Where are Things Changing?

Using data from the American Community Survey for Historical Analyses

By Doug Hillmer

Independent consultant

Abstract:

Currently, there are 14 years of 1-year data releases covering the entire U.S. population from the American Community Survey (ACS), 2006-2019. This rich body of data can be used for many types of historical analyses. But, there are questions that arise when attempting these analyses, including: which data product(s) from the annual releases should be used; which type of geography is best for the analysis; how do we detect "change" over time? This paper documents the author's experiences when attempting a data analysis of educational attainment trends at the county level using all 14 years of a single ACS table. The paper includes ideas and suggestions that may be useful for anyone planning to do their own historical analysis of a topic using the ACS data. Several details along with references to other relevant work are included in appendices to this paper.

1. Introduction and background

Recently, I was doing some data analysis to help a political candidate running for a House seat in New York state. I came across some estimates on educational attainment that surprised me. In this congressional district, a fairly high proportion of white men were not completing high school, and, in contrast, white women were improving on that measure. I had already been interested in using the ACS data to look for trends over time for some characteristic. So, I decided to do this for a few basic measures of educational attainment.

I want to be sure to note here that this paper, as the title suggests, is about both geography and time: "where" indicates a level (or multiple levels) of geography to focus on; "changing" suggests time; i.e., years in the 2006-2019 period. This paper makes no attempt to explain why things are changing. That is a task better left to seasoned educational data analysts and beyond the scope of this paper.

Also, with regard to the word "changing", it is clear that one cannot demonstrate that change has really occurred for some area just by analyzing one sample-based dataset. Only by using other relevant data and information can such a change be verified.

2. How the ACS data was used

I decided to focus on two simple measures of educational attainment representing the two "extremes" of educational attainment in adults: no high school completion (including GED); and, achieving a Bachelors degree or higher. Also, because the ACS is well known for publishing results annually for many sub-state geographic areas, I wanted to explore one type of sub-state geography. I decide to use only the counties that met the 65,000 population threshold; i.e., those counties for which the ACS would publish (some) results each year. This decision is examined in more detail in section 3 below. I also decided to the two measures I had chosen by gender for the entire population and for three sub-populations as well: white alone, black alone, and Hispanic origin. Adding the race/Hispanic origin and gender dimensions makes these results potentially more interesting, but it also means fewer sample cases supporting these measures – i.e., results that are statistically less reliable.

Then, there was the “time” decision. I had 14 years of annual 1-year ACS releases available. For the sake of simplicity, I decided to use the entire 2006-2019 period. Since I had no prior knowledge of any government programs or other reason to focus on a sub-period, I used all 14 years of data. There may be objections to this decision. One objection that comes to mind is that using all 14 years may cause me to miss trends that begin later in the time period. This is because the years before the trend begins may not show any trend, but those years are still included in the 14-year period. Although I did not look at any sub-periods, there are some simple ways to investigate sub-periods for trends that emerge during those periods. However, one question that arises (discussed in more detail below) is how many years (data points) should we require to do some kind of “trend analysis”? And, to what extent is that number of years dependent on the topic being studied?

3. WHERE – deciding on the sub-state geographic type to use

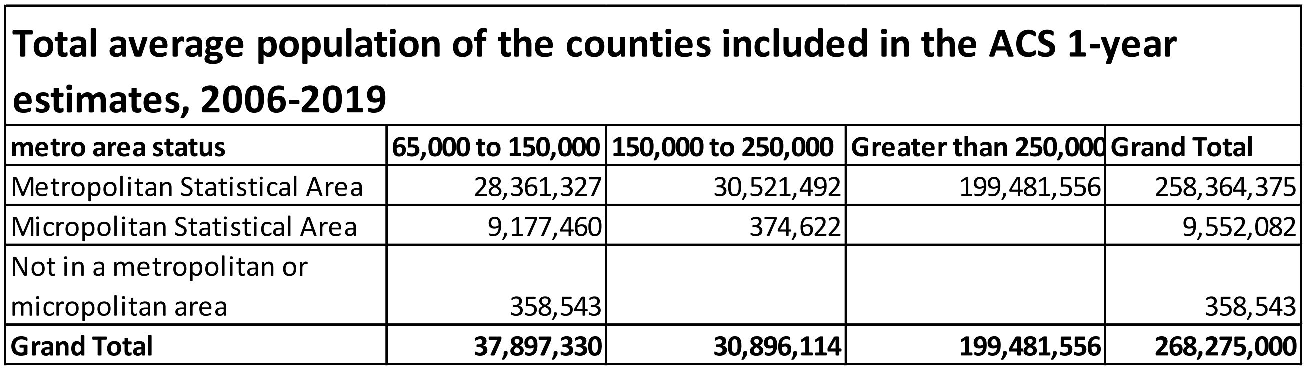
3.1. Why counties?

Counties are political jurisdictions, not subject to change because of a Census. There are many other possible sub-state geographic area types to choose from: cities; school districts; PUMAs; etc. However, all statistical geographic areas (those dependent on Census results for their boundary definitions) are subject to changes right in the middle of the 2006-2019 period. School districts are complex to use across all states because of the three types of school districts (elementary, secondary, and unified) and because some areas have one type but not the other types. While cities are another political jurisdiction, their boundaries are much more likely to change than county boundaries. In fact, there was only one change affecting counties eligible for 1-year ACS estimates in the 2006-2019 period: Bedford City Virginia which had been an independent city was absorbed into Bedford County, Virginia adding about 10% to the population of Bedford County. The biggest change to county boundaries in the last 20 years was the creation of Broomfield County, Colorado causing the boundaries of three other counties to change. However, this change occurred in 2001. Detailed descriptions of county boundary changes covering the 2006-2019 period can be found at the “county changes“ link under

https://www.census.gov/programs-surveys/geography/technical-documentation/boundary-change-notes.html .

Table 1 shows the sum of the total population for each county with a population of 65,000 or more and therefore eligible for 1-year ACS estimates. The total population for each county is calculated as an average over the 14-year period. The grand total represents about 80% of the 2020 Census count of the U.S. population.

Table 1 total population in counties eligible for 1-year ACS estimates by population size and metropolitan area status



Total average population of the counties included in the ACS 1-year estimates, 2006-2019

metro area status	**65,000 to 150,000**	**150,000 to 250,000**	**Greater than 250,000**	**Grand Total**
Metropolitan Statistical Area	28,361,327	30,521,492	199,481,556	258,364,375
Micropolitan Statistical Area	9,177,460	374,622		9,552,082
Not in a metropolitan or micropolitan area	358,543			358,543
Grand Total	**37,897,330**	**30,896,114**	**199,481,556**	**268,275,000**

3.2. Can PUMS data be useful?

There are two ways to use PUMS data from the 1-year ACS: use a pre-aggregated table for the PUMA geography level; use the 1-year PUMS data. There are definite advantages to using the estimates from the full ACS sample at the PUMA level. Since all PUMAs are defined to be approximately 100,000 total population in size, there are estimates published for each PUMA from the 1-year ACS. However, when population subgroups are being analyzed, many of these estimates may be suppressed because there is simply not enough sample to support a reliable estimate. For example, the table I used to generate estimates for the Black alone population did, indeed, suffer losses due to data suppression at the county level. However, the loss was even worse at the PUMA level (approximately 30% of the PUMAs).

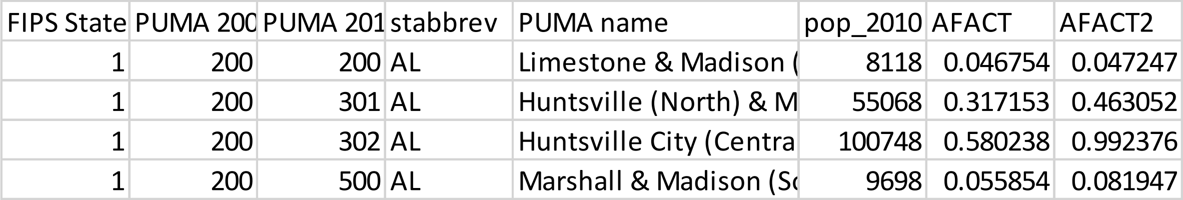
That leads to the question can we use the PUMS microdata itself for each year to detect trends? There is one major advantage to using the PUMS data. You are guaranteed to get a result of some kind for each PUMA; i.e., you have “wall-to-wall” geographic coverage. However, the statistical reliability of your results will be less than results from the full sample since the PUMS sample is a significant reduction in sample size.

Perhaps the biggest hurdle to overcome when using PUMS data or data from the full sample at the PUMA level is the fact that PUMAs are re-defined 2 years after a census. In general, this results in a major change to PUMA boundaries. To see how much change is actually occurring, you can use a software application known as GEOCORR from the Missouri Census Data Center (https://mcdc.missouri.edu/applications/geocorr.html). GEOCORR allows the user to view the amount of overlap between two different types of geography. One way to use GEOCORR is to compare the 2000-based PUMA definitions with the 2010-based definitions. This comparison can be done in both directions. As it says in the GEOCORR documentation,

“*The standard output correlation list from Geocorr has a single AFACT allocation factor variable, which indicates the decimal portion of the source geocodes contained within the target geocodes. It may also be useful to know how this works going in the other direction, i.e., to know what portion of the target area (the complete target area, not just the part within the source area) is contained in the source geocodes (AFACT2).*”

Here is an example of the GEOCORR output for one 2000-based PUMA in Alabama.

Figure 1 an example of the GEOCORR output from one 2000-based PUMA in Alabama



FIPS State	PUMA 200	PUMA 201	stabbrev	PUMA name	pop_2010	AFACT	AFACT2
1	200	200	AL	Limestone & Madison (	8118	0.046754	0.047247
1	200	301	AL	Huntsville (North) & M	55068	0.317153	0.463052
1	200	302	AL	Huntsville City (Centra	100748	0.580238	0.992376
1	200	500	AL	Marshall & Madison (S	9698	0.055854	0.081947

Note that four 2010-based PUMAs are needed to cover the 2000-based PUMA. The AFACT column gives the proportion of the 2000-based PUMA that is covered by the 2010-based PUMA in that row. The AFACT2 column does the same but in the opposite direction: 2010-based PUMA coverage by 2000-based PUMA. In Fig. 1, the numbers in the AFACT column add up to 1 (with slight rounding error), but the numbers in the AFACT2 column are not additive since we do not see all of the 2000-based PUMAs contributing to each of the 2010-based PUMAs.

4. How I investigated change over time for educational attainment at the county level

I chose educational attainment because there is already a great deal known about changes in educational attainment over the past years and decades. As mentioned above, I wanted to analyze results for two basic measures, both as proportions: no high school completion; and, Bachelors degree or higher. A great deal of analysis had already been done at the National Center on Educational Statistics (NCES). (For example, see this NCES table on high school completion over a 100-year period - https://nces.ed.gov/programs/digest/d18/tables/dt18_104.10.asp)

However, I did not find data over time at the county level on the educational attainment measures I chose to study. There is an ACS table that supports this analysis and did not change during the 2006-2019 period. (It is worth noting here that the ACS website contains very useful documentation on the structure and universe of each table for all years since the beginning of ACS in full production. See this link for more detailed information: https://www.census.gov/programs-surveys/acs/technical-documentation/table-shells.html)The table is B15002 and its race/Hispanic origin iterations, B15002A-I. However, this table is more detailed than what I needed. Fortunately, a "collapsed" version (with less detail) of the table is also available: C15002 and C15002A-I. Table 2 shows the template for C15002 and for one of the iterations, C15002A. The 11 cells of the iterated template (C15002A shown here) contain sufficient detail to enable the calculation of both measures by gender and race/Hispanic origin groups. And, the full table, C15002, supports these calculations for the entire population. In general, when considering alternative sources from the pre-aggregated data for your area of interest, it is best to choose the source with the least amount of detail since that source will be suppressed less often. Note that the universe for these tables is "Population 25 years and over". This restriction can affect how the results are interpreted, and it is discussed in section 8 below.

