point from 2015. The asterisk in the column labeled “2016 - 2015 Statistical Significance” identifies a significant difference between the two estimates at a 90 percent confidence level. If there is no asterisk in the column between the 2 comparison years, for example the poverty rate of individuals aged 65 years and over, you can interpret the current year’s statistic as unchanged—in this case at about 8 percent in 2016 (see Figure 5.7).

In 2016, Michigan's poverty rate for all residents was 15 percent, a drop of nearly 1 percentage point from 2015. The asterisk in the column labeled "2016 - 2015 Statistical Significance" identifies a significant difference between the two estimates at a 90 percent confidence level. If there is no asterisk in the column between the 2 comparison years, for example the poverty rate of individuals aged 65 years and over, you can interpret the current year's statistic as unchanged—in this case at about 8 percent in 2016 (see Figure 5.7).



By clicking on earlier years of this report, at the top of the page, you can see that the poverty rate for Michigan residents (all people) has been steadily declining since 2014. Figure 5.8 shows the results of this analysis, which was featured in a recent article in the *Detroit Free Press*.<sup>42</sup>



<sup>42</sup> Kristi Tanner, *Detroit Free Press*, "Census data: For a fourth year, Michiganders see incomes rise," 2017, <[www.freep.com/story/news/2017/09/14/michiganders-making-more-cash-even-detroit-new-stats-say/660518001/](http://www.freep.com/story/news/2017/09/14/michiganders-making-more-cash-even-detroit-new-stats-say/660518001/)>.

Comparison Profiles are a great way to look for statistically significant differences in estimates over time. They allow you to say with a certain amount of confidence, in this case for individual poverty, that rates continue to decline.

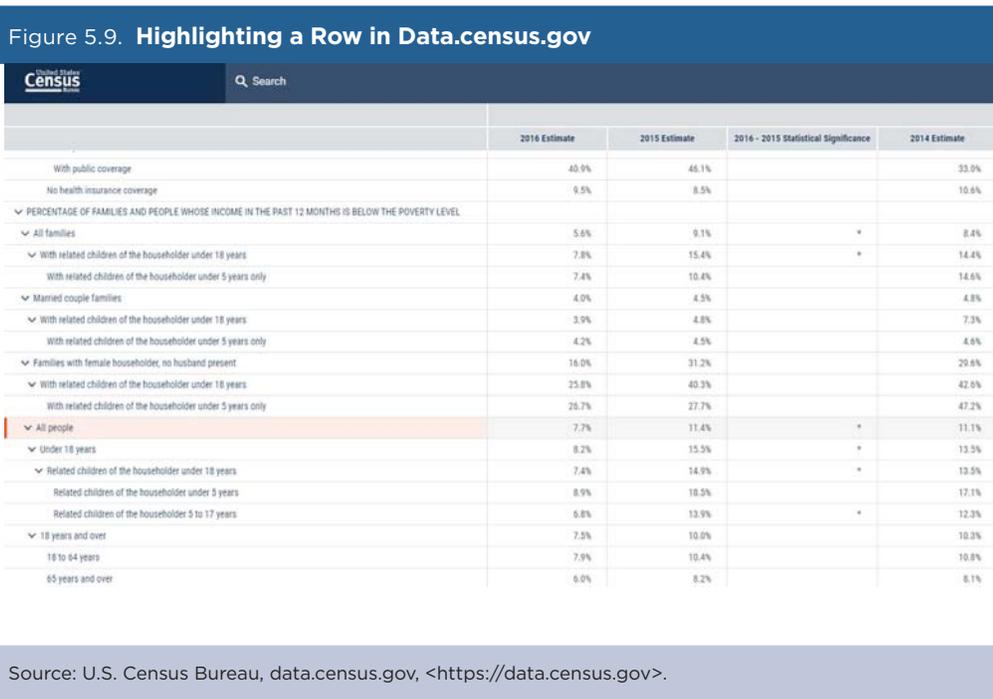
To see how Michigan’s poverty trend compares to the U.S. poverty trend:

- Scroll to the top of the table and select “1 Geography” to open the geography filter.
- Click on “Nation” and then select “United States.”
- “United States” and “Michigan” should now be in your “Selected Geographies.”
- Click on the “Close” button in the lower right corner to view the table with your updated geography selections.
- Check year-over-year trends to see if the change in poverty for all U.S. residents is statistically significant.

