Visualizing Income Distribution in the United States

Humberto Barreto and Sang T. Truong

Department of Economics and Management
DePauw University, Greencastle, IN 46135
Working Papers 2020-3

September 4, 2020

Abstract: Visit research.depauw.edu/econ/incomevis to see a novel, eye-catching visual display of the income distribution in the United States that conveys fundamental information about the evolution and current level of income inequality to a wide audience. We use IPUMS CPS data to create household income deciles adjusted for price level and household size for each of the 50 states and the District of Columbia from 1976 to 2018. We adjust for state price differences from 2008 to 2018. Plotting these data gives a 3D chart that provides a startling picture of income differences within and across states over time. Those interested in further customization can use our Python visualization toolbox, *incomevis*, available at pypi.org/project/incomevis.

JEL Codes: A10, A20, J10.

Keywords: visualization, household income, inequality, microdata, big data, data-mining, Python.

1 Introduction

We create a 3D chart to convey information about the distribution of real household income for a general audience. Income inequality is most often communicated by focusing on a single extreme, such as top 10% or 1% shares, or by a statistic which purports to capture some kind of overall measure of inequality, for example, the Gini coefficient or Theil index. The former approach leaves out a great deal of information and the latter requires expert-level knowledge, with no clear best option.

This short note shows the original application of the 3D display to international income data, applies the same visualization to real household income in the United States (U.S.), and explains the effect of sampling variability in interpreting the display over time.

2 The Sutcliffe Graph

(Sutcliffe, 2001) tackled the problem of conveying information about income inequality to a wide, non-expert audience with a data visualization (reproduced as Figure 1) that is eye catching, easily understood, and thought provoking. "This striking image shows the world distribution of income as a kind of city. ... Most prominently in the far corner we see the skyscraper ... which represents the income of the richest tenth of the U.S. population" (Sutcliffe, 2001, p. 17). The width of the city blocks represents population and there are ten blocks for each country, one for each income decile.

Sutcliffe's graph has been updated and enhanced by the CoreEcon project (The CORE Team, 2020). Figure 2 shows how color improves the display and conveys even more information. The side-by-side comparison of the two charts makes it easy to see China's rise since it keeps its red color, indicative of a poor country in 1980, but by 2014 has moved right, past the middle of the chart. The colors help explain how the distribution has changed from 1980 to 2014. For more years and interactive controls, see (Blundell, 2014).

3 A Sutcliffe Graph for Real Household Income by States

We create a 3D chart with block widths reflecting population to display data on household income by state in the U.S. from 1976 to 2018, as shown in Figure 3. As it does for world income distribution, our version of the Sutcliffe graph captures attention and delivers a great deal of information to a wide, non-expert audience in a single chart. We also take advantage of color to give viewers an overall sense of how different states have changed over time. We set the color scheme for all years based on median real household income in 1976 to make changes in the rankings easy to see.

Some users may want to engage with these data at a deeper level so we published our visualization as a Python package, *incomevis*, freely and openly available at pypi.org/project/incomevis. This ensures complete transparency, enables replication, and allows for customization of the visualization. Color schemes can be changed and the perspective on the chart can be adjusted. Animation is fully available (with playback controls) at rb.gy/yiaexf.