Table 2 templates for the detailed tables used to generate the measures of interest

C15002			**SEX BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER**
C15002			*Universe: Population 25 years and over*
C15002	1	C15002_001	Total:
C15002	2	C15002_002	Male:
C15002	3	C15002_003	Less than 9th grade
C15002	4	C15002_004	9th to 12th grade, no diploma
C15002	5	C15002_005	High school graduate (includes equivalency)
C15002	6	C15002_006	Some college, no degree
C15002	7	C15002_007	Associate's degree
C15002	8	C15002_008	Bachelor's degree
C15002	9	C15002_009	Graduate or professional degree
C15002	10	C15002_010	Female:
C15002	11	C15002_011	Less than 9th grade
C15002	12	C15002_012	9th to 12th grade, no diploma
C15002	13	C15002_013	High school graduate (includes equivalency)
C15002	14	C15002_014	Some college, no degree
C15002	15	C15002_015	Associate's degree
C15002	16	C15002_016	Bachelor's degree
C15002	17	C15002_017	Graduate or professional degree
C15002A			**SEX BY EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER (WHITE ALONE)**
C15002A			*Universe: White alone population 25 years and over*

C15002A	1	C15002A_001	Total:
C15002A	2	C15002A_002	Male:
C15002A	3	C15002A_003	Less than high school diploma
C15002A	4	C15002A_004	High school graduate (includes equivalency)
C15002A	5	C15002A_005	Some college or associate's degree
C15002A	6	C15002A_006	Bachelor's degree or higher
C15002A	7	C15002A_007	Female:
C15002A	8	C15002A_008	Less than high school diploma
C15002A	9	C15002A_009	High school graduate (includes equivalency)
C15002A	10	C15002A_010	Some college or associate's degree
C15002A	11	C15002A_011	Bachelor's degree or higher

Disaggregating the two measures by gender and the population subgroups (including total population) resulted in 16 separate measures. Then, for each measure by gender and population subgroup, I added a ratio measure of men to women. That brought the total number of measures to 24, shown here in Table 3.

Table 3 Variable names for measures and descriptions

variable name	description
bach_all_f_pct	total population: females with a Bachelors degree or higher as a proportion of all females 25 years or older
bach_all_m_pct	total population: males with a Bachelors degree or higher as a proportion of all males 25 years or older
bach_all_ratio	total population: ratio of men to women with bachelors degree or higher
bach_black_f_pct	Black alone population: females with a Bachelors degree or higher as a proportion of all females 25 years or older
bach_black_m_pct	Black alone population: males with a Bachelors degree or higher as a proportion of all males 25 years or older
bach_black_ratio	Black alone population: ratio of men to women with bachelors degree or higher
bach_hisp_f_pct	Hispanic population: females with a Bachelors degree or higher as a proportion of all females 25 years or older
bach_hisp_m_pct	Hispanic population: males with a Bachelors degree or higher as a proportion of all males 25 years or older
bach_hisp_ratio	Hispanic population: ratio of men to women with bachelors degree or higher
bach_white_f_pct	White alone population: females with a Bachelors degree or higher as a proportion of all females 25 years or older
bach_white_m_pct	White alone population: males with a Bachelors degree or higher as a proportion of all males 25 years or older
bach_white_ratio	White alone population: ratio of men to women with bachelors degree or higher
no_hs_all_f_pct	total population: females with no high school completion as a proportion of all females 25 years or older
no_hs_all_m_pct	total population: males with no high school completion as a proportion of all males 25 years or older
no_hs_all_ratio	total population: ratio of men to women no high school completion
no_hs_black_f_pct	Black alone population: females with no high school completion as a proportion of all females 25 years or older
no_hs_black_m_pct	Black alone population: males with no high school completion as a proportion of all males 25 years or older
no_hs_black_ratio	Black alone population: ratio of men to women no high school completion
no_hs_hisp_f_pct	Hispanic population: females with no high school completion as a proportion of all females 25 years or older
no_hs_hisp_m_pct	Hispanic population: males with no high school completion as a proportion of all males 25 years or older
no_hs_hisp_ratio	Hispanic population: ratio of men to women no high school completion
no_hs_white_f_pct	White alone population: females with no high school completion as a proportion of all females 25 years or older
no_hs_white_m_pct	White alone population: males with no high school completion as a proportion of all males 25 years or older
no_hs_white_ratio	White alone population: ratio of men to women no high school completion

5. CHANGING

Analyzing change over time presents two important questions: what do we mean by "change", and what kind of time period are we interested in studying? This section addresses both questions.

5.1. Time period: Why use 1-year data instead of 5-year data?

The biggest drawback to using the 1-year ACS data is the population threshold of 65,000 total population. In addition, there is another drawback: suppression of certain tables occurs in the 1-year products but not in the 5-year products – c.f., American Community Survey Office Data Suppression (census.gov). Depending on the population subgroup being examined, data suppression can have a minimal or very large impact on the analysis.

However, the obvious drawback to using the 5-year ACS estimates is that (as of today) there are only two non-overlapping 5-year periods: 2006-2010; and, 2011-2015. Of course, 2016-2020 will be added later in 2021. However, that is still only 3 data points. There is another serious drawback to using the 5-year data when a geographic area qualifies for 1-year estimates as well. The 5-year estimates have a "smoothing" effect that can make it more difficult or impossible to see trends that are actually taking place. For example, consider this fictional table of 10 annual results for the percent of people not completing high school:

Year	1-year estimates	5-year estimates
2006	0.25	0.23
2007	0.24	0.23
2008	0.23	0.23
2009	0.22	0.23
2010	0.21	0.23
2011	0.2	0.196
2012	0.27	0.196
2013	0.18	0.196
2014	0.18	0.196
2015	0.15	0.196

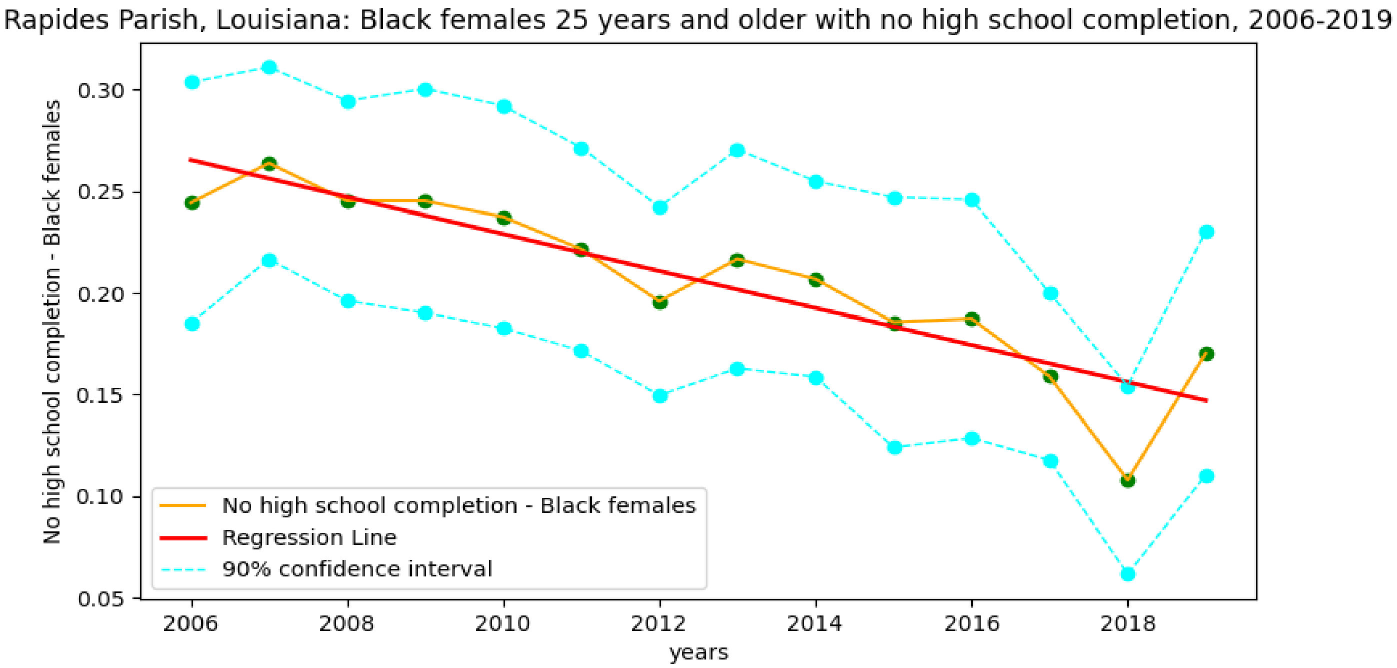
In the 10-year period there is only 1 year in which the proportion goes up (2012) from previous years. Thus, the full set of 10 data points probably satisfies almost any statistical definition of a "trend". However, the two data points from the 5-year periods mask out most of this trend

5.2. What do we mean by "Change"?

If we are looking for change that is more than change due to random sampling error, it is common to talk about a "trend". But, what is a "trend"? It seems that a tend implies change happening, more or less, in a consistent manner; i.e., either some measure going up or going down over time. Therefore, we would look for a monotonically increasing or decreasing curve. Of course, there are changes that are not a single monotonic curve, but a curve in one direction followed by another curve in a different direction. Such a combination of curves could be seen in economic analysis of a boom period followed by a recession. Or, if we are studying certain characteristics of an age cohort over time. For example, we might expect fertility to rise as an age cohort of young women go through normal child-bearing years and begin to decline as the cohort continues to grow older.

However, my analysis of two measures of educational attainment, proportion not completing high school and proportion with a Bachelor's degree or higher, is aimed at looking for two monotonic trend curves, one sloping downward and the other sloping upwards. Also, I am examining data over a 14-year period. So, even a gradual slope degree can result in a significant trend over that time period. Thus, it is possible that the year-to-year changes are not statistically significant at the 90% or 95% level of confidence, but there is still a trend underway. Here is one example (see the note in Appendix 1 on my use of regression in this paper).

Figure 2 Rapides Parish, LA: Change in proportion of females 25 years and older with no high school completion



The slope of the regression line in Fig 2 is about 0.006, and the r-squared value of the trendline is about 0.8. However, the Z score for the 2006 estimate vs the 2019 estimate is 1.44; i.e. not statistically significant at even the 90% level of confidence. In fact, none of the Z scores comparing two succeeding years is significant at the 90% level. Yet, it seems clear that a downward trend is occurring.

The main statistical test for a trend that I used is the Mann-Kendall (MK) nonparametric trend test. I set the p-value to 0.001. As a nonparametric test MK makes no assumptions about the underlying distribution of the data. That does not mean there is no evident distribution, but only that no assumption about that distribution is made. It may seem unusual (ill-advised?) to use a nonparametric measure when we are dealing with a “parametric” environment in which the underlying distribution may be known. Most references to MK applications are in studies related to the environment and engineering. The formula for the measure used in MK is quite simple:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \operatorname{sgn}(X_j - X_k)$$

with

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases}$$

Note that the S measure is not based on magnitude of change. So, a very small slope over time could have the same S value as another instance with a steeper slope. The test for a trend computes a Z statistic and then does a hypothesis test at a given level of significance (the p-value based on the normal distributions defined in statistics textbooks). See the references in Appendix 3 for more details.

Using the example for Rapides Parish, Louisiana shown in Fig. 2, here is what the S value results would look like as we go through each of the 14 years:

Table 4 Each of the 13 sums of the signs of the differences from the first year to 2019 starting with 2006

year	**Sum of sign value of the differences , first year to 2019**
2006	
2007	-7
2008	-12
2009	-11
2010	-10
2011	-9
2012	-8
2013	-3
2014	-6
2015	-5
2016	-2
2017	-3
2018	0
2019	1
S value	**-75**

And, indeed, the MK test finds the above example for Rapides Parish, Louisiana to be a trend.

Note that the maximum value (minimum value) for the S measure equates to $\sum_1^{n-1} k$ = n(n-1)/2. So, for n=14 years, we get 91 (-91) as the maximum (minimum) value for S.

6. An overview of the results

I only reviewed results for counties with 10 or more years of data; i.e., 10-14 years. There are 817 such counties. Data were published in all these counties for the White alone population. However, because of data suppression, results for the Black alone population were available in only 489 counties and for the Hispanic origin population only 513 counties. Table 3 below summarizes the results by trend for each measure.