If you need additional years of data, follow the same steps used to select 2016 estimates by clicking on your “Product” selection at the top of the page.

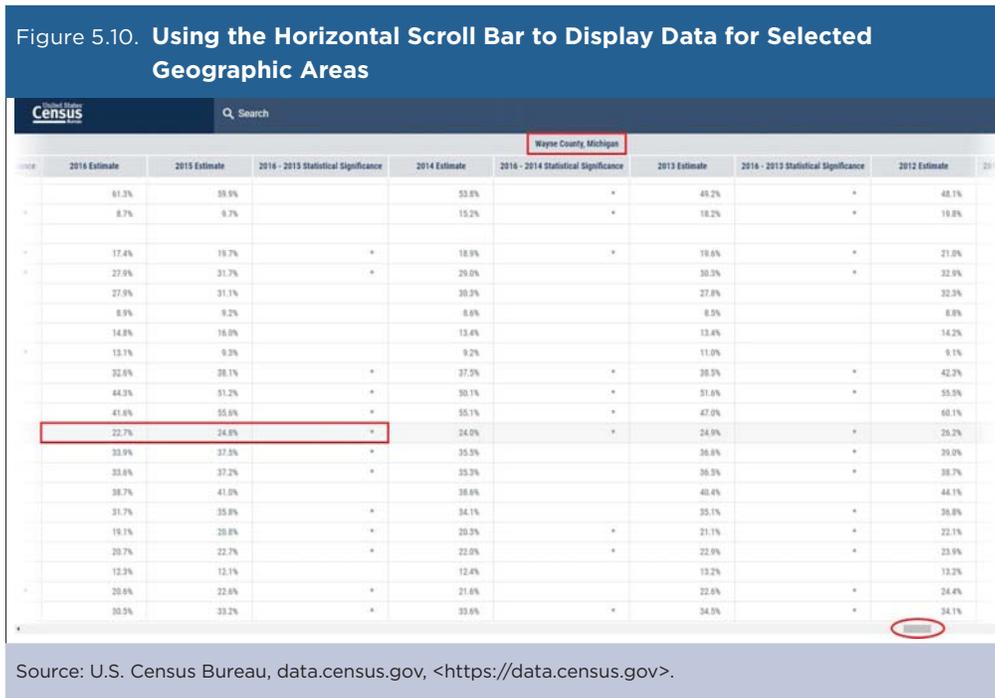
You can also use Comparison Profiles to compare trends across multiple geographic areas. To determine, for example, which large Michigan counties saw poverty rates drop in 2016, complete the following steps:

- Scroll to the top of the table and select “2 Geographies” to open the geography filter.
- Click on “County” and then select “Michigan” from the list.
- Select “All counties in Michigan.”<sup>43</sup>
- Close the filter by clicking on the “Close” button in the lower right corner.
- Scroll to the bottom of the table and click on the cell that says “All people” below the “Percentage of Families and People Whose Income in the Past 12 Months is Below the Poverty Level.” This highlights the desired row of data (see Figure 5.9).



<sup>43</sup> There are 83 counties in Michigan. Since the ACS 1-year data set is selected, only counties with 65,000 or more residents will be available—a total of 29. For statistics on less populous counties, use the ACS 5-year estimates.

Use the left/right arrow keys or the horizontal scroll bar at the bottom of the page to view the data for different counties in Michigan (see Figure 5.10). In Michigan, six counties had statistically significant changes in their poverty rate in 2016; four declined and two increased. Wayne County, the state's largest county, saw a 2.1 percentage-point decline in its poverty rate, from 24.8 percent in 2015 to 22.7 percent in 2016.



