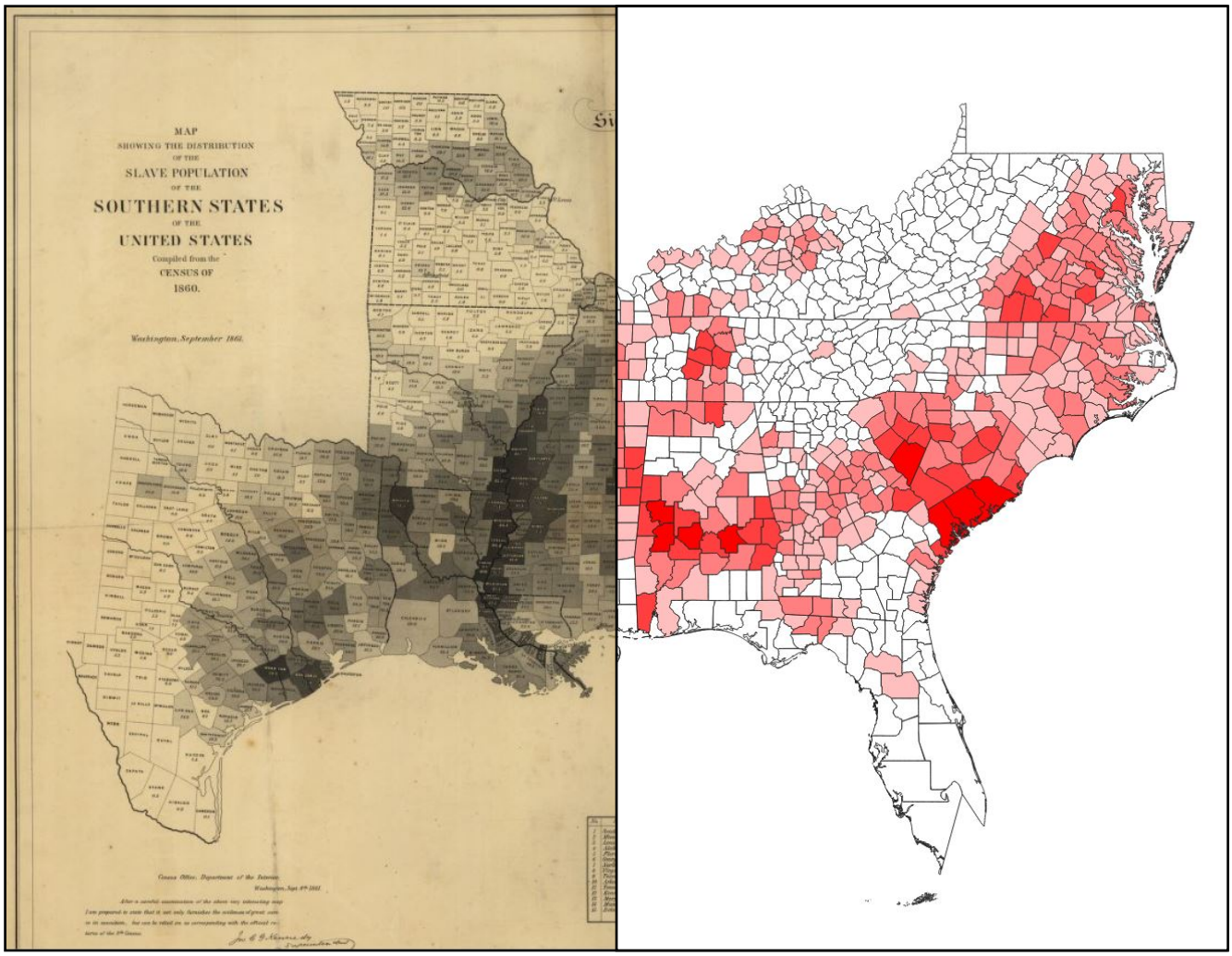
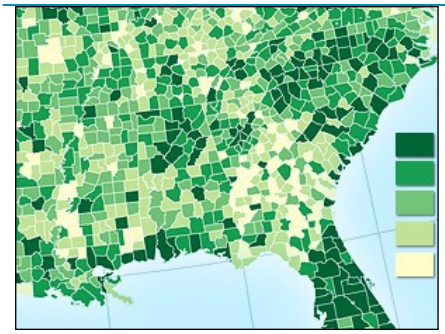
Census Data & Choropleths

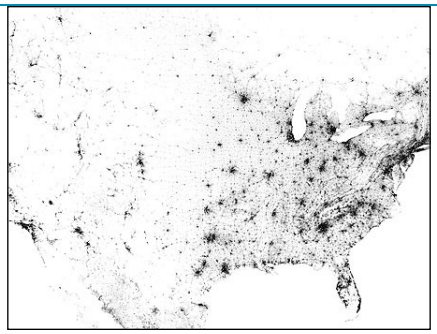


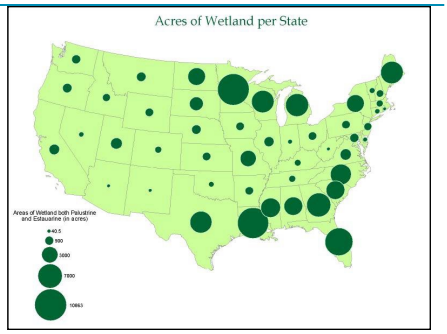
Introduction

Historical census data are among the richest sources of social and environmental information available for historical GIS. Until recently, its greatest limitation has been the cost- and labor-intensive process of transcribing printed materials and translating disparate formats to modern, GIS ready databases. Over the last ten years, governmental and academic institutions have released detailed digitized census archives dating to the 18th century (and earlier) in an number of geographic locations in the US and many other countries. In this exercise, you will use historical agricultural census data to examine census attribute information and produce choropleth maps of the American South on the eve of Civil War.