Table 5 For each measure, number of counties by trend

Results for 24 measures of educational attainment for counties published from the 1-year ACS tables, 2006-2019

Notes: 1) The columns labelled "Avg S values" refer to the average of the Mann-Kendall S estimates for assessing a trend.

2) Only counties with 10 or more years of published results are included.

3) A p-value threshold of 0.001 was used in the Mann-Kendall test to test the hypothesis that a trend existed.

	decreasing		increasing		no trend		Total	
Measure	Counties	Avg S value	Counties	Avg S value	Counties	Avg S value	Counties	Avg S value
bach_all_f_pct			317	73.00	500	44.78	817	55.73
bach_all_m_pct			102	70.92	715	29.36	817	34.55
bach_all_ratio	40	-67.85			777	-27.81	817	-29.77
bach_black_f_pct			33	69.55	456	23.52	489	26.62
bach_black_m_pct			6	68.00	483	14.15	489	14.81
bach_black_ratio					489	-5.91	489	-5.91
bach_hisp_f_pct			39	70.95	474	20.80	513	24.61
bach_hisp_m_pct			19	67.63	494	17.68	513	19.53
bach_hisp_ratio	1	-63.00			512	-7.73	513	-7.84
bach_white_f_pct			286	71.72	531	42.82	817	52.94
bach_white_m_pct			75	69.88	742	27.62	817	31.50
bach_white_ratio	31	-68.94			786	-26.55	817	-28.16
no_hs_all_f_pct	212	-70.20			605	-41.53	817	-48.97
no_hs_all_m_pct	149	-69.70			668	-37.30	817	-43.21
no_hs_all_ratio					817	9.56	817	9.56
no_hs_black_f_pct	45	-70.96			444	-27.63	489	-31.62
no_hs_black_m_pct	27	-69.59			462	-23.97	489	-26.49
no_hs_black_ratio	1	-63.00			488	6.81	489	6.67
no_hs_hisp_f_pct	61	-71.72			452	-21.44	513	-27.42
no_hs_hisp_m_pct	47	-70.83			466	-22.31	513	-26.76
no_hs_hisp_ratio	3	-62.67			510	-7.42	513	-7.75
no_hs_white_f_pct	164	-69.66			653	-38.60	817	-44.83
no_hs_white_m_pct	94	-68.32	1	65.00	722	-34.38	817	-38.16
no_hs_white_ratio			1	65.00	816	11.46	817	11.53

For each of the 24 measures, the number of counties with no trend is larger than the counties with the increasing or decreasing trends. The two largest trend totals are for all women with a Bachelor's degree or higher and White alone women with a Bachelor's degree or higher. The ratio measures are clearly weaker than the proportion measures in showing a trend. In fact, two of the ratio measures (bach_black_ratio and no_hs_all_ratio) showed no trend for any of the counties.

Had I run the MK statistic with a lower threshold for the trend hypothesis test (i.e., with a p value larger than 0.001), there would be considerably more "with trend" results. In fact, I first did these trend tests with a p-value threshold of 0.01, resulting in almost 3 times as many trend cases.

Admittedly, adjusting the p-value as I have done for the MK test is quite subjective. The main advantage of using a lower p-value threshold is that we are less likely to include “false positives”, things that may appear to be trends but are due to sampling error. However, identifying trends is definitely an area needing further research.

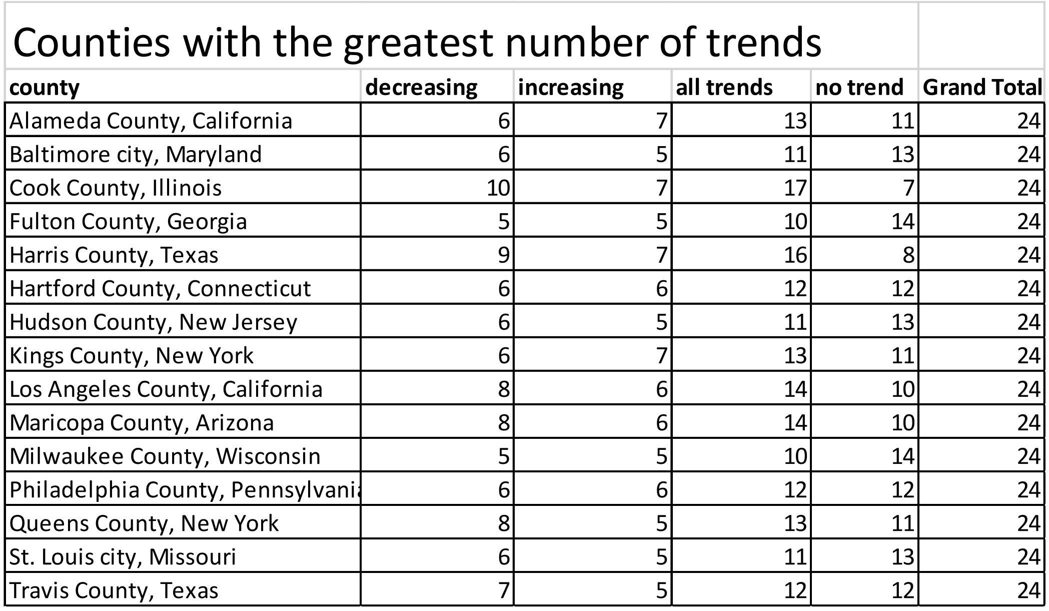
Table 6 Average S statistic value and average p value along with total number of occurrences for each trend result category

	Trend result	**Avg S value**	**Avg p value**	**Number of occurrences**
Results across all measures and all counties	decreasing	-69.78	0.0003	875
	increasing	71.69	0.0002	879
	no trend	-3.41	0.3306	16,030
			Total	**17,784**

There were 1,754 cases satisfying these criteria across all measures and all counties. This represents about 10% of all the measures across all counties. Table 7 shows the number of trends for each measure.

We can also look at these results from the counties perspective. Here are two tables from that perspective:

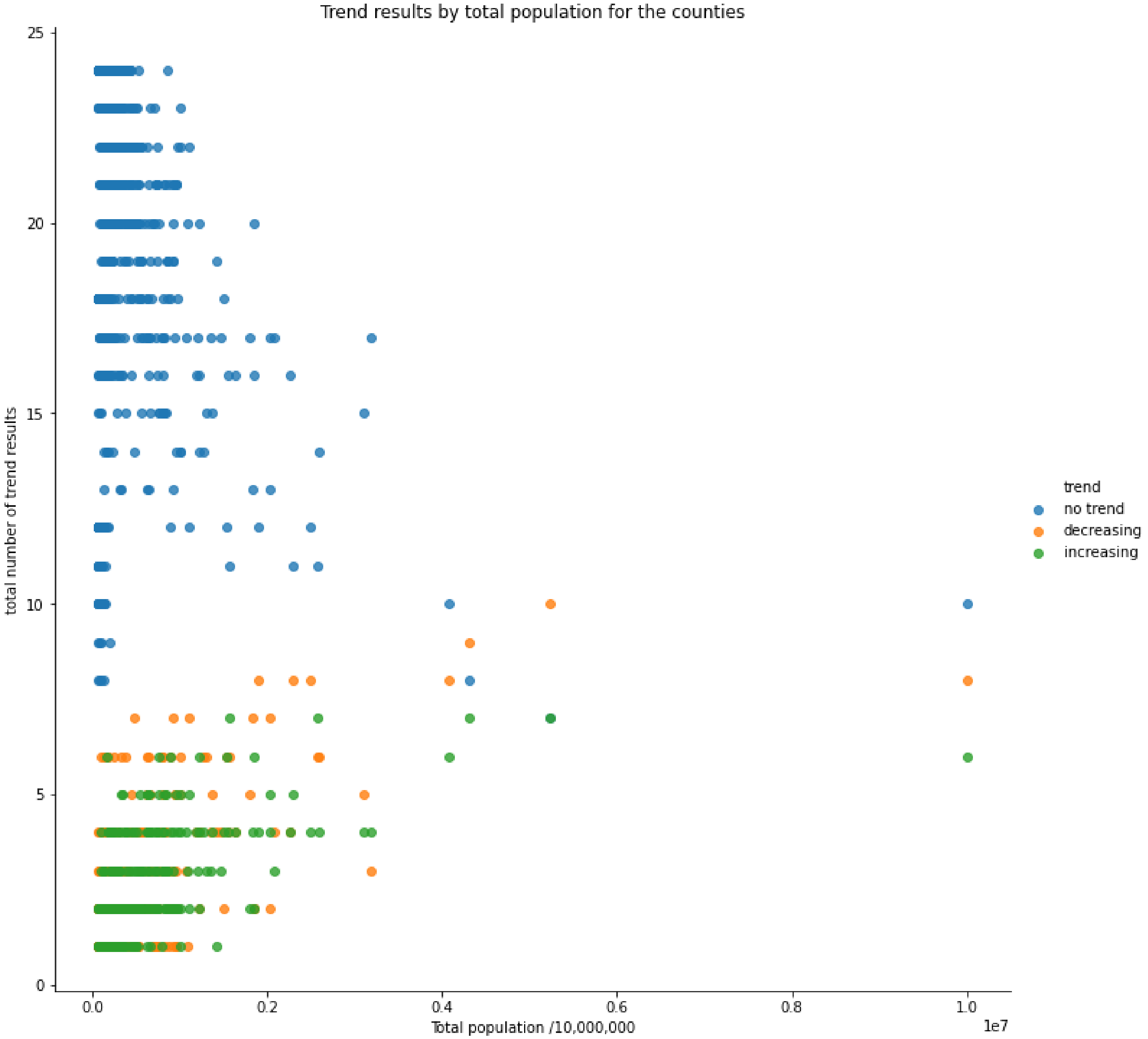
Table 7 Counties with 5 or more instances of both decreasing and increasing trends across all measures



Counties with the greatest number of trends					
county	**decreasing**	**increasing**	**all trends**	**no trend**	**Grand Total**
Alameda County, California	6	7	13	11	24
Baltimore city, Maryland	6	5	11	13	24
Cook County, Illinois	10	7	17	7	24
Fulton County, Georgia	5	5	10	14	24
Harris County, Texas	9	7	16	8	24
Hartford County, Connecticut	6	6	12	12	24
Hudson County, New Jersey	6	5	11	13	24
Kings County, New York	6	7	13	11	24
Los Angeles County, California	8	6	14	10	24
Maricopa County, Arizona	8	6	14	10	24
Milwaukee County, Wisconsin	5	5	10	14	24
Philadelphia County, Pennsylvani	6	6	12	12	24
Queens County, New York	8	5	13	11	24
St. Louis city, Missouri	6	5	11	13	24
Travis County, Texas	7	5	12	12	24