This example used Comparison Profile reports to analyze poverty trends for residents of Michigan and the United States, including identifying statistically significant changes in poverty rates by county. Data users working with other ACS estimates—such as those from the Detailed Tables—can test for significant difference among estimates by using the Census Bureau’s Statistical Testing Tool.<sup>44</sup>

<sup>44</sup> U.S. Census Bureau, Statistical Testing Tool, <[www.census.gov/programs-surveys/acs/guidance/statistical-testing-tool.html](http://www.census.gov/programs-surveys/acs/guidance/statistical-testing-tool.html)>.

## Case Study #3: Census Reporter

**Skill Level:** Advanced

**Subject:** File Transfer Protocol (FTP), Table structure, geographic areas

**Type of Analysis:** Working with large ACS data sets; repackaging ACS data

**Tools Used:** PostgreSQL, FTP

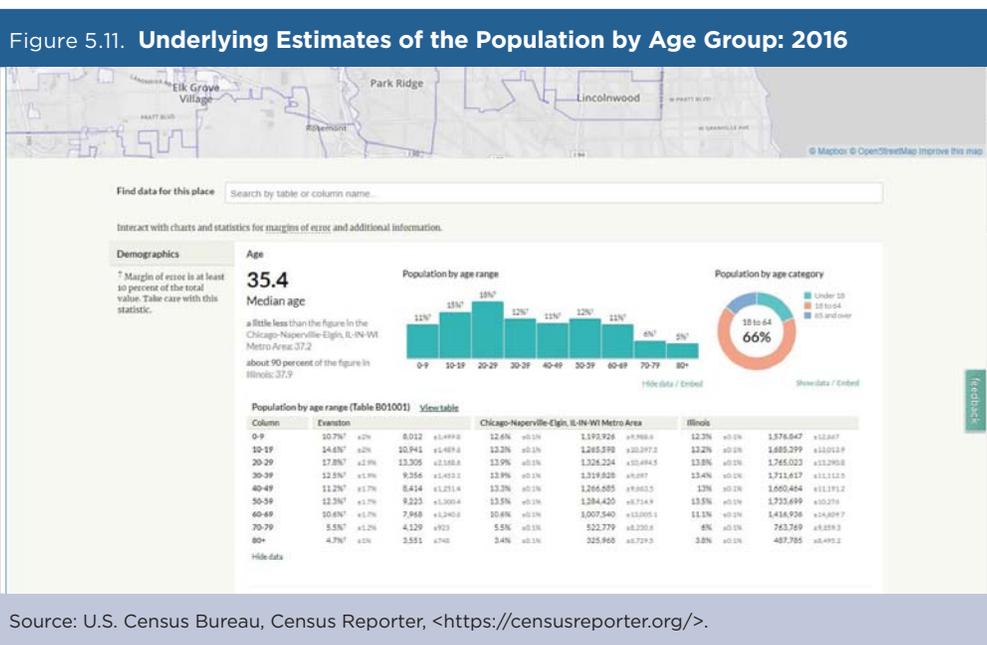
**Author:** Joe Germuska, Census Reporter Project Lead, Knight Lab at Northwestern University

While open government data have recently become trendy, the U.S. Census Bureau has been releasing troves of American Community Survey (ACS) data for years—through raw data files as well as pretabulated tables available through the data.census.gov Web site. The Census Bureau’s data.census.gov site is pure gold for journalists, but for novice users, it can be hard to mine.

Our team at Northwestern University’s Knight Lab created Census Reporter to make it easier for journalists to write stories using ACS data.<sup>45</sup> With Census Reporter, we have the freedom to highlight certain data points and leave out others. We are applying the same principles that we have used to build news applications for a general public: we prefer clarity over completeness.