Choropleth Maps | Dot Density Maps | Proportional Symbols







Choropleth Maps

Choropleth maps are among the most popular maps employed today (and consequently among the most abused). This is likely because people tend to think of themselves in spaces defined by administrative boundaries (states/provinces/countries), which in this context we can call *enumeration units*. According to cartographic convention, choropleths must:

1. **Be organized according to enumeration units**
2. **Show ratios/proportions/rates etc. (data normalized over space)**
3. **Be able to measure the mapped theme anywhere on its surface**

One does not have to employ choropleths, however. Other thematic mapping styles offer appropriate, visually clear data visualizations. **Dot density** maps are simple and intuitive methods of showing clustered/raw data. (caution: they should always be mapped on equal area projections)/ **Graduated** or **Proportional Symbol** maps are also used to display raw data. As opposed to dot density maps, it is visually easier to estimate relative magnitude of a phenomenon with proportional symbols.

Instructions:

Today, we will explore the use of census data to produce correct and incorrect choropleth visualizations using data from Rick Thomas's book *In Time and Place*. This census dataset is available through the HGIS website.

1. Create a new folder entitled "Week 6" on your computer. All of the data you need for this exercise is stored in a shared google drive. You can access it here: https://drive.google.com/drive/folders/1GHpu3ym3JNuW5R-lQS6M815OSwurGK9m?usp=sharing
Go to the shared Google Drive folder for Week 6 (US South 1860). Download and save all the files to the Week 6 folder on your computer. If you haven't already, you should bookmark the site.

2. Open QGIS. Feel free to load a web base layer. Import the **USSouth1860.shp** shapefile. ay attention to the file extension, here. .shp refers to "shapefile." Click “layer" from the top

toolbar, then ”add vector layer", click the ellipses and navigate to the folder you just downloaded from google drive. Click the file that says "USSouth1860.shp", if there are two, look at the file type and it should say "SHP file” or “ESRI shape file” on a Mac. Click "add".

3. The shapefile includes county level data for 14 historic states (currently 15) . *Where is the state we've added since 1860*?

4. Right now, your map is uniformly shaded the same color. We can change the symbology of the map to a choropleth visualization. We know that at least one of the criteria for choropleths has already been met. This map has **enumeration units** (in this case, county boundaries). Let's proceed from there.

5. The first step when dealing with census data (or really any data you are trying to visualize) is to look at the attribute table. Right click USSouth1860 and open the table. This is a lot more detailed than the attribute table you created last week, right? Still, it is fundamentally no different to create this. What do you see?

6. Some of the categories will be intuitive. "State, county" etc. But many more won't be. Three important categories that may not be immediately recognizable to you are columns 18, 19, and 20 “GISJOIN” "GISJOIN2" and "GISJOIN_1." Although it isn't apparent from the shapefile shared in the Google Drive, census data you download online often does not arrive packaged as county or state shapefiles (its just a spreadsheet). Census data in tabular formats need to be “**joined**” (sometimes called concatenated) to georeferenced administrative polygon shapefile (called “boundary files). The column “GISJOIN” contains numbers that refer to counties in a system employed by the National Historical GIS (NHGIS). To link the census **table** with its **boundary files,** you would use this “join field,” since linking tables requires at least one column with similar information. Many of you will be using census data with your final project and there is a tutorial available for joining census data to shapefiles on our course website. If all of this seems like Greek to you, it won't be later this semester when we create datasets from the census.

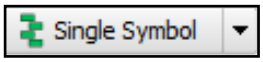
4	290	0.00000000000	370	0290	902.00000000000	1746289.291399...	58112.25771180...	G3700290	3700290	G3700290	1860	North Carolina	Camden	5343	2942	27
5	1090	0.00000000000	470	1090	1895.00000000000	669290.8963760...	-230307.042255...	G4701090	4701090	G4701090	1860	Tennessee	McNairy	14732	12810	2
6	470	0.00000000000	510	0470	1081.00000000000	1549804.878470...	258089.4648930...	G5100470	5100470	G5100470	1860	Virginia	Culpeper	12063	4959	42
7	2770	0.00000000000	480	2770	1722.00000000000	38433.14850600...	-431635.457654...	G4802770	4802770	G4802770	1860	Texas	Lamar	10136	7294	
8	130	0.00000000000	470	0130	1580.00000000000	1050539.465950...	-57711.3691807...	G4700130	4700130	G4700130	1860	Tennessee	Campbell	6712	6281	6
9	410	0.00000000000	290	0410	1967.00000000000	258736.1543940...	229916.0007810...	G2900410	2900410	G2900410	1860	Missouri	Chariton	12562	9672	5
10	1170	0.00000000000	220	1170	352.00000000000	560550.5408959...	-725369.708424...	G2201170	2201170	G2201170	1860	Louisiana	Washington	4708	2996	2
11	2250	0.00000000000	480	2250	1710.00000000000	54731.72934150...	-690812.986507...	G4802250	4802250	G4802250	1860	Texas	Houston	8058	5239	
12	3070	0.00000000000	480	3070	1729.00000000000	-317881.556408...	-698613.923825...	G4803070	4803070	G4803070	1860	Texas	McCulloch	0	0	