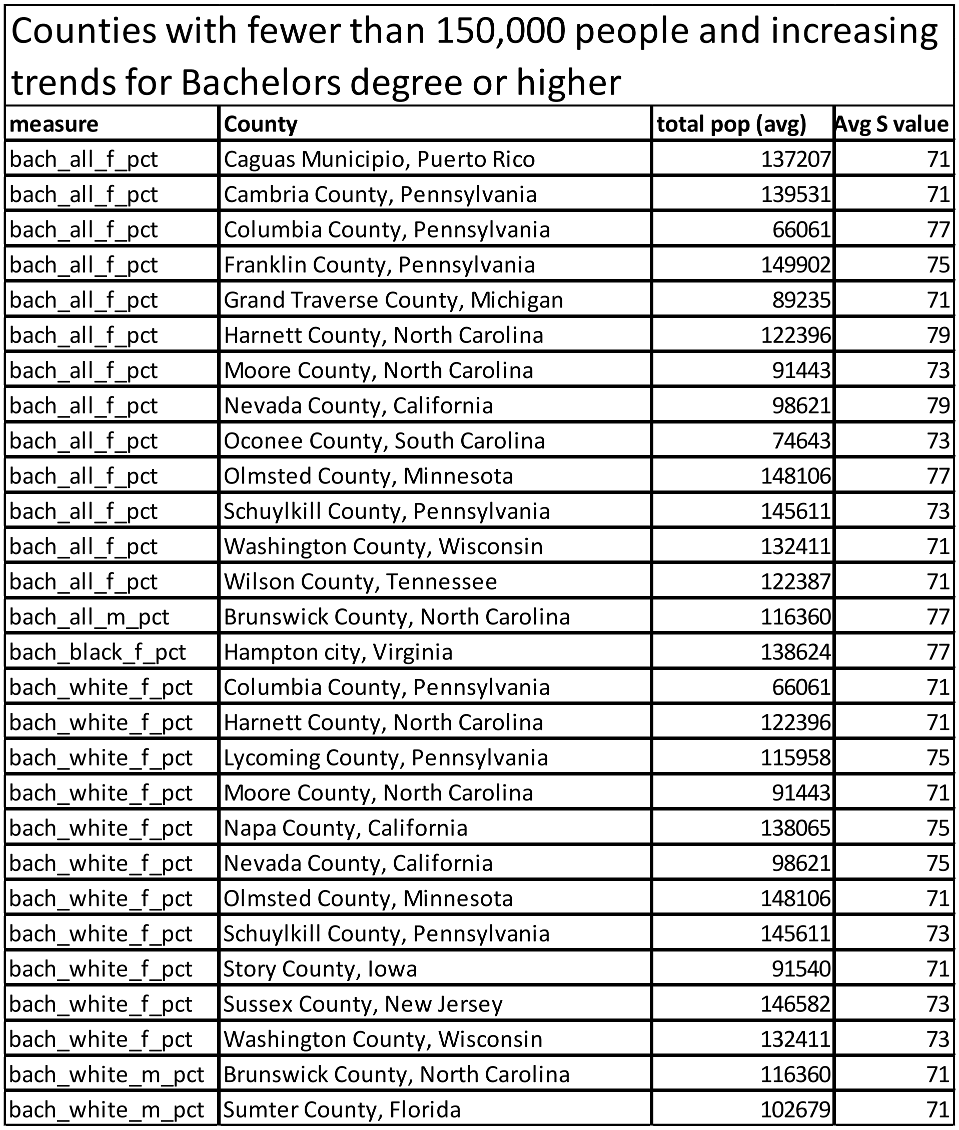
The decreasing trends are all related to no high school completion, and the increasing ones are all for bachelor’s degree or higher. These are all very large counties (including the two county equivalents Baltimore city and St. Louis city). 171 counties had no trends for any of the measures. Fig. 3 shows all trend results by county population size. Note that the blue dots for “no trend” are concentrated in the left hand part of the graph where the counties have smaller populations.

Figure 3 Trend results for all 24 measures by the total population (average over 2006-2019 period) for each county
NOTE: x-axis numbers are given as a fraction of 10,000,000



Here is a list of the smaller counties with the most trends for Bachelor's degree or higher:

Table 8 Smaller counties with increasing trends with an average S value > 70 for Bachelor's degree or higher



Counties with fewer than 150,000 people and increasing trends for Bachelors degree or higher

measure	County	total pop (avg)	Avg S value
bach_all_f_pct	Caguas Municipio, Puerto Rico	137207	71
bach_all_f_pct	Cambria County, Pennsylvania	139531	71
bach_all_f_pct	Columbia County, Pennsylvania	66061	77
bach_all_f_pct	Franklin County, Pennsylvania	149902	75
bach_all_f_pct	Grand Traverse County, Michigan	89235	71
bach_all_f_pct	Harnett County, North Carolina	122396	79
bach_all_f_pct	Moore County, North Carolina	91443	73
bach_all_f_pct	Nevada County, California	98621	79
bach_all_f_pct	Oconee County, South Carolina	74643	73
bach_all_f_pct	Olmsted County, Minnesota	148106	77
bach_all_f_pct	Schuylkill County, Pennsylvania	145611	73
bach_all_f_pct	Washington County, Wisconsin	132411	71
bach_all_f_pct	Wilson County, Tennessee	122387	71
bach_all_m_pct	Brunswick County, North Carolina	116360	77
bach_black_f_pct	Hampton city, Virginia	138624	77
bach_white_f_pct	Columbia County, Pennsylvania	66061	71
bach_white_f_pct	Harnett County, North Carolina	122396	71
bach_white_f_pct	Lycoming County, Pennsylvania	115958	75
bach_white_f_pct	Moore County, North Carolina	91443	71
bach_white_f_pct	Napa County, California	138065	75
bach_white_f_pct	Nevada County, California	98621	75
bach_white_f_pct	Olmsted County, Minnesota	148106	71
bach_white_f_pct	Schuylkill County, Pennsylvania	145611	73
bach_white_f_pct	Story County, Iowa	91540	71
bach_white_f_pct	Sussex County, New Jersey	146582	73
bach_white_f_pct	Washington County, Wisconsin	132411	73
bach_white_m_pct	Brunswick County, North Carolina	116360	71
bach_white_m_pct	Sumter County, Florida	102679	71

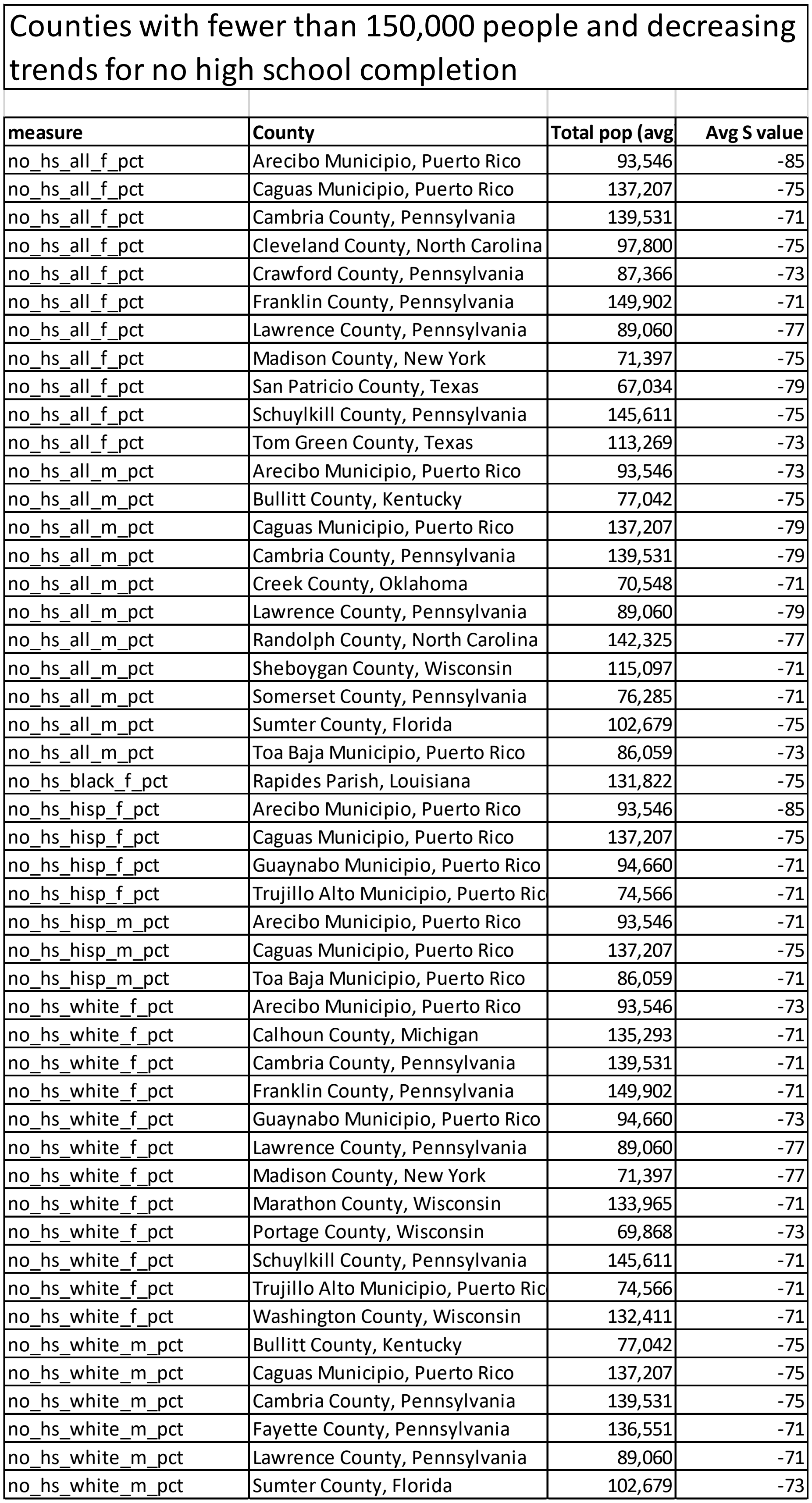
Table 8 is sorted by measure. So, it is more difficult to see the counties that occur more than once in the table:

- Brunswick County, North Carolina
- Columbia County, Pennsylvania
- Harnett County, North Carolina
- Schuykill County, Pennsylvania
- Washington County, Wisconsin.

Two other things to note about table 8: Except for Hampton city, Virginia, all trends are either for the entire population or for the White alone population. Also, note that there are only two counties that show an increasing trend in Bachelor's degree or higher for males.

Table 9 for decreasing trends in no high school completion in "small" counties is the counterpart to Table 8.

Table 9 Smaller counties with decreasing trends with an average S value < -70 for no high school completion



Counties with fewer than 150,000 people and decreasing trends for no high school completion

measure	County	Total pop (avg	Avg S value
no_hs_all_f_pct	Arecibo Municipio, Puerto Rico	93,546	-85
no_hs_all_f_pct	Caguas Municipio, Puerto Rico	137,207	-75
no_hs_all_f_pct	Cambria County, Pennsylvania	139,531	-71
no_hs_all_f_pct	Cleveland County, North Carolina	97,800	-75
no_hs_all_f_pct	Crawford County, Pennsylvania	87,366	-73
no_hs_all_f_pct	Franklin County, Pennsylvania	149,902	-71
no_hs_all_f_pct	Lawrence County, Pennsylvania	89,060	-77
no_hs_all_f_pct	Madison County, New York	71,397	-75
no_hs_all_f_pct	San Patricio County, Texas	67,034	-79
no_hs_all_f_pct	Schuylkill County, Pennsylvania	145,611	-75
no_hs_all_f_pct	Tom Green County, Texas	113,269	-73
no_hs_all_m_pct	Arecibo Municipio, Puerto Rico	93,546	-73
no_hs_all_m_pct	Bullitt County, Kentucky	77,042	-75
no_hs_all_m_pct	Caguas Municipio, Puerto Rico	137,207	-79
no_hs_all_m_pct	Cambria County, Pennsylvania	139,531	-79
no_hs_all_m_pct	Creek County, Oklahoma	70,548	-71
no_hs_all_m_pct	Lawrence County, Pennsylvania	89,060	-79
no_hs_all_m_pct	Randolph County, North Carolina	142,325	-77
no_hs_all_m_pct	Sheboygan County, Wisconsin	115,097	-71
no_hs_all_m_pct	Somerset County, Pennsylvania	76,285	-71
no_hs_all_m_pct	Sumter County, Florida	102,679	-75
no_hs_all_m_pct	Toa Baja Municipio, Puerto Rico	86,059	-73
no_hs_black_f_pct	Rapides Parish, Louisiana	131,822	-75
no_hs_hisp_f_pct	Arecibo Municipio, Puerto Rico	93,546	-85
no_hs_hisp_f_pct	Caguas Municipio, Puerto Rico	137,207	-75
no_hs_hisp_f_pct	Guaynabo Municipio, Puerto Rico	94,660	-71
no_hs_hisp_f_pct	Trujillo Alto Municipio, Puerto Ric	74,566	-71
no_hs_hisp_m_pct	Arecibo Municipio, Puerto Rico	93,546	-71
no_hs_hisp_m_pct	Caguas Municipio, Puerto Rico	137,207	-75
no_hs_hisp_m_pct	Toa Baja Municipio, Puerto Rico	86,059	-71
no_hs_white_f_pct	Arecibo Municipio, Puerto Rico	93,546	-73
no_hs_white_f_pct	Calhoun County, Michigan	135,293	-71
no_hs_white_f_pct	Cambria County, Pennsylvania	139,531	-71
no_hs_white_f_pct	Franklin County, Pennsylvania	149,902	-71
no_hs_white_f_pct	Guaynabo Municipio, Puerto Rico	94,660	-73
no_hs_white_f_pct	Lawrence County, Pennsylvania	89,060	-77
no_hs_white_f_pct	Madison County, New York	71,397	-77
no_hs_white_f_pct	Marathon County, Wisconsin	133,965	-71
no_hs_white_f_pct	Portage County, Wisconsin	69,868	-73
no_hs_white_f_pct	Schuylkill County, Pennsylvania	145,611	-71
no_hs_white_f_pct	Trujillo Alto Municipio, Puerto Ric	74,566	-71
no_hs_white_f_pct	Washington County, Wisconsin	132,411	-71
no_hs_white_m_pct	Bullitt County, Kentucky	77,042	-75
no_hs_white_m_pct	Caguas Municipio, Puerto Rico	137,207	-75
no_hs_white_m_pct	Cambria County, Pennsylvania	139,531	-75
no_hs_white_m_pct	Fayette County, Pennsylvania	136,551	-71
no_hs_white_m_pct	Lawrence County, Pennsylvania	89,060	-71
no_hs_white_m_pct	Sumter County, Florida	102,679	-73