Although Census Reporter is not a substitute for all the detailed ACS data available on the Census Bureau’s Web site, it provides a good place to start for journalists who want to explore ACS data for a given topic or geographic area. Since it was launched in 2014, Census Reporter has been widely used by journalists looking for background information for their stories. For example, the site was recently used in stories by *U.S. News & World Report* (on the best states to live), *The Texas Observer* (on the effects of Hurricane Harvey on Port Arthur, TX), and *Marketplace* (on gentrification in a Los Angeles neighborhood called Highland Park).<sup>46</sup>

We chose to focus on the ACS—as opposed to other federal data sources—because it provides the best combination of recent and local data. Census Reporter presents the latest ACS estimates for the nation, states, and many substate areas, down to the block group level. Charts, maps, and other data visualizations provide a friendly interface for navigating these data (see Figure 5.10). Users can also click on a “Show data” link below each chart to get more information about the underlying estimates and margins of error (see Figure 5.11).



<sup>45</sup> Census Reporter, <<https://censusreporter.org/>>.

<sup>46</sup> Casey Leins, *U.S. News & World Report* (March 1, 2017), “New Hampshire Benefits From Neighbor as a Leading State,” <[www.usnews.com/news/best-states/articles/2017-03-01/new-hampshire-benefits-from-neighbor-as-a-leading-state](http://www.usnews.com/news/best-states/articles/2017-03-01/new-hampshire-benefits-from-neighbor-as-a-leading-state)>; Michael Barajas, *The Texas Observer* (March 28, 2018), “Without State Aid, Advocates Worry Port Arthur Will Bleed Residents Long After Harvey,” <[www.texasobserver.org/port-of-no-return/](http://www.texasobserver.org/port-of-no-return/)>; *Marketplace*, Wealth and Poverty Desk, “York & Fig: At the Intersection of Change,” <<https://features.marketplace.org/yorkandfig/>>.

---

## **Data Behind the Scenes**

While Census Reporter is easy for journalists to use, it is also a good case study in how more advanced users can adapt ACS data to other purposes. The growing availability of open-source software has created new opportunities for journalists with programming skills to analyze and filter large data sets like the ACS to create new applications.

One of the most important goals behind the scenes at Census Reporter is making a flexible system to fetch ACS data so that the team at Census Reporter can quickly refresh their database and Web site. The amount of data that comes directly from the Census Bureau is huge and unwieldy for real-time queries, so our first experiment was to try loading these data into a set of “PostgreSQL” databases. PostgreSQL is a language that programmers can use to store, manipulate, and retrieve data.<sup>47</sup>

We based our database schema on Lee Hachadoorian’s census-postgres project, giving us a good place to start.<sup>48</sup> His scripts created a flexible schema to load and query all ACS tables for the 2006–2010 and 2007–2011 ACS 5-year data sets.

Since we were interested in comparisons over longer periods of time, we “forked” his project (created a copy of the code repository where we could make modifications without changing the original code) and added scripts to load data from additional ACS releases.

To make the process of loading these data more repeatable and reliable, we wrote several scripts that are meant to be run on a server. After the scripts finish running, you end up with a PostgreSQL database with complete, nationwide ACS data split across several thousand tables queryable with SQL. Combining these data with the Census Bureau’s TIGER geographic data, we can make geographic, topical, and temporal queries against the ACS data.

### **Building the ACS Database**

The data for Census Reporter are from the ACS Summary Files—a set of comma-delimited text files that contain all of the Detailed Tables for the ACS data releases—and were retrieved from the Census Bureau’s File Transfer Protocol (FTP) site.<sup>49</sup> The FTP client allows users to download large numbers of files or entire folders containing the Summary File data for each ACS release. Beginning with the 2011 ACS, the Census Bureau has made it easier to download the entire Summary File for an ACS release in two “TAR” files (Tape Archive files used in UNIX-based operating systems). Users should note that the TAR files are large and, depending on system constraints, may take some time to download.