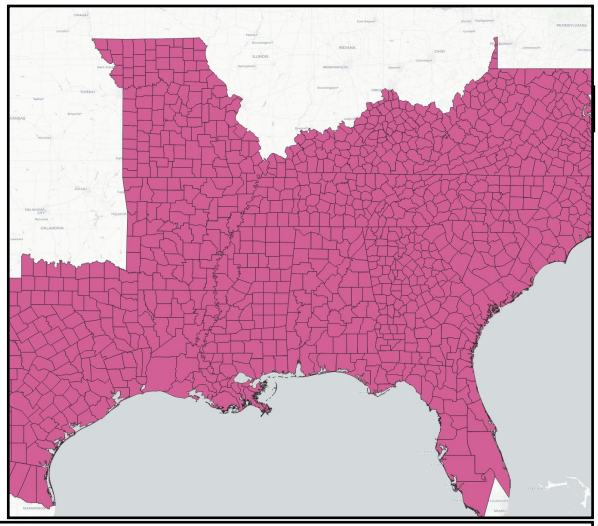
7. What are the other categories? DECADE is obvious, “NHGISNAM” refers to county (also an alternate spelling under “ICPSRNAM”). STATENAM is state name, etc. *What would you assume the other categorizes mean*?

8. Each of these later categories lists either the value of cash crops, acres dedicated to its cultivation, the number of white, non-white free, none-white slave, and total population. It is important at this point to acknowledge that this is raw data. Crop data refers to the total amount of crop production, regardless of acres in a county. Demographic data includes total numbers of people of a given race, regardless of how many people live in the county.
Why might a choropleth visualization be a poor choice, and ultimately misrepresent these raw numbers?
The alternative would be "normalized" data - or raw data as a rate, percentage, or portion of total acreage or population. Choropleths should ONLY be used with normalized data, but that doesn't mean you can't visualize raw numbers. L*ook at the alternatives to choropleth mapping above (dot density or proportional dots) when dealing with RAW data.*

Right click USSouth1860 and open properties. Look to the left hand side of the window. Select "symbology."

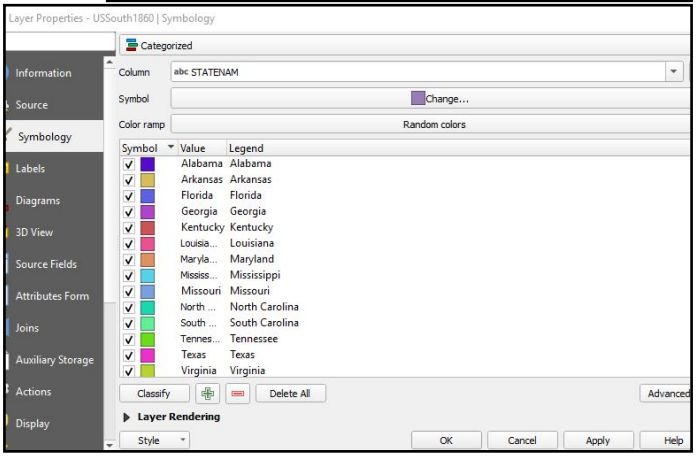
9. The default symbology option for your polygon is "single symbol." You can play around with the symbology, but the changes you make will apply equally to all of your data. Single symbol simply means that all of your data will be "styled" the same way. Go ahead and change the color of your map to see what a change to "single symbol" styling does.





10. In what contexts might it be useful to maintain this type of single symbol styling (hint - think about the maps you've produced so far…)

11. Let's try styling a different set of attributes from our attribute table. What if we want to visualize the different states and give them each a unique color? The attribute table lists states associated with each county. Rather than single symbol, this time choose "categorized.



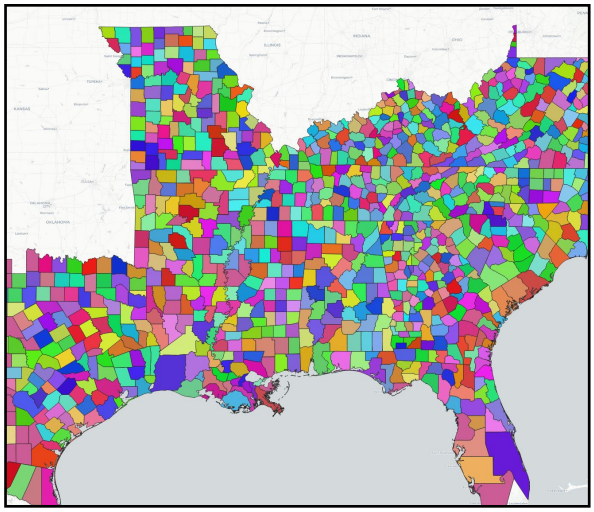
12. Click on the down arrow to the right of"single symbol" and select "categorized."

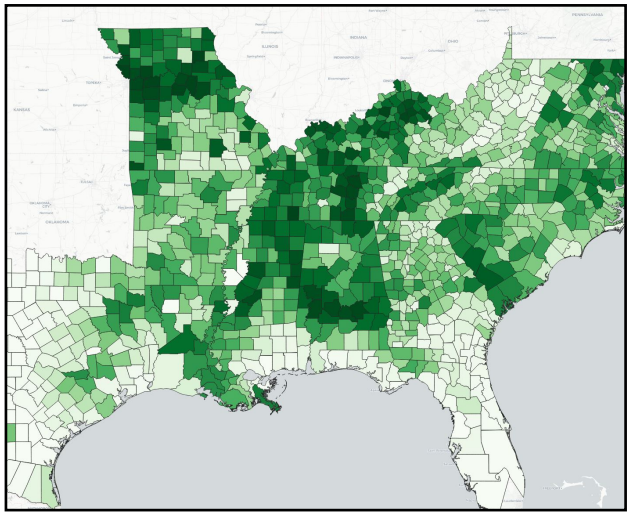
13. Categorized means that the attribute

features in the layer will be shown in different colors based on unique values in an attribute field. The column we are interested in "categorizing" is our STATENAM column. Under "column" select STATENAM. Click "classify." This populates the central box with the State name values. You'll notice that the last box is empty. You can either uncheck the box or leave it be. Click Ok.