The documentation of *incomevis* describes, in detailed steps, exactly how to acquire and transform the raw household income data (HHINCOME at cps.ipums.org/cps-action/variables/HHINCOME) into percentiles and output it for rendering as a 3D chart, as shown in Figure 3. In brief, we use IPUMS CPS data (Flood et al., 2020) to estimate real household income (2018$) (RHH) and real household income per (equivalized) person (ERHH) by state and year. Raw, CPS ASEC household income is adjusted for inflation over time with the Consumer Price Index and the square root equivalence scale (Johnson et al., 2005, p. 13) is used to adjust real household income for household size. Note that sample years 1977 to 2019 give household income for the previous year. Household sampling weights are incorporated.
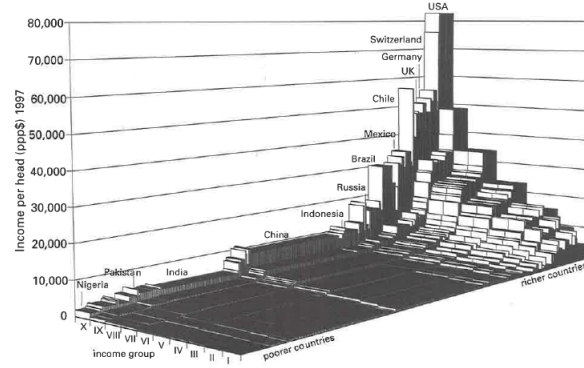
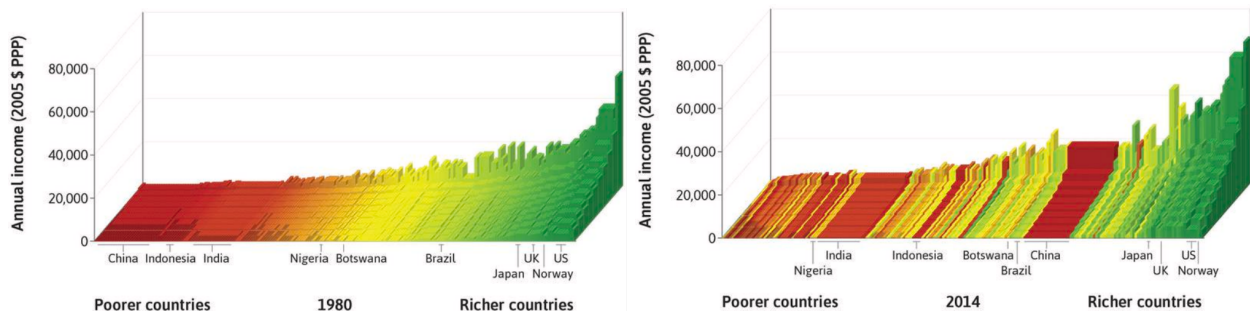Figure 1: World Distribution 1: The unequal city (Sutcliffe, 2001).

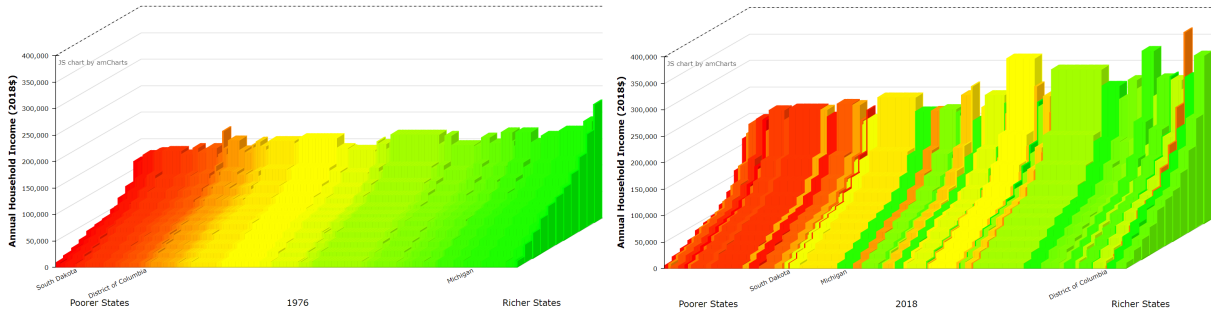Figure 2: World Income Distribution in 1980 (left) and 2014 (right) (The CORE Team, 2020).

Figure 3: U.S. Household Income Distribution in 1976 (left) and 2018 (right).
Source: research.depauw.edu/econ/incomevis

After computing real household income, the charts in Figure 3 are made by ordering the states from lowest (left) to highest (right) median real household income. Percentiles from 5th to 95th by 10, along with the median, are plotted from front to back. The display width of each state is determined by the number of households in that state in that year.

The Sutcliffe graphs for real household income in the U.S. in 1976 versus 2018 dramatically show the changing landscape of income distribution in the U.S. Figure 3 communicates the following facts about income distribution in the U.S:

1. The distribution of household income has become more unequal over time. In 1976, the chart was much more like a flat plane and each year's strip was roughly linear (until the back wall). In 2018, the curvature from front (5th percentile) to back (95th percentile) is much more pronounced and the

2

blocks at the back are much taller. This increase in curvature means the gap between the lowest and highest household incomes has increased.