Each data product (for example, 2012–2016 ACS 5-year data files) can be processed as one large file, but the data are horizontally partitioned by state and are vertically separated into “sequences” (chunks of data spanning 256 columns or fewer). A detailed explanation of the table structure and sequences can be found in the Summary File Documentation provided with each ACS data release.<sup>50</sup>

The collection of geography and sequence files makes for a large number of tables that have to be bulk loaded if a user wants to work with multiple files or multiple geographic areas. Import routines assume that all ACS data are separated into small geographic areas (tracts and block groups) and large geographic areas (all other areas), but file names are reused for both types of geographic areas. To distinguish between small and large geography files, the import routines assume that the two types of files are separated into directories named “All\_Geographies\_Not\_Tracts\_Block\_Groups” and “Tracts\_Block\_Groups\_Only.”

In each case, the parent directory name must match the name of the database schema where these data will be stored. For Census Reporter data work, we name the schemas after the data sets folder name on the Census Bureau FTP server (for example, “acs2016\_5yr”).

---

<sup>47</sup> PostgreSQL, <[www.postgresql.org](http://www.postgresql.org)>.

<sup>48</sup> GitHub, PostgreSQL schema and import scripts for recent U.S. Census Bureau data, <<https://github.com/leehach/census-postgres>>.

<sup>49</sup> U.S. Census Bureau, American Community Survey (ACS), Data via FTP, <[www.census.gov/programs-surveys/acs/data/data-via-ftp.html](http://www.census.gov/programs-surveys/acs/data/data-via-ftp.html)>.

<sup>50</sup> U.S. Census Bureau, American Community Survey (ACS), Summary File Documentation, <[www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html](http://www.census.gov/programs-surveys/acs/technical-documentation/summary-file-documentation.html)>.

The resulting database is used to create the profile and comparison pages at our Census Reporter site. Since we did all this hard work to get Census Reporter started, we wanted to share our methods with others. We made the scripts available to the public for anyone else who wants to replicate the process or modify it for their own work (see Figure 5.12). You can find the scripts on the GitHub Web site.<sup>51</sup>

**Figure 5.12. Census-Postgres Scripts on the GitHub Web Site**

**census-postgres scripts**

A set of scripts to make it easier to set up census-postgres on an Amazon EC2 instance.

**Use Existing Data**

If you just want to use American Community Survey data on your own PostgreSQL machine, it's recommended to use the dumps that Census Reporter provides. [Read more about those dumps in our Tumblr post.](#)

Use the next section only if you want to go through the process of rebuilding these data dumps from scratch.

**From Scratch**

These are the steps I follow when I want to start from scratch and load all ACS releases into the database.

1. Launch a `c1.xlarge` instance using the most recent Ubuntu 14.04 image, making sure to connect all four of the ephemeral storage to block devices during the setup walkthrough. If you have the `aws` command line tool installed and configured, this command should do it:

```
aws ec2 request-spot-instances --dry-run \
--spot-price 1.5 \
--instance-count 1 \
--launch-specification '{\
  "InstanceType": "c1.xlarge",\
  "ImageId": "ami-xxxxxx",\
  "BlockDeviceMappings": [\
    {\
      "VirtualName": "ephemeral0", "DeviceName": "/dev/sdb",\
      "VirtualName": "ephemeral1", "DeviceName": "/dev/sdc",\
      "VirtualName": "ephemeral2", "DeviceName": "/dev/sdd",\
      "VirtualName": "ephemeral3", "DeviceName": "/dev/sde"\
    }\
  ]\
}'
```

2. Connect to it and immediately launch `screen`

**Set up disk**  
([from this link](#))

Source: GitHub, Scripts used to set up census-postgres on an Amazon EC2 instance, <<https://github.com/censusreporter/census-postgres-scripts>>.

<sup>51</sup> GitHub, Scripts used to set up census-postgres on an Amazon EC2 instance, <<https://github.com/censusreporter/census-postgres-scripts>>.