14. Your map now shows different colors for different states. Importantly, if you click the dropdown arrow in the "layers" toolbar, you will note that your new classifications are listed as well. You can use this as your legend until you produce a map.

15. Categorizing is a relatively simple, straightforward classification scheme. *Are there any other attribute categories that would work well? Why would individual crops be poor choices for these categorized views? Why might counties be a poor choice? Which would be the best option then?*





16. If we categorized an individual crop like corn, what would we get? Initially, we would get a random assortment of colors with a huge assortment of classified values. Go ahead and try it. Just change the column you are categorizing to "corn." Click Classify and accept QGIS's warnings that the map will be virtually unreadable. Since the attribute table lists total acreage devoted to corn, every different area amount has to be accounted for. We can apply SOME order to this chaos by applying a color ramp. One could visualize the increasing amount of corn acreage somewhat effectively this way. *Why is this still a problematic approach to choropleth mapping?*

17. Two reasons.

1. It gives the impression that corn is produced in some counties much more than others (this may be true, but it may simply be because some counties are just BIGGER

COUNTIES). We would need to divide these values by total acreage to estimate *intensity* of corn production.

2. If you look at the layer tab's drop down menu, you'll note how many different values this map is showing. This is unclassified **raw** data. The legend is thus unwieldy and makes the map difficult to read and interpret.

Are there any columns in the attribute table that are already normalized?

18. PctCorn is likely our best bet. Why? Its percentage data. Large or productive counties (with relatively small amounts of corn) aren't disproportionately represented. The values have now been rounded to the nearest 10th of a percentage. In other words, they are normalized, making the differences more stark and the legend more manageable.

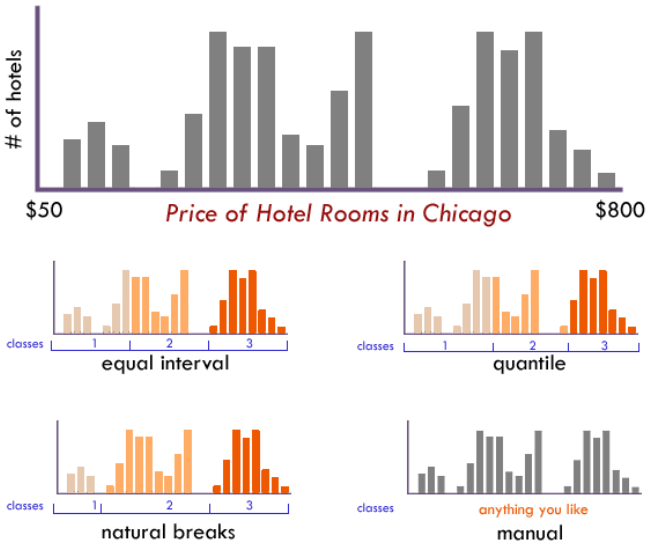
19. Does selecting a column of "normalized" data make a difference? Lets explore this by selecting a third styling option. "Graduated."

20. Graduated styling is useful when you want to do your own **classification** of data. By classifying, I mean statistically organizing (or grouping) your data. We already know that raw data is not ideal for choropleth mapping, so let's examine our last column, PctCorn. This gives us the percentage of cash crop acreage dedicated to corn. Since its a percentage, it is ideal for choropleths.

21. Return to the properties dialogue for USSouth1860 and select "graduated."

****A NOTE ON CLASSIFICATION****

In general, you want to classify data on choropleth maps to give an accurate representation of the theme you are describing. This means the type of classification you choose depends on your data, and depends on the



The form of this histogram suggests that 3 or 4 data classes seem most appropriate. Lacking any other insight, the "dips/gaps" suggest natural places to break the data.

EQUAL INTERVAL divides the data into equal size classes (e.g., 0-10, 10-20, 20-30, etc.) and works best on data that is generally spread across the entire range. **CAUTION:** Avoid equal interval if your data are skewed to one end or if you have one or two really large outlier values. Outliers in that case will likely produce empty classes, wasting perfectly good classes with no observations in them. Since the hotel data above doesn't have really large outliers, this is a data distribution that works well with equal interval.

QUANTILES will create attractive maps that place an equal number of observations in each class: If you have 30 counties and 6 data classes, you'll have 5 counties in each class. The problem with quantiles is that you can end-up with classes that have very different numerical ranges (e.g., 1-4, 4-9, 9-250...the last class is huge). Quantile can also separate locations with very similar rates and group together places that have very different rates, which is very undesirable, so use the histogram to see if this is happening. **CAUTION:** In the hotel room example above, the quantile produced a questionable class break by lumping a portion of the third cluster back into class 2, despite it being much closer (numerically) to the other observations in class