2. Household income has risen over time, but it has not been even across the distribution. The front blocks (poorest) have stayed short and it is easy to see that the back blocks (richest) have grown a great deal. In 1976, the blocks in the back had a height of about $150,000. In 2018, they are much taller, with the 95th percentile in the richer states over $250,000 and the District of Columbia (DC), the thin slice with the tallest block, over $350,000.

3. There is a lot of variation in household income across states. By hovering the mouse over the online version graph, we can see that, in 2018, Mississippi had the lowest median household income at $42,763 (using RHH, thus, without adjusting for household size or state price differences). The highest ranked state was Massachusetts, with a median household income of $86,568 that is more than double that of Mississippi in 2018. The user can also choose a percentile and then move horizontally, from left to right, to see how the household income for that percentile rises from the poorest to the richest state.

4. There has been movement since 1976 in the ranking of states by household income, but mobility is limited. The colors are based on each state's median household income rank in 1976. Since the colors are mixed in 2018, this means some states have risen while others have fallen. The District of Columbia has moved up (to the right) a great deal, while Michigan's ranking has gone down. Will South Dakota continue to rise? Several states, however, seem stuck at the bottom.

The users can access the interactive version at research.depauw.edu/econ/incomevis and hover their mouse over different parts of the chart to get a popup with information. The down arrow in the top-right corner of the screen allows the users to download the data for the chart. They can apply individual state labels; choose the year; and display either real household income (RHH) or real household income per (equivalized) person (ERHH). The latter can be used to demonstrate the same four points as above, but controls for household size so it gives a better measure of household income per person. The online version of the chart invites comparison and generates questions.

Many sources and statistics provide evidence that income inequality has risen in the U.S. in the last few decades. However, communicating this to a non-expert audience is difficult. A Sutcliffe graph offers a way to bridge this gap. It attracts attention while displaying a great deal of information. The minimal processing (household incomes are adjusted for inflation and then sorted) is another advantage in that there is no analysis to explain. It is simply a display of sorted raw data. Our interactive version allows for customizing the graph for a class use or exploration.

4   Sampling Variability

Our Sutcliffe graph is based on data from the CPS ASEC, which produces a sample estimate of the true, population income distribution in the U.S. Unfortunately, this adds a layer of complication to the chart that we need to explain: sampling variability.

The height of each block in Figure 3 is a sample estimate of the RHHINCOME percentile for a state and, therefore, it has a standard error (SE). To get a handle on the margin of error in the height of each block depicted in the charts, we found bootstrapped SEs for the District of Columbia, with the smallest sample size of 320, and California (CA), the largest n = 5,412, in the 1977 CPS sample. Full documentation is available in the Excel workbook, 1976RHHINCOME.xlsm, at research.depauw.edu/econ/incomevis.

Table 1 shows that the dollar value of the standard error of the percentile estimate rises as the percentile rises and falls as the sample size rises. This means that there is relatively little sampling variability in the front of the Sutcliffe graphs in Figure 3, but there is quite a bit of bounce in the tall blocks in the back. The highlighted cells in Table 1 show the standard deviation of 10,000 resampled 95th percentiles. For CA (and other highly populated states with wide blocks), the margin of error is several thousand dollars for the 95th percentile, but for states with thin slices (e.g., DC and SD), the bounce is in the tens of thousands of dollars. In other words, if the Bureau of Labor Statistics had carried out a second survey in 1977, we would see the 95th percentile of real household income in DC vary by plus or minus roughly $13,500.

n = 320 for DC in 1977 CPS					
percentile	5	25	50	75	95
RHHINCOME ($)	8,600	29,762	49,861	83,379	172,087
bootstrapped SE ($)	1,322	2,859	2,392	3,807	13,487

n = 320 for CA in 1977 CPS					
percentile	5	25	50	75	95
RHHINCOME ($)	12,161	30,887	57,451	92,212	166,192
bootstrapped SE ($)	338	586	812	927	3,080