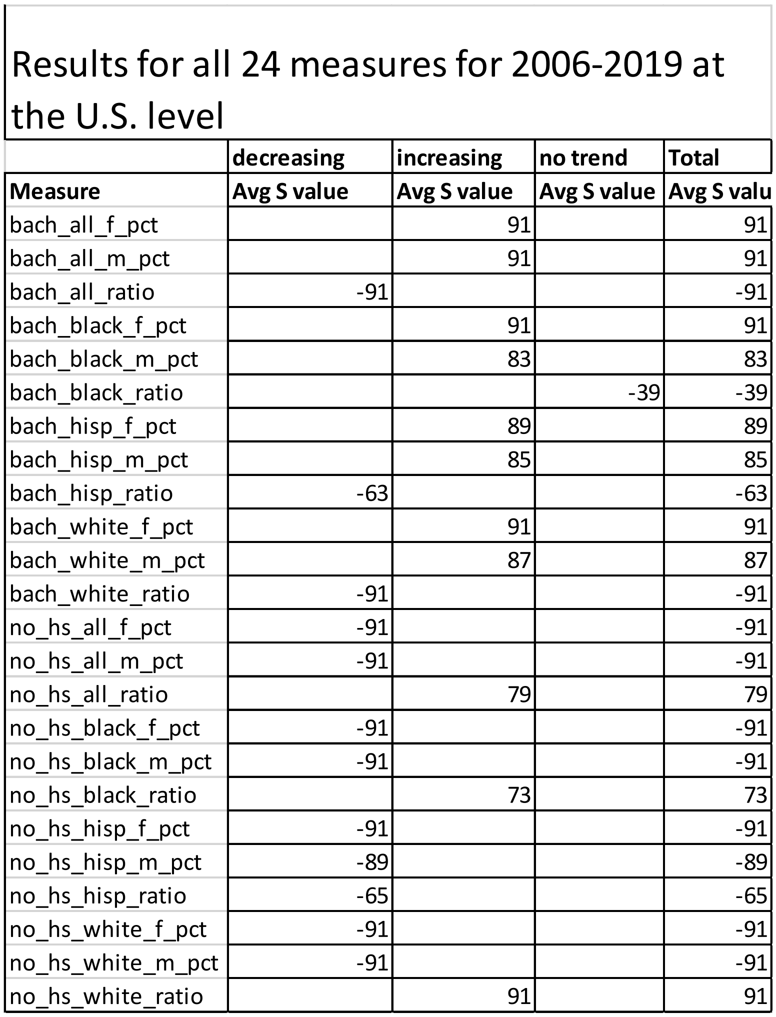
One final note about counties with trends: Pennsylvania had the greatest number of county-level trends at 153.

7. Results of creating the 24 measures at higher levels of geography

I created each measure at three other geographic levels: U.S.; states; and metro and non-metro parts of each state. These results are summarized in this section.

7.1. U.S.

Table 10 Results of measures for 2006-2019 at U.S. level



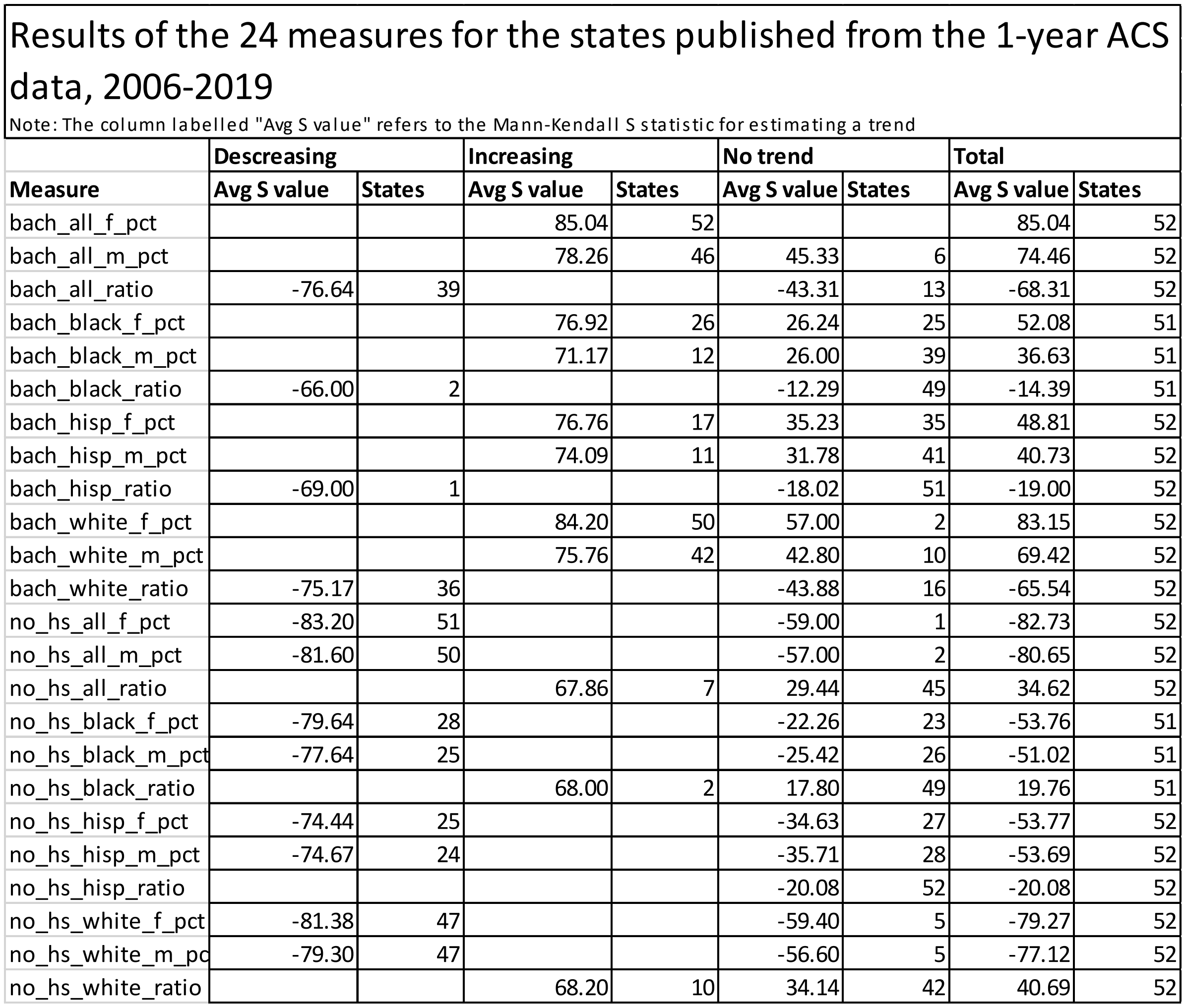
Results for all 24 measures for 2006-2019 at the U.S. level

	decreasing	**increasing**	**no trend**	**Total**
Measure	**Avg S value**	**Avg S value**	**Avg S value**	**Avg S valu**
bach_all_f_pct		91		91
bach_all_m_pct		91		91
bach_all_ratio	-91			-91
bach_black_f_pct		91		91
bach_black_m_pct		83		83
bach_black_ratio			-39	-39
bach_hisp_f_pct		89		89
bach_hisp_m_pct		85		85
bach_hisp_ratio	-63			-63
bach_white_f_pct		91		91
bach_white_m_pct		87		87
bach_white_ratio	-91			-91
no_hs_all_f_pct	-91			-91
no_hs_all_m_pct	-91			-91
no_hs_all_ratio		79		79
no_hs_black_f_pct	-91			-91
no_hs_black_m_pct	-91			-91
no_hs_black_ratio		73		73
no_hs_hisp_f_pct	-91			-91
no_hs_hisp_m_pct	-89			-89
no_hs_hisp_ratio	-65			-65
no_hs_white_f_pct	-91			-91
no_hs_white_m_pct	-91			-91
no_hs_white_ratio		91		91

The only surprising result is that there is no trend for the ratio of Black alone men to Black alone women with Bachelors degree or higher. The other results are consistent with results already recorded in other analyses. It is worth noting that 14 measures have an absolute S value of 91; i.e., the linear curve is strictly monotonic (increasing or decreasing) since 91 is the maximum number of values that can be compared in the S measure for a 14-year period.

7.2. States

Table 11 Results of the 24 measures for 2006-2019 at the state level ‘



Results of the 24 measures for the states published from the 1-year ACS data, 2006-2019

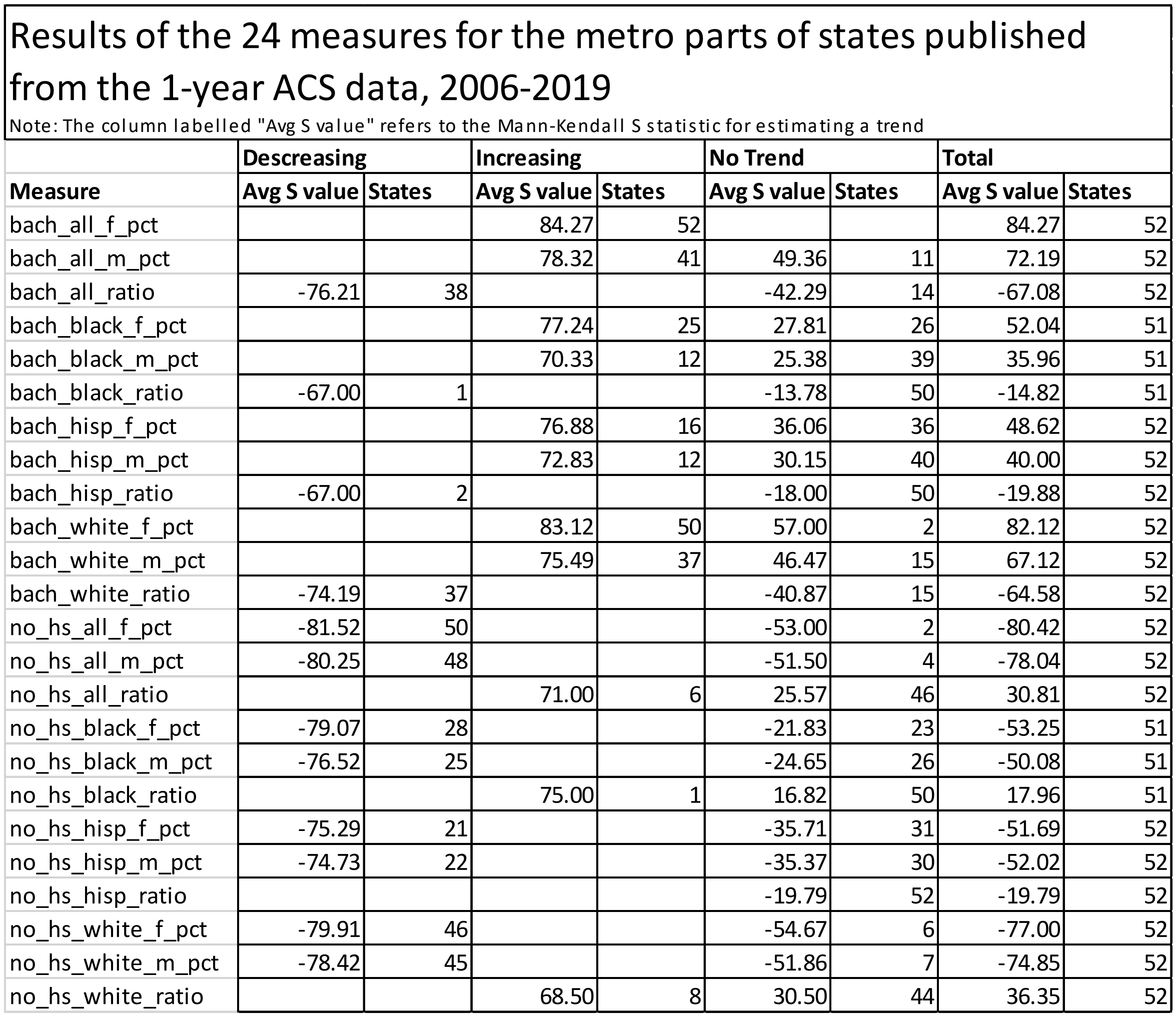
Note: The column labelled "Avg S value" refers to the Mann-Kendall S statistic for estimating a trend