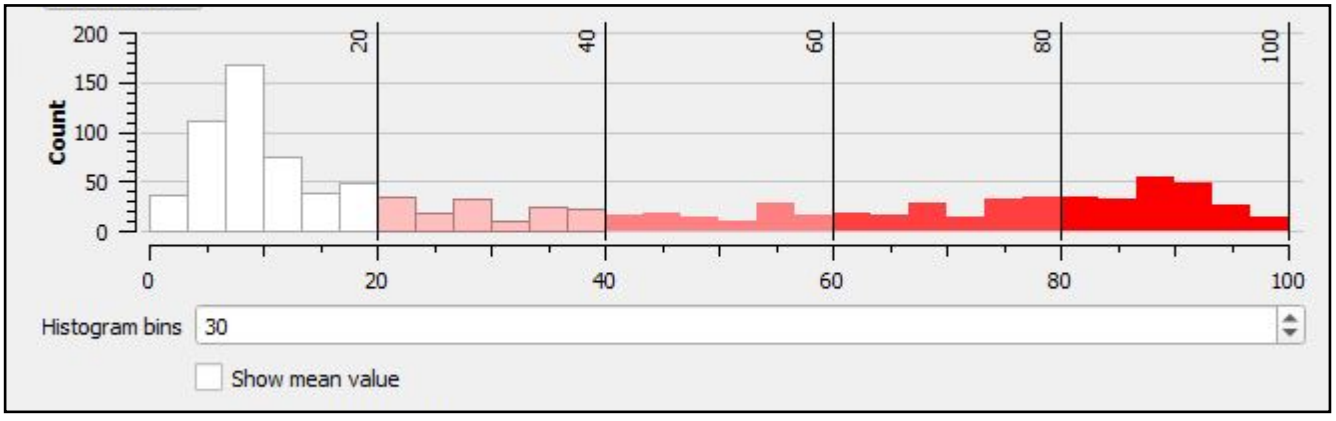
NATURAL BREAKS is a kind of "optimal" classification scheme that finds class breaks that (for a given number of classes) will minimize within-class variance and maximize between-class differences. One drawback of this approach is each dataset generates a unique classification solution, and if you need to make comparison across maps, such as in an atlas or a series (e.g., one map each for 1980, 1990, 2000) you might want to use a single scheme that can be applied across all of the maps.

purpose of the map. That said, in general, you want to "group" similar phenomena together and maximize the difference between different phenomena. This will make your map more legible.

There are a number of different explanations about how and why to choose the classification you do. One of my favorites comes from indiemapper (above).

22. So how can we use this information in QGIS? Under column, select PctCorn. Select a suitable color ramp to show gradual change in relative intensity of corn production. Change the number of classes to 5. Generally, you don't want to have a choropleth with less than 3 or more than 7 classes.

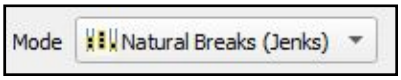
23. The default classification scheme is equal interval. Notice how the values are distributed so that an equal amount of values are distributed to each category? Is this the appropriate classification scheme? In order to find out, let's look at the histogram.



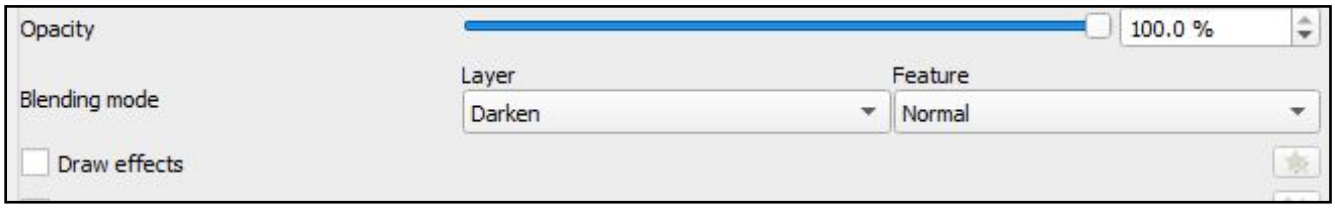
23. Click the histogram tab and click "load values" on the next tab. Equal interval captures much of the first "bin" of information, but it tends to split up the outliers.

24. You can also select "show mean value" or "show standard deviation." This may give you a sense of how appropriate the standard deviation model may be. Go ahead and switch your classification scheme to "standard deviation. Does this represent your data more effectively?

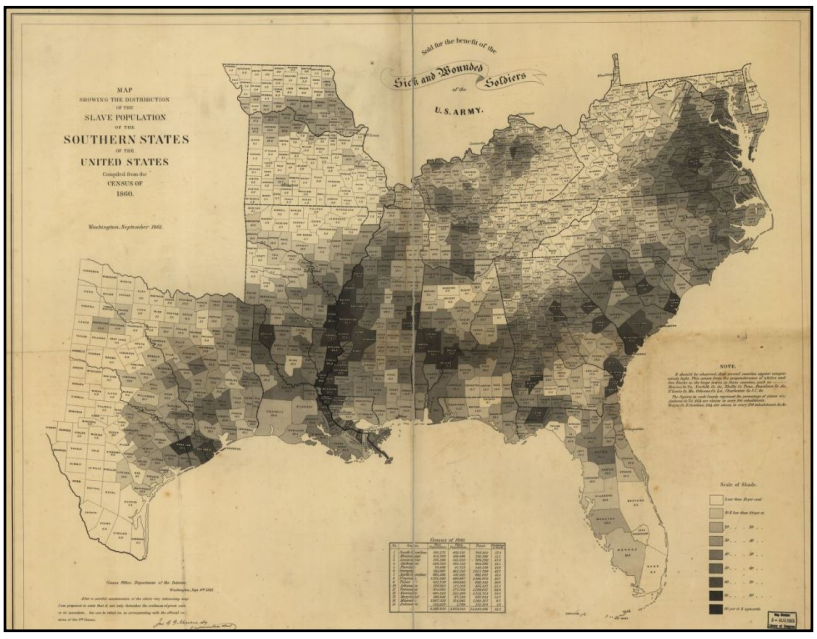
25. Now experiment with Natural Breaks (Jenks), Quantile (number of values in each class are the same), and Pretty Breaks ("pretty" meaning the values are round numbers). You can change the data classification options by clicking the “classes” tab. Of these options, which do you think best displays your data? Why? *Refer back to the "Note on Classification"*



26. Feel free to explore the "layer rendering" options at the bottom of the Layer Properties window. If you're using a base map, you can "blend" your layers together (rather than simply changing transparency.