Table 1: Bootstrapping to Estimate Margin of Error (in 2018$).
Source: 1976RHHINCOME.xlsm available at research.depauw.edu/econ/incomevis

Creating the Sutcliffe graph with CPS data forces us to confront the fact that there is sampling variability in the chart. We should interpret the heights of the blocks as estimates and understand that the heights are subject to sampling fluctuation that increases as we go from front to back. A sharp-eyed viewer may notice that the back wall in 2018 (Figure 3) is not only taller, but more jagged than it was in 1976. We must proceed with caution in responding to this observation. This cannot be naively interpreted as if the back wall was based on the true, population 95th percentile—sampling variability must be part of the discussion.
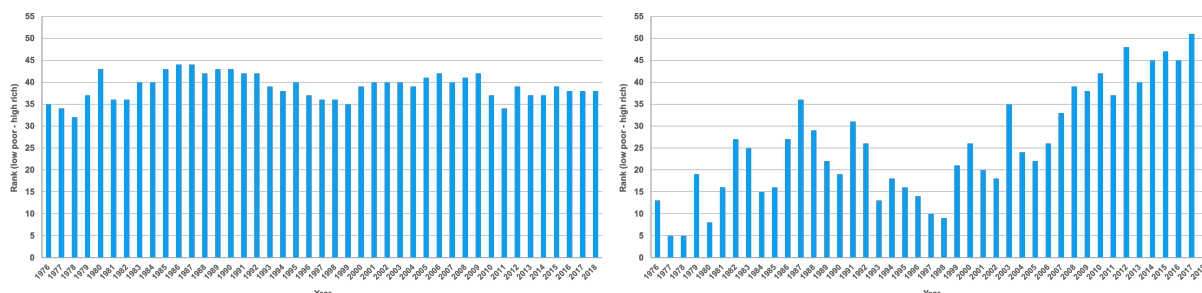

Figure 4: Ranking of CA (left) and DC (right) based on Median RHHINCOME over time.
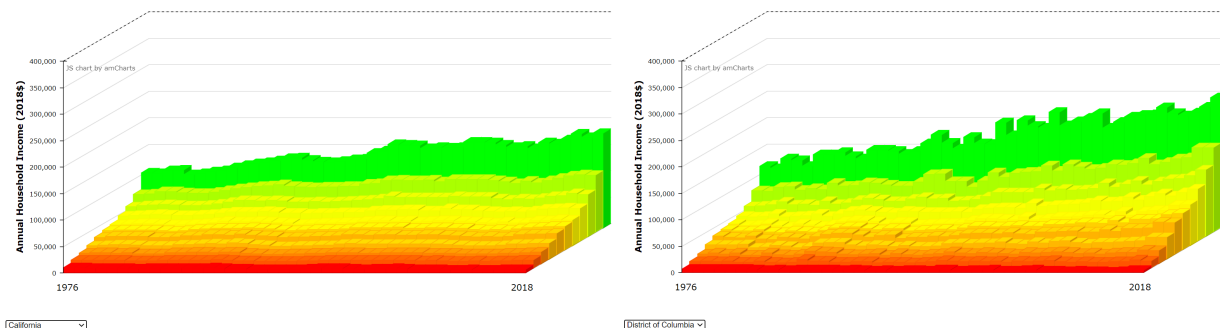

Figure 5: U.S. Household Income Distribution in CA (left) and DC (right).
Source: research.depauw.edu/econ/incomevis

One simple way to show the effect of sampling variability is through animation. The website at re-

4

search.depauw.edu/econ/incomevis has three videos that show how the graph changes over time. There is so much movement in the rankings (and colors) for one year to the next that it cannot be attributed to real factors. It is true, for example, that DC's median real household income has increased from 1976 to 2018 relative to other states, but the bouncing around from one year to the next is due to sampling variability in the estimate (Figure 4).

Another way to understand this point is by looking at a single state over time. Select the *Individual State* option at research.depauw.edu/econ/incomevis to get a state's income distribution from 1976 to 2018. The width of the blocks is the same (no population information is included). Figure 5 compares the distributions of CA and DC.