	Descreasing		Increasing		No trend		Total	
Measure	Avg S value	States	Avg S value	States	Avg S value	States	Avg S value	States
bach_all_f_pct			85.04	52			85.04	52
bach_all_m_pct			78.26	46	45.33	6	74.46	52
bach_all_ratio	-76.64	39			-43.31	13	-68.31	52
bach_black_f_pct			76.92	26	26.24	25	52.08	51
bach_black_m_pct			71.17	12	26.00	39	36.63	51
bach_black_ratio	-66.00	2			-12.29	49	-14.39	51
bach_hisp_f_pct			76.76	17	35.23	35	48.81	52
bach_hisp_m_pct			74.09	11	31.78	41	40.73	52
bach_hisp_ratio	-69.00	1			-18.02	51	-19.00	52
bach_white_f_pct			84.20	50	57.00	2	83.15	52
bach_white_m_pct			75.76	42	42.80	10	69.42	52
bach_white_ratio	-75.17	36			-43.88	16	-65.54	52
no_hs_all_f_pct	-83.20	51			-59.00	1	-82.73	52
no_hs_all_m_pct	-81.60	50			-57.00	2	-80.65	52
no_hs_all_ratio			67.86	7	29.44	45	34.62	52
no_hs_black_f_pct	-79.64	28			-22.26	23	-53.76	51
no_hs_black_m_pct	-77.64	25			-25.42	26	-51.02	51
no_hs_black_ratio			68.00	2	17.80	49	19.76	51
no_hs_hisp_f_pct	-74.44	25			-34.63	27	-53.77	52
no_hs_hisp_m_pct	-74.67	24			-35.71	28	-53.69	52
no_hs_hisp_ratio					-20.08	52	-20.08	52
no_hs_white_f_pct	-81.38	47			-59.40	5	-79.27	52
no_hs_white_m_pc	-79.30	47			-56.60	5	-77.12	52
no_hs_white_ratio			68.20	10	34.14	42	40.69	52

Even at the state level, the results are quite different form the U.S. level, and we see many instances of measures with no trends for several states. This table also reveals that the ratio measures are weaker than the others with regard to trends. Also, for every group and for both genders the proportion not completing high school decreased in many states. (The C15002B table for the Black alone population was suppressed for too many years in Montana to allow for an MK trend test.)

7.3. Metro and non-metro parts of the states

Table 12 Results of the 24 measures for 2006-2019 for the metropolitan areas within states



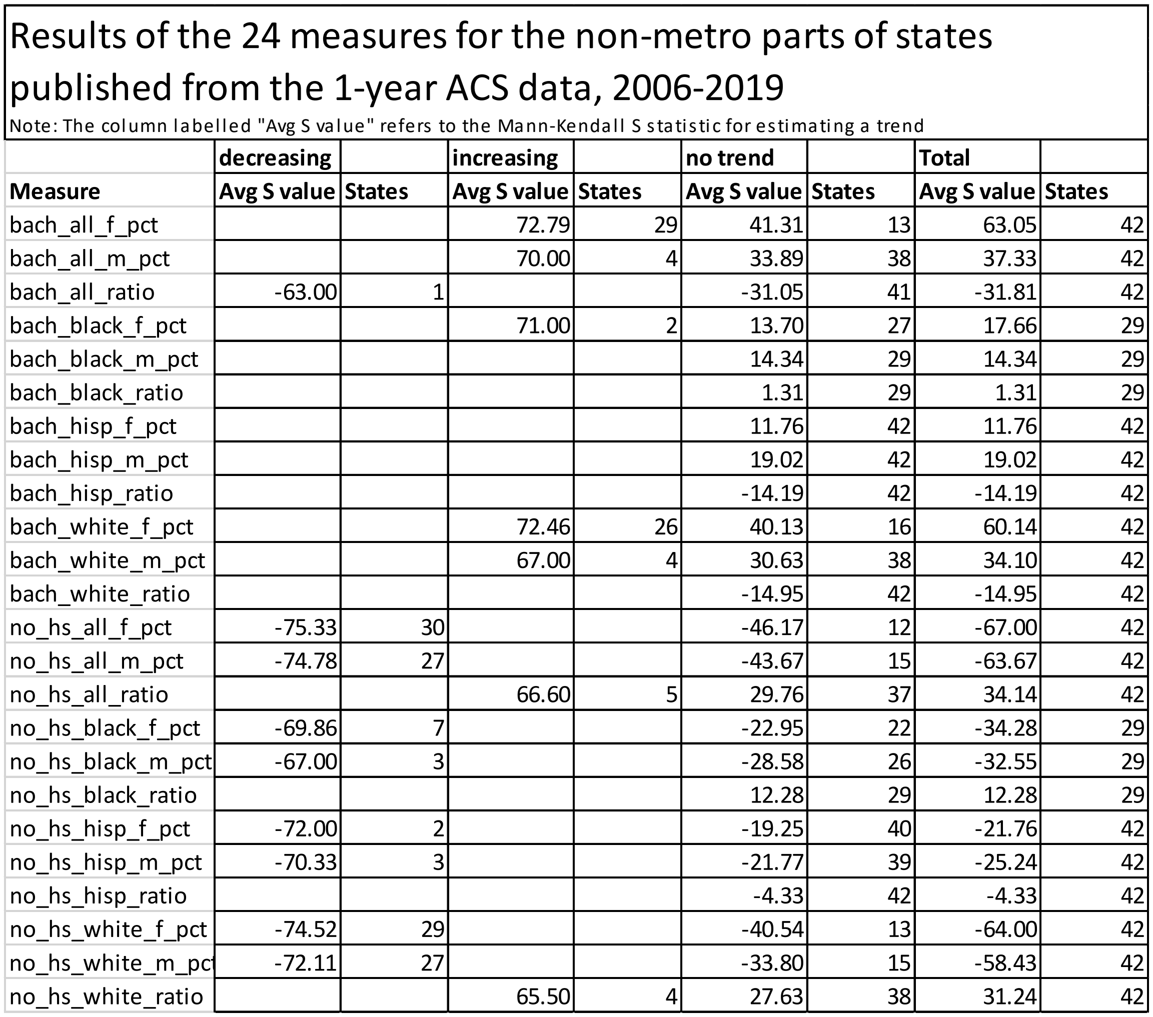
Results of the 24 measures for the metro parts of states published from the 1-year ACS data, 2006-2019

Note: The column labelled "Avg S value" refers to the Mann-Kendall S statistic for estimating a trend

	Descreasing		**Increasing**		**No Trend**		**Total**	
Measure	**Avg S value**	**States**	**Avg S value**	**States**	**Avg S value**	**States**	**Avg S value**	**States**
bach_all_f_pct			84.27	52			84.27	52
bach_all_m_pct			78.32	41	49.36	11	72.19	52
bach_all_ratio	-76.21	38			-42.29	14	-67.08	52
bach_black_f_pct			77.24	25	27.81	26	52.04	51
bach_black_m_pct			70.33	12	25.38	39	35.96	51
bach_black_ratio	-67.00	1			-13.78	50	-14.82	51
bach_hisp_f_pct			76.88	16	36.06	36	48.62	52
bach_hisp_m_pct			72.83	12	30.15	40	40.00	52
bach_hisp_ratio	-67.00	2			-18.00	50	-19.88	52
bach_white_f_pct			83.12	50	57.00	2	82.12	52
bach_white_m_pct			75.49	37	46.47	15	67.12	52
bach_white_ratio	-74.19	37			-40.87	15	-64.58	52
no_hs_all_f_pct	-81.52	50			-53.00	2	-80.42	52
no_hs_all_m_pct	-80.25	48			-51.50	4	-78.04	52
no_hs_all_ratio			71.00	6	25.57	46	30.81	52
no_hs_black_f_pct	-79.07	28			-21.83	23	-53.25	51
no_hs_black_m_pct	-76.52	25			-24.65	26	-50.08	51
no_hs_black_ratio			75.00	1	16.82	50	17.96	51
no_hs_hisp_f_pct	-75.29	21			-35.71	31	-51.69	52
no_hs_hisp_m_pct	-74.73	22			-35.37	30	-52.02	52
no_hs_hisp_ratio					-19.79	52	-19.79	52
no_hs_white_f_pct	-79.91	46			-54.67	6	-77.00	52
no_hs_white_m_pct	-78.42	45			-51.86	7	-74.85	52
no_hs_white_ratio			68.50	8	30.50	44	36.35	52

Notice that the differences between men and women are not as great for the metropolitan areas of states as they are at the county level. In fact, for one group, Hispanic origin, the number of counties showing decreasing trends in no high school completion for men is slightly greater than the same result for women.

Table 13 Results of the 24 measures for 2006-2019 for the non-metropolitan areas within states



Results of the 24 measures for the non-metro parts of states published from the 1-year ACS data, 2006-2019

Note: The column labelled "Avg S value" refers to the Mann-Kendall S statistic for estimating a trend

	decreasing		increasing		no trend		Total	
Measure	**Avg S value**	**States**	**Avg S value**	**States**	**Avg S value**	**States**	**Avg S value**	**States**
bach_all_f_pct			72.79	29	41.31	13	63.05	42
bach_all_m_pct			70.00	4	33.89	38	37.33	42
bach_all_ratio	-63.00	1			-31.05	41	-31.81	42
bach_black_f_pct			71.00	2	13.70	27	17.66	29
bach_black_m_pct					14.34	29	14.34	29
bach_black_ratio					1.31	29	1.31	29
bach_hisp_f_pct					11.76	42	11.76	42
bach_hisp_m_pct					19.02	42	19.02	42
bach_hisp_ratio					-14.19	42	-14.19	42
bach_white_f_pct			72.46	26	40.13	16	60.14	42
bach_white_m_pct			67.00	4	30.63	38	34.10	42
bach_white_ratio					-14.95	42	-14.95	42
no_hs_all_f_pct	-75.33	30			-46.17	12	-67.00	42
no_hs_all_m_pct	-74.78	27			-43.67	15	-63.67	42
no_hs_all_ratio			66.60	5	29.76	37	34.14	42
no_hs_black_f_pct	-69.86	7			-22.95	22	-34.28	29
no_hs_black_m_pct	-67.00	3			-28.58	26	-32.55	29
no_hs_black_ratio					12.28	29	12.28	29
no_hs_hisp_f_pct	-72.00	2			-19.25	40	-21.76	42
no_hs_hisp_m_pct	-70.33	3			-21.77	39	-25.24	42
no_hs_hisp_ratio					-4.33	42	-4.33	42
no_hs_white_f_pct	-74.52	29			-40.54	13	-64.00	42
no_hs_white_m_pc	-72.11	27			-33.80	15	-58.43	42
no_hs_white_ratio			65.50	4	27.63	38	31.24	42

It is important to note that there are 9 states (including D.C.) with either no land that is not in a metropolitan or micropolitan area or with a population under 65,000. Therefore, the rightmost column for the non-metro portion of states is at most 42. Also, the column for the number of states with no trend in this table is much greater than the corresponding column in the table for the metropolitan areas in the states. Even for measures of the total population, trend results are much weaker here than in the metro parts of the states. The non-metro table also shows that data for the Black alone population outside the metro areas is much more limited; only 29 states have results for the Black population measures. Although the sample size for the non-metro parts of states is much smaller than the metro parts for most states, most of these areas have more than 250,000 residents. Nevertheless, there is greater sampling variability here, and that may be affecting these results. Further examination of sampling variability, perhaps calculating a median coefficient of variation value for each measure, would be needed to see the extent to which sampling variability is affecting these results.