27. Now that we've explored the potential for using census data to produce choropleth maps according to various classifications, let's return to our subject for the week. Racialized territories have often been visualized since the mid 19th century using census statistics.

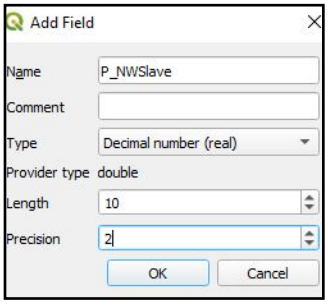


Hergesheimer, "Map showing the distribution of the slave population of the southern states of the United States. Compiled from the census of 1860" (1861)

28. We're going to recreate this map. We have a problem, though. Look at the legend of this historical map. This map does NOT show raw data. Its a choropleth map that shows normalized data. In this case, the percent of total population that were enslaved in 1860. Look at your attribute table.

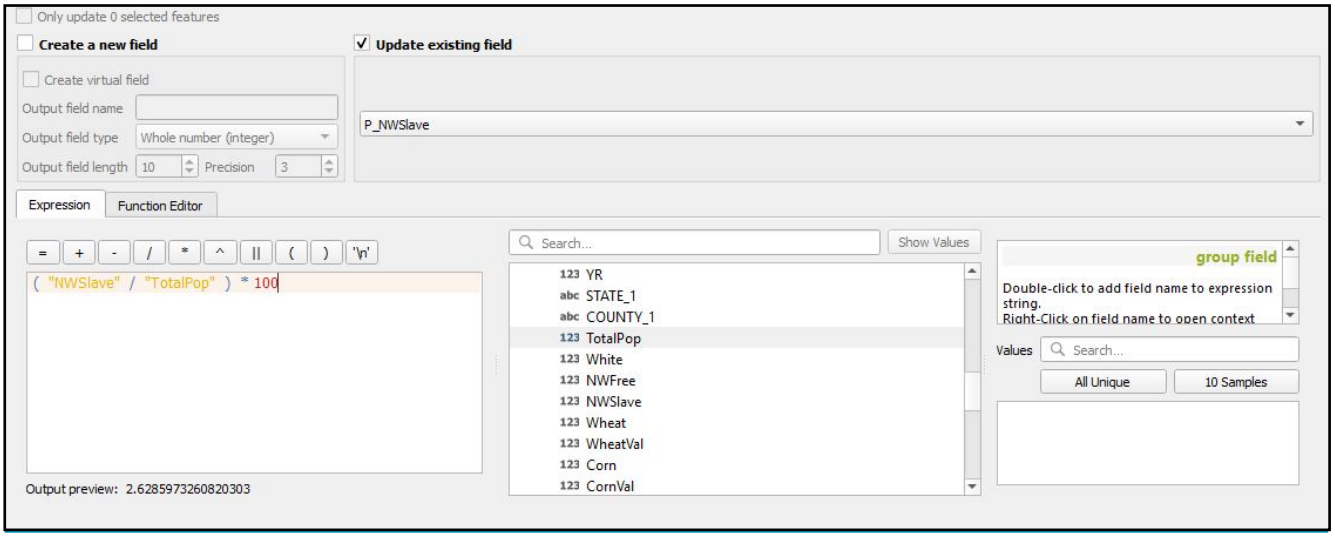
Do you see a column that shows normalized counts of people? No. We need to make one.

29. To do this, we will have to edit our attribute table. Open the attribute table. Find the pencil "edit" tool in the upper left of the attribute table window. Next, select "new field" on the same toolbar. This will create a new column in your shapefile dataset. Here, you need to give your new column a name "P_NWSlave" (percent non white slave), ignore comment, make sure the type is "Decimal number (real)." This is important because it will determine whether your results will be integers (i.e whole numbers), text (like state names), or date. Length refers to the total number of digits, precision refers to the total number of decimals. Make length 10, precision 2. When you click ok, this new column will appear on the far right of the attribute table.

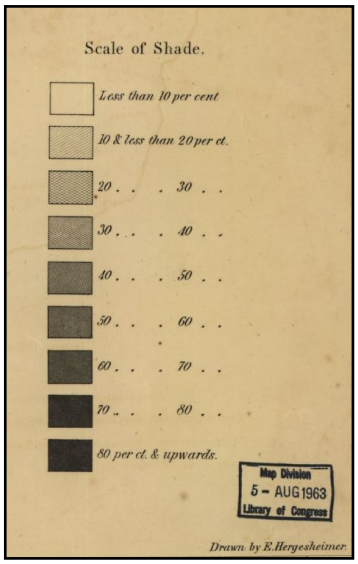


29. Next we need to calculate the percentage of county population that was enslaved. To do this, click the field calculator icon on the upper toolbar of the attribute table.