It is easy to see that most of the fluctuations in DC's back wall are too large to be explained by real economic forces—these ups and downs reflect the effects of random sampling. CA's back wall is smoother than DC's because its sample size is much bigger, 5,412 versus 320 observations.

Individual state charts reinforce the first two of the four points made above and they enable us to ask a new question, related to the first point: exactly how are the blocks rising from front to back? Is the increase linear, convex, or concave? California's 2018 distribution (farthest right in Figure 5 shows the blocks rising roughly linearly (or maybe even concave in the middle percentiles) and then taking off, rising faster and faster. It is obvious for CA and DC that the curvature is much more pronounced in 2018 than it was in 1976 and this provides direct evidence of rising household income inequality. Users can download the data to compute the exact change in heights for different years and different states.

5   Controlling for State Price Differences

We routinely adjust for inflation to make a better comparison of household income over time. We can do the same for geographic differences in prices. It is true, for example, that household income is much higher in California than Alabama (AL), but prices are also much higher in CA than AL. The Bureau of Economic Analysis offers regional price parities (RPP) to enable "comparisons of buying power across the 50 states and the District of Columbia" (Bureau of Economic Analysis, 2020), but only since 2008. In 2018, for example, AL's state price level for all items is 86.4. CA was a third higher at 115.4. Thus, adjusting for price differences across states matters. We know of no state price level data stretching back to 1976.

Our interactive website provides an option for selecting RHHRPP and ERHHRPP, real household income adjusted for state prices and real household income per (equivalized) person adjusted for state prices, respectively, from 2008 to 2018. Figure 6 shows real household income adjusted for state prices in 2018 in the bottom left panel. The top left panel reproduces Figure 3 (2018) for convenience in comparison.

Although it cannot be used to convey the dramatic change in income inequality over decades because state price level data are not available, controlling for price differences across states does provide a better snapshot of the variation in real household income in the U.S. South Dakota is much farther right in the bottom than the top graph because their households face prices about 12% lower than nationally (SD's 2018 RPP is 87.9). Since lower household income states typically have lower price levels, their adjusted incomes rise and richer states see their adjusted incomes fall so that the entire graph is more level in the bottom than top left—the effect is especially noticeable along the back wall.

The right panels in Figure 6 show real household incomes per (equivalized) person. The District of Columbia moves to the top because it has more small households of one or two persons than other states. The bottom right graph, real household income per (equivalized) person adjusted for state price differences (ERHHRPP), best speaks to the question of income inequality per person because it controls for both
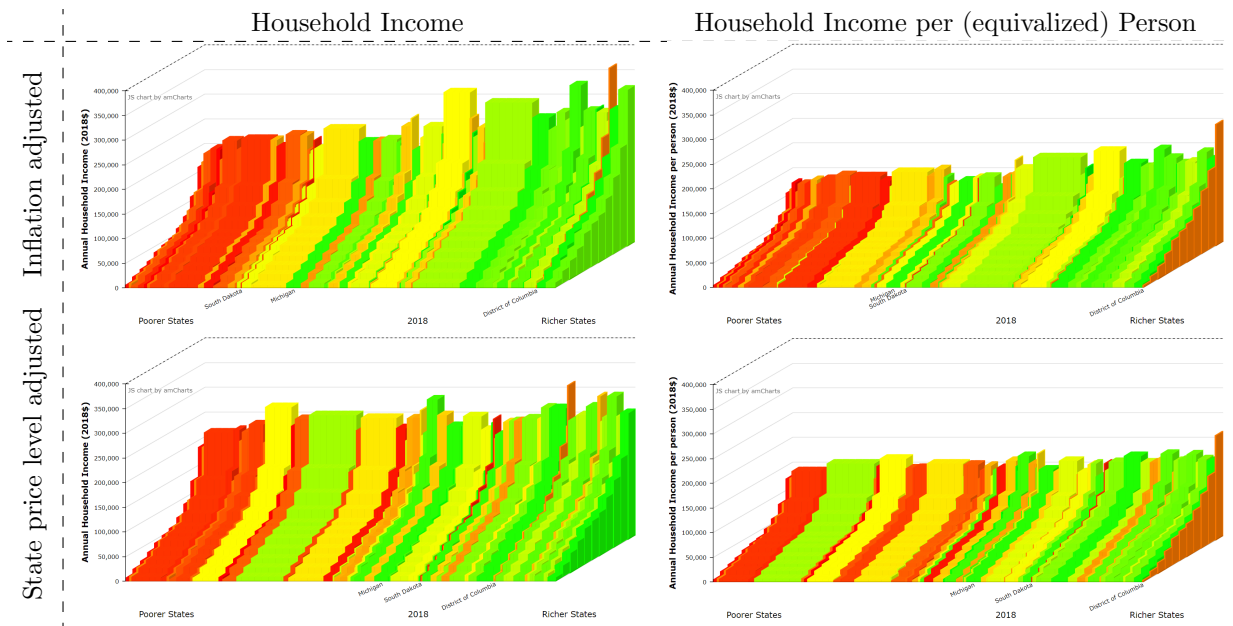
Figure 6: Four Snapshots of Household Income in 2018. Source: research.depauw.edu/econ/incomevis