8. Findings

- My results are consistent with results on both measures at the state and national level. A few references are given in Appendix 3 of this paper.
- A limitation of the high school completion results: based solely on the population 25 years and older. So, a downward trend (= improvement in high school completion) may not be due to any program or policy implemented by a school system or county. For example, in-migration of people could affect these results.
- It would be helpful to further disaggregate these results by age groups (e.g. 25-34, 35-50, etc.). However, that may lead to so much data loss and reduced statistical quality in measures at the county level to force the analysis to be done at a higher geographic level.
- Many more "no trend" results in the non-metro portions of states versus the metro portions is a result requiring further analysis. For example, the proportion of Hispanic women not completing high school showed a decreasing trend in the non-metro parts of only 7 states, whereas the same measure had a decreasing trend in the metro parts of 32 states. Can this be because the metro portions of states are much more populous and thus the measure estimates are less variable?

9. Conclusions and lessons learned

When I conceived of the idea for this project I thought it would be fairly straightforward. However, the work brought with it many surprises and challenges. The easy part was downloading the data to start with once I had determined that I could use the same detailed tables for each of the 14 years. It is very simple to get the data from the Census Bureau's API. Some of the problems I encountered were self-inflicted since I had switched from using R to using Python and was still a Python neophyte. However, Python met all my needs; it just took me a while to figure out how to make that happen.

I have tried to make a convincing case for using counties instead of other geographic types, but the primary reason for using counties was the fact that the 2010 Census occurred during the 2006-2019 period. This ruled out using PUMS data or the full sample data products at the PUMA geographic level.

The most difficult part of the project was defining what I call a "trend" and finding a way to detect trends. I would not be surprised to learn that there is a better way to detect trends, but I found MK to be very understandable, and I could easily test a trend hypothesis (I used a module for Python called pymannkendall.py).

When I do my next "trend detection" project, I will try to incorporate what I have learned and answer these questions before I even start to download data:

a. Can I use the pre-aggregated data at the PUMA geographic level; i.e., am I using a time period in which the PUMA boundary definitions do not change?
b. Is there a table in one of the ACS data products that supports what I want to measure and does not change throughout the time period? If not, can I use the PUMS microdata to create my estimates without sacrificing too much in reliability?
c. Is there a way for me to check my results against some already created data product, perhaps at the margins of a table or only at a higher geographic level?
d. Should I use the 1-year ACS data products, or are there enough non-overlapping 5-year periods allowing me to use the 5-year ACS estimates? A related question: is there some way to use the 5-year estimates to "fill in the hole" of the counties too small for 1-year estimates? Or, can I "fill the hole" using the 1-year PUMS data for each year in the time period?
e. Is the population subgroup I plan to study distributed fairly evenly across geography, or in a very skewed manner (as is the case for the Black alone population)?

f. Is there an alternative to using the pre-aggregated data? For example, could the annual PUMS data be used to follow an age cohort over time?

Appendices

Appendix 1. Graphs of trends for some measures in selected counties

A note about the use of linear regression in the graphs: even though the “year” is treated as the independent variable in this simple linear regression for a measure (the dependent variable), this does not mean that I think the mere passage of time influences the results for a measure. This use of simple linear regression is aimed solely at finding a “best-fit” line (aka “trendline”) and using the r-squared value to evaluate how well this line minimizes the sum of the squared residuals; i.e., how close the “best-fit” line comes to the actual data points.

These first two graphs are for Black alone males and females in Baltimore city without high school completion. Both measures met the two criteria to be rated as a trend for 2006-2019. (These results are being reviewed now by an analyst at the Baltimore Metropolitan Council.)

Figure 4 decreasing trend: S value of 79 and p-value around 0.0002

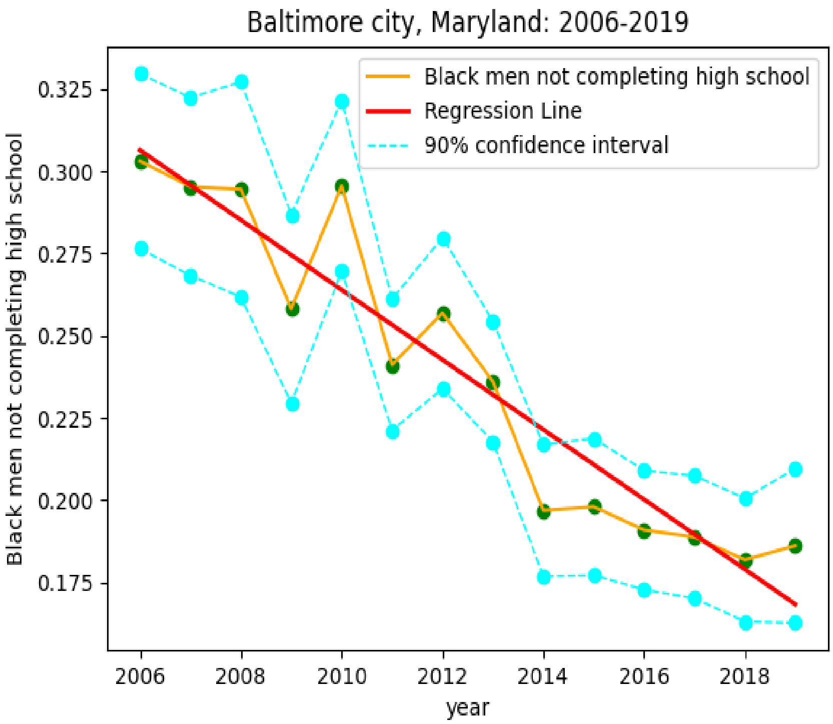


Figure 5 decreasing trend: S value of 79 and p-value around 0.0002

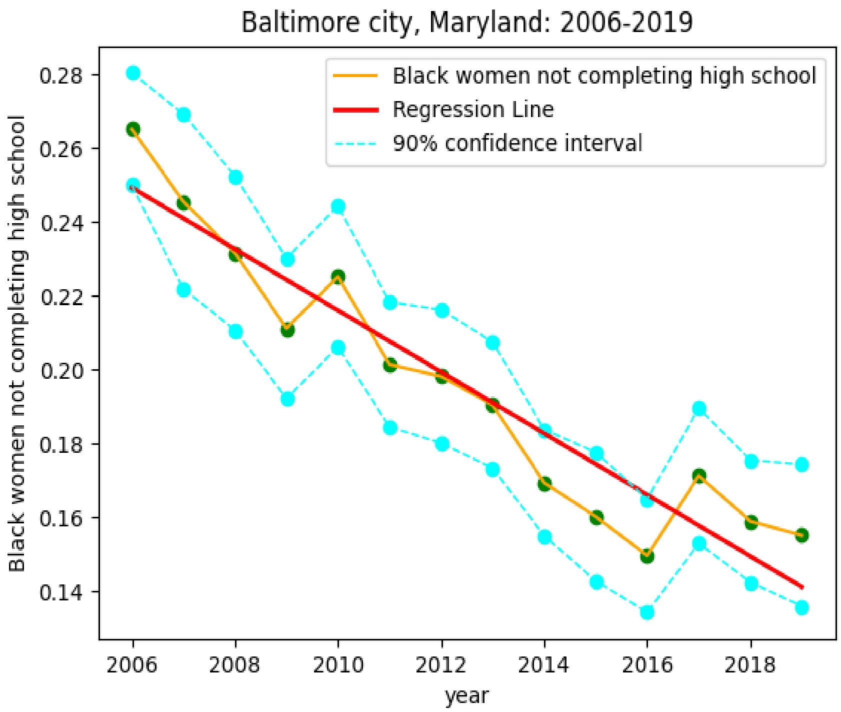


Figure 6 Example of a strong trend for a county just above the 65,000 threshold. Columbia County population is ~ 66,000.

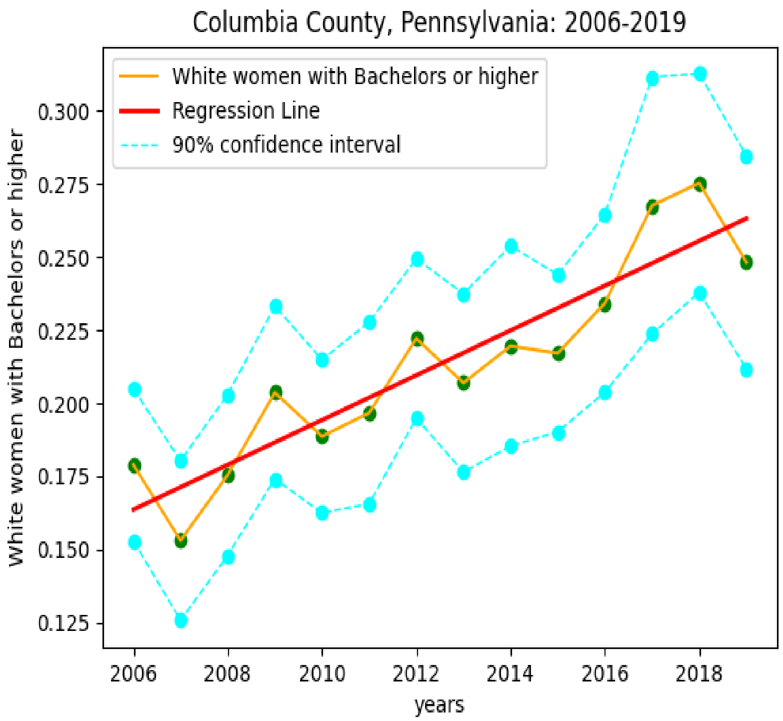


Figure 7 An example of a county that appears to have a decreasing trend for the proportion of Black women with a Bachelor's degree or higher. However, this case did not meet the p-value threshold of 0.001 Its p-value was 0.004, and the S value was 53. Therefore, it is a "no trend" case.

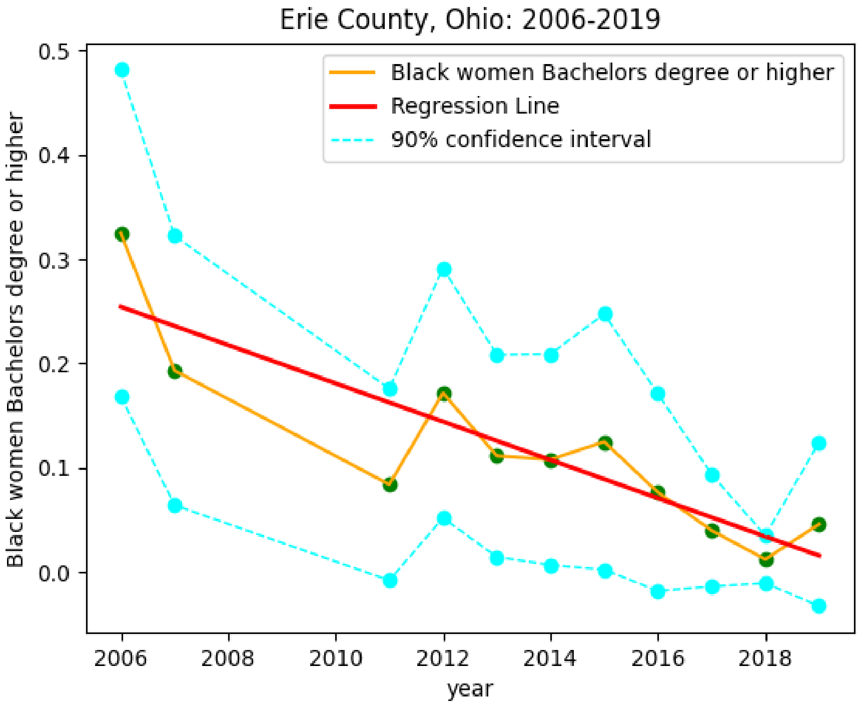


Figure 8 This case was judged to be an increasing trend even though there is considerable variation over the years. The S value is 73 and the p-value is about 0.0009. Confidence intervals not shown for every year because denominator of estimate was controlled starting in 2010.