30. The field calculator in QGIS is a very robust multitool, but today we will be focusing on a just a few key features. First, make sure to uncheck the box "create new field." We have already done this when we made the P_NWSlave column, though alternatively, that could be done in this window. The "update existing field" box should automatically be checked. Next, make sure that "P_NWSlave" is the field you are updating using the dropdown menu below. Look to the middle panel. You will see a number of different options for inputs into your "calculator." We are interested in the percent of total population that were non-white and enslaved. To calculate this, click on "Fields and Values" and you will see a dropdown menu of all of your field options (these are your attribute table columns). Create an equation that divides non-white enslaved people by total population. (see below). The equation divides

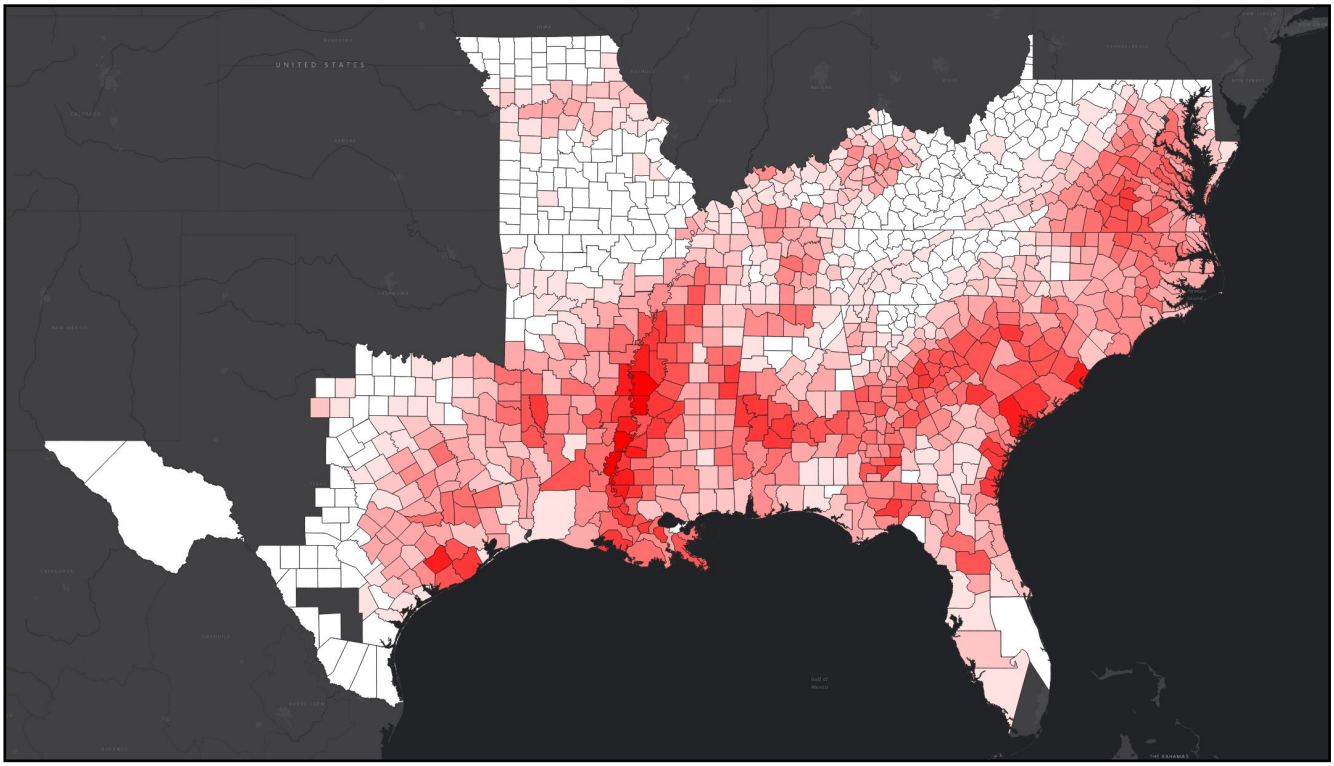


non-white slave by the total population. Multiplying this result by 100 results in a percentage. Click OK. Save your edits by clicking the pencil icon again.



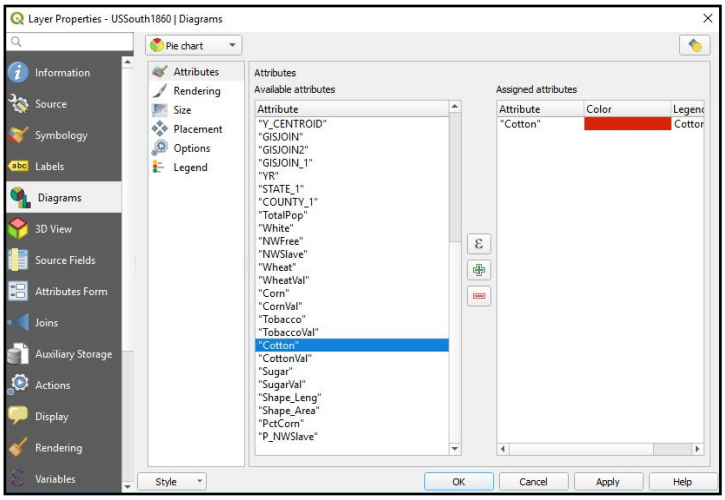
31. Your new column now has figures that: 1. are divided according to enumeration units (counties) 2. show ratios/proportions/rates etc. (data normalized over space) 3. measures a mapped theme anywhere on its surface. Go ahead and choose an appropriate classification for this new column and build a new choropleth map that shows the "distribution of the slave population of the southern states of the United States." What classification scheme did you choose? (remember to look at the historic map for hints). Make sure you number of classes matches the number on Hergesheimer's map.