inflation over time and differences in state price levels.

6 Limitations and Further Research

Comparing household income over time is not without its problems. The CPS ASEC survey is not exactly the same every year. Questions change and so do collection methods (e.g. see a discussion of major changes in 2017 at (Rothbaum & Edwards, 2019)). Another problem is that adjusting with CPI is best thought of as an approximate attempt to control for price changes over time. The index suffers from a series of biases (new good, quality, outlet, and substitution) that cause it to overstate inflation. Finally, RPP are only available back to 2008.

Our data visualization can be extended by incorporating other variables that affect income distribution over time in the U.S. Perhaps the most important is race. When surveyed about perceived income differences by race and ethnicity, "The evidence indicates that the magnitude of these misperceptions is substantial, with respondents estimating that for every \$100 in wealth held by a White family, a Black family has \$90, when, in reality, that Black family has \$10; misperceptions about the Latinx–White wealth gap are just as large" (Kraus et al., 2019, p. 917). We had hoped to create Sutcliffe graphs of racial and ethnic subgroups (and by education), but the CPS ASEC sample size of roughly 60,000 households is too small to generate reliable percentiles for each state broken down by race and ethnicity.

Income (and wealth) distribution is a complex issue. This short note is not meant to perform sophisticated analyses. We do not control for changes in the age distribution or explore higher order concepts like convergence among states. Our contribution is to provide an entrée to the study of income distribution. A Sutcliffe graph of household income in the U.S. over time offers a good place to start because it lays bare the facts, captures attention, and stimulates many questions about causes and remedies.

References

Blundell, J. (2014). *Visualisation of the global income distribution since 1980.* Retrieved 2020-08-20, from

https://jackblun.github.io/Globalinc/

Bureau of Economic Analysis. (2020). *Regional price parities by state and metro area.* Retrieved 2020-08-20, from https://www.bea.gov/data/prices-inflation/regional-price-parities-state-and-metro-area

Flood, S., King, M., Rodgers, R., Ruggles, S., & Warren, J. R. (2020). *Integrated public use microdata series, current population survey: Version 6.0 [dataset].* Retrieved 2020-08-20, from doi.org/10.18128/D030.V6.0

Johnson, D., Smeeding, T., & Torrey, B. (2005). Economic inequality through the prisms of income and consumption. *Monthly Labor Review*, 11–24. Retrieved 2020-08-20, from www.bls.gov/opub/mlr/2005/04/art2full.pdf

Kraus, M., Onyeador, I., Daumeyer, N., Rucker, J., & Richeson, J. (2019). The misperception of racial economic inequality. *Perspectives on Psychological Science*, 899–921. doi: doi.org/10.1177/1745691619863049

Rothbaum, J., & Edwards, A. (2019). *Survey redesigns make comparisons to years before 2017 difficult.* Retrieved 2020-08-20, from https://www.census.gov/library/stories/2019/09/us-median-household-income-not-significantly-different-from-2017.html

Sutcliffe, B. (2001). *100 ways of seeing an unequal world.* London: Zed Books.

The CORE Team. (2020). *The economy.* Retrieved 2020-08-20, from https://core-econ.org/the-economy/book/text/19.html