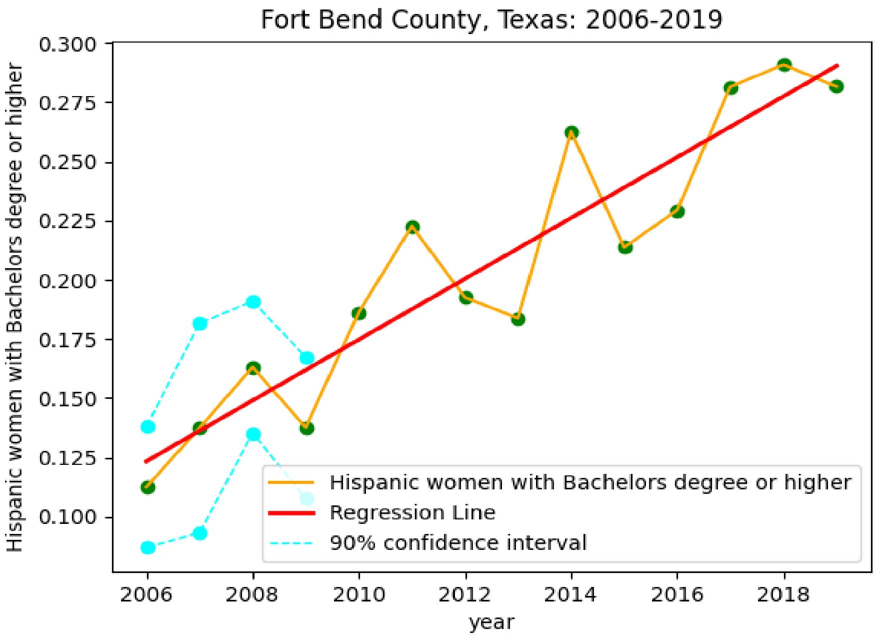
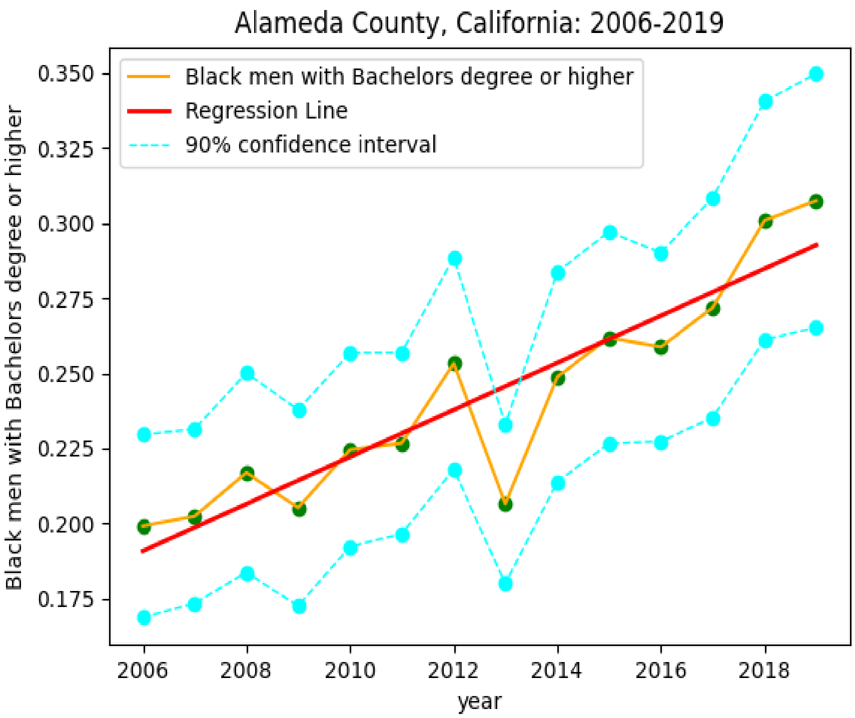


Figure 9 Increasing trend example



Appendix 2. U.S. county map with counties used in the study highlighted

The map image below shows the 817 counties that were available from the 1-year ACS data, 2006-2019. Each state had at least one county in the 817, and Texas had the most with 52. Each of these counties has at least 10 years of data for the measure analyzed, proportion of white women with no high school completion. The pink markers are for counties with a trend. All trends were decreasing. The second map is a detail from the full U.S. map. This map shows a number of counties near Pittsburgh, PA and Youngstown, OH with trend results, most of them showing a trend. The underlying thematic map is for the 2019 median household income based on the 5-year ACS estimates. A tool like Social Explorer can be useful in the early stages of looking for possible "independent" variables that might influence the results for the measure because the underlying theme can be easily changed to another characteristic while the markers remain on the map unchanged.

Figure 10 county level median household income map from Social Explorer with markers for counties analyzed for trend for measure no_hs_white_f_pct

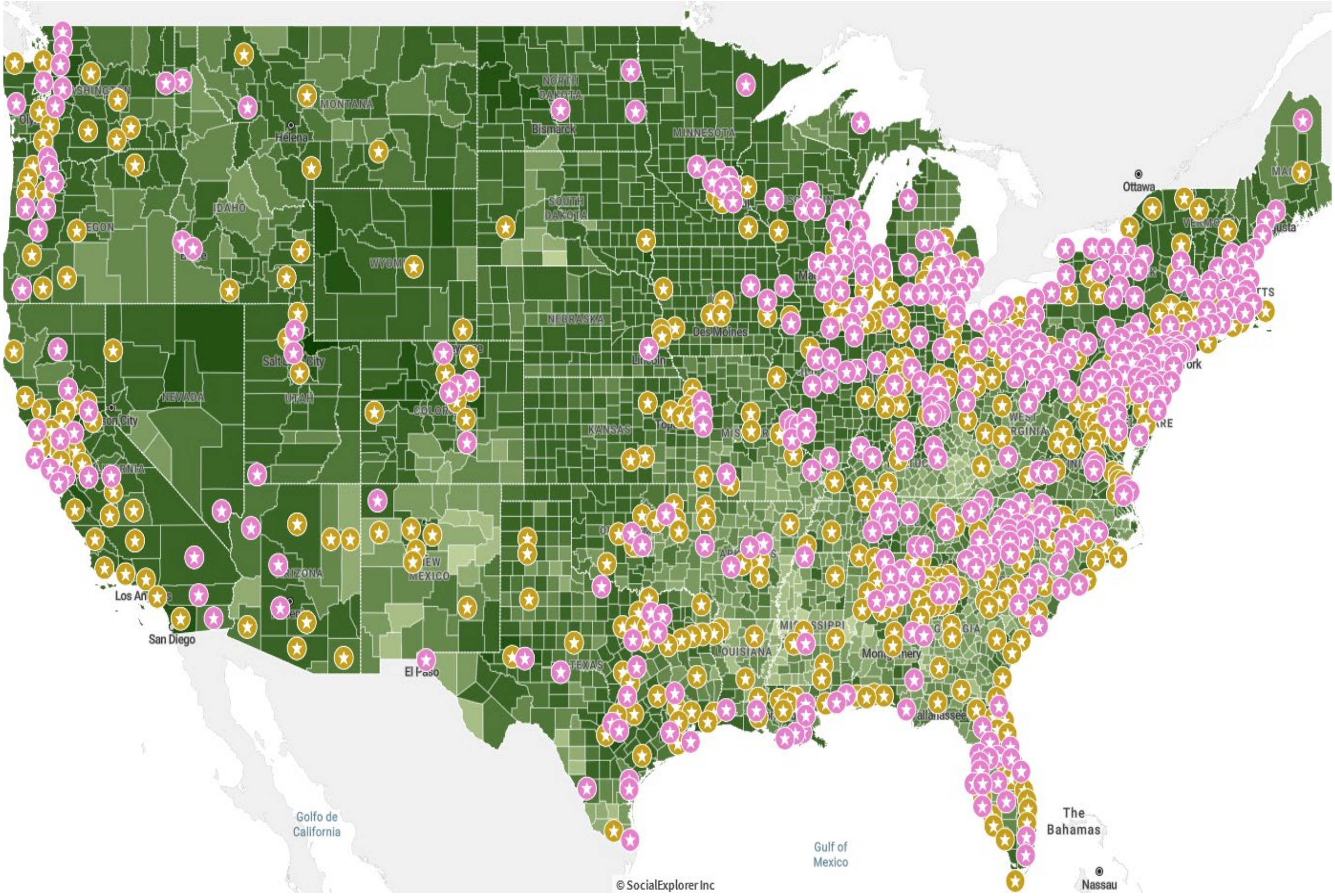
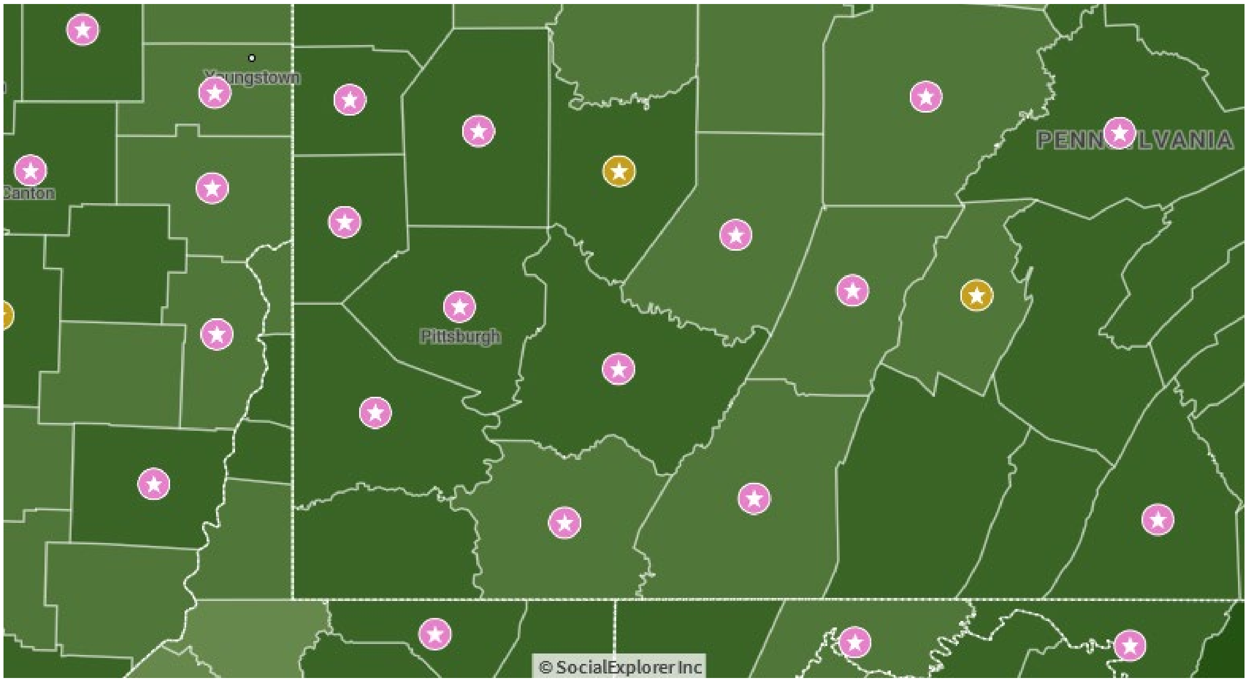


Figure 11 counties in Western PA and Eastern OH with trend results



Appendix 3. Notes and references

A good description of the Mann-Kendall trend test can be found in https://cran.r-project.org/web/packages/trend/vignettes/trend.pdf

A more thorough description of Mann-Kendall is at https://vsp.pnnl.gov/help/vsample/design_trend_mann_kendall.htm

Python code for the Mann-Kendall test can be found at https://github.com/mmhs013/pyMannKendall/blob/master/pymannkendall/pymannkendall.py

High school dropout rates measured over 2006-2018 are summarized at https://nces.ed.gov/programs/coe/indicator_coj.asp

Changes in educational attainment for various levels from 2000 to 2019 are shown by gender and by race and Hispanic origin groups at https://nces.ed.gov/programs/coe/indicator_caa.asp