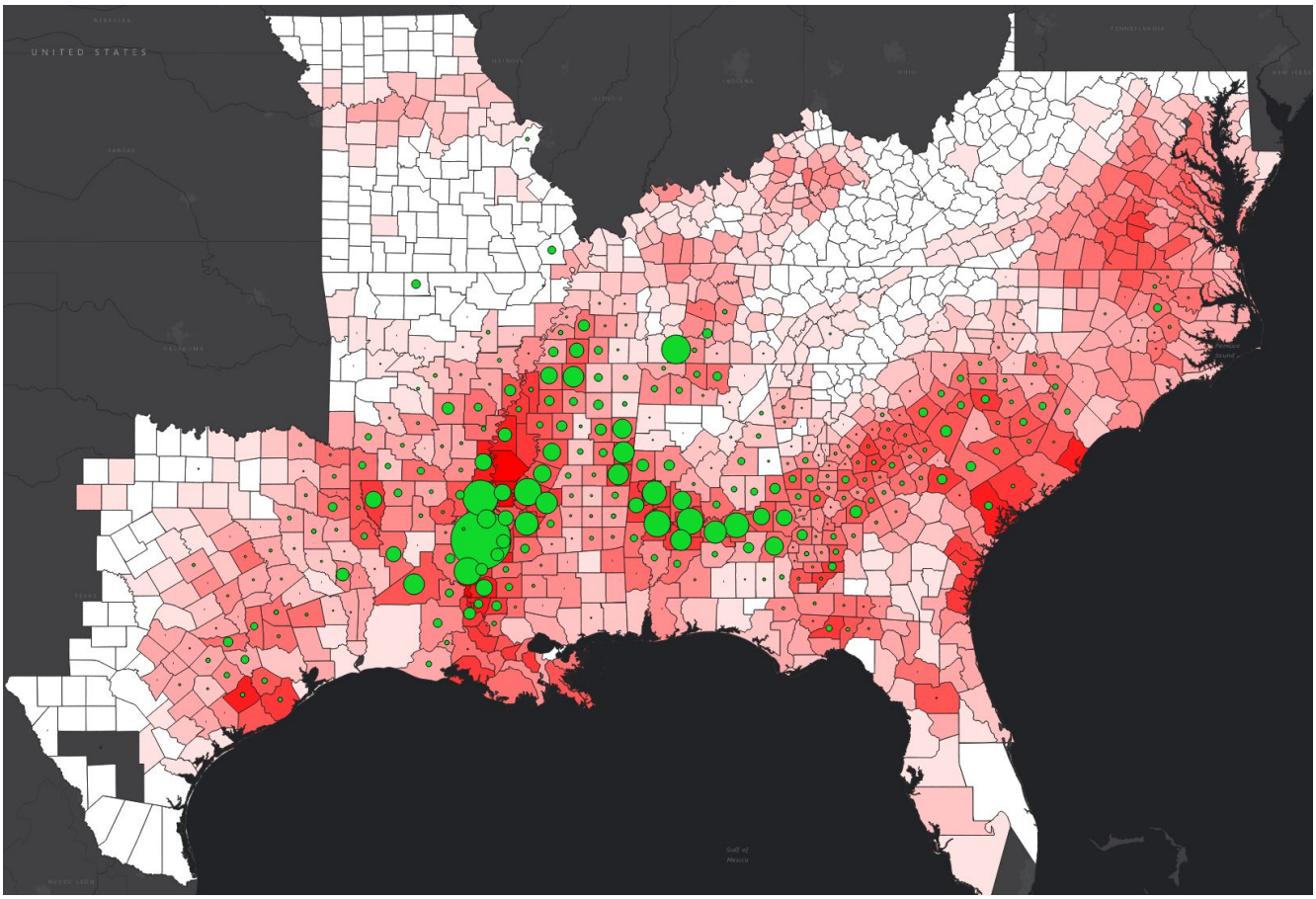
32. We've now recreated an historical map. Screenshot and save this. Can it help us answer an historical question? For instance, what type of cash crops were most tightly correlated to the areas with highest slave populations. In other words, what kinds of plantations drove the slave system? Let's start with cotton.



33. We could create more choropleth maps by dividing total bushels of cotton by acreage. I want to show you another way can visualize this data, however. Sometimes, you WANT to visualize raw, unnormalized data. When you do, use proportional dots or dot density (see pg 2) We will use proportional dots.

34. Click on your USSouth1860 layer to open its properties again. This time, click "diagrams" on the left toolbar. Click "no diagrams" on the top and change it to "pie chart" from the drop down menu. Below that, click "attributes." This is selecting the column you want to visualize with proportional dots. Then double click "Cotton" from the available attributes", this adds it to your assigned attributes. This also selects the color your dots will appear as.





35. Next, click "size" on the left menu. Select "scaled size" and make sure it is scaled according to the "attribute" cotton. For maximum value, click "find".
Size should equal 20. click apply. Caution, this may take some time to load.
36. Pretty clear spatial relationship right? Screenshot and save this. Now follow the same steps to overlay proportional dots for the other cash crops. Post the maps and answer the questions below and submit it to the blog.

Don't forget to save your project!

QUESTIONS FOR BLOG:

1. How would you describe the spatial relationship between slavery and counties with high production of:
    a. Corn
    b. Wheat
    c. Tobacco
    d. Cotton
    8. Sugar

2. How might you determine the relative importance of these crops for the confederacy?