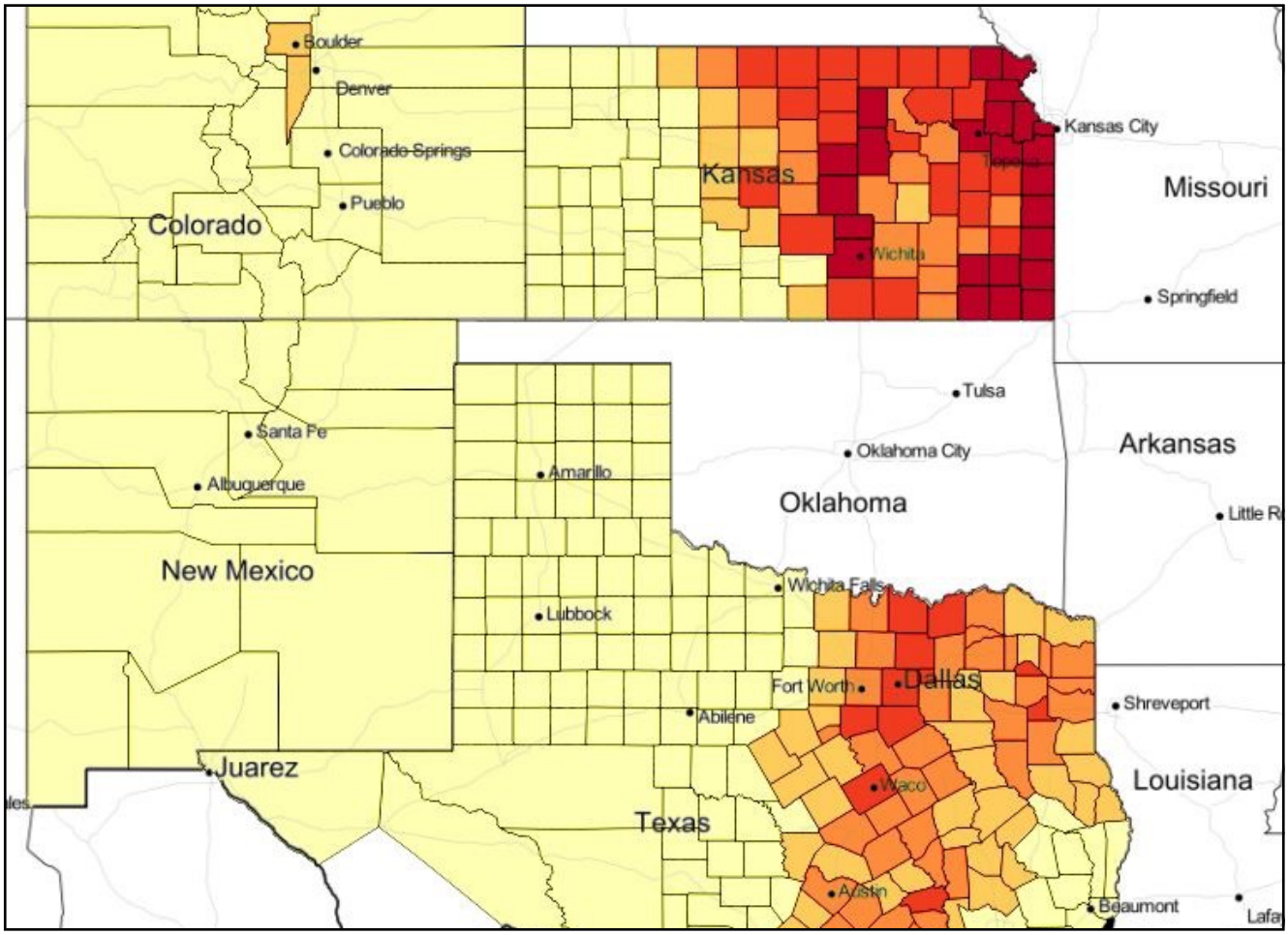
Census Data & Choropleths

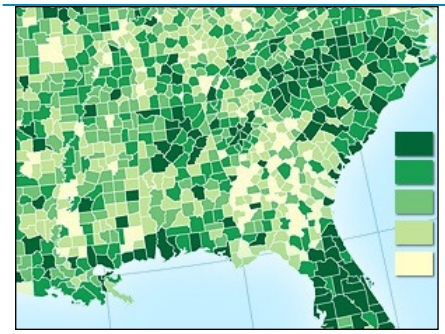
Mapping Rural Land Use of the Great Plains

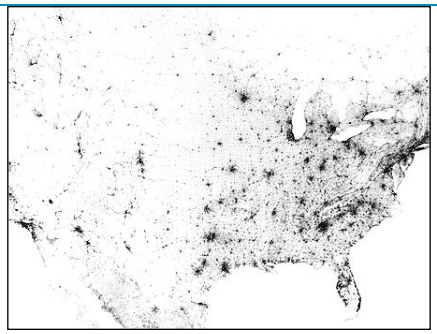


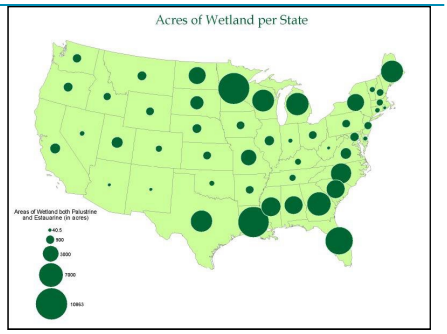
Introduction

Historical census data are among the richest sources of social and environmental information available for HGIS. 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 southern portion of the great plains.

Choropleth Maps | Dot Density Maps | Graduated Symbol Maps







Choropleth Maps

Choropleth maps are among the most popular maps employed today (and are also some of 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 divided 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 Geoff Cunfer's analysis of land use on the southern Plains. This dataset is associated with the ESRI publication *Placing History* (2008), which was itself drawn from the Inter-university Consortium for Political and Social Research (ICPSR) "Great Plains Population and Environment Data Series."

1. Create a new folder entitled "Week 7" on your computer. Go to the shared Google Drive folder for Week 7. Download and save the files to your computer. If you haven't already, you should bookmark the site. https://drive.google.com/drive/folders/0B5F--ticb5UAeHEyeVMxb0tZaHc?usp=sharing

2. Open QGIS. The county boundaries are likely familiar, but if you'd like, feel free to load a web base layer. Import the 1880.shp shapefile.

3. The shapefile includes county level data for 3 states and one territory. Use the "identify features tool to determine which state is lacking data. Why do you suppose this is?

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 1880 and open the table. What do you see?

	GISJOIN	NHGISNAM	STATENAM	UNFIPS	NAME	YEAR	AREA	CORN	WHEAT	HAY	OATS	COTTON	BARLEY	RYE	PctCorn	PctCorn2
0	2000650	Graham	Kansas	20065	GRAHAM	1880	574164	4925	1013	2370	60	0	14	63	0.01	0
1	2000670	Grant	Kansas	20067	GRANT	1880	365360	0	0	0	0	0	0	0	0.00	0
2	2000710	Greeley	Kansas	20071	GREELEY	1880	513217	0	0	0	0	0	0	0	0.00	0
3	2000750	Hamilton	Kansas	20075	HAMILTON	1880	639189	0	0	0	0	0	0	0	0.00	0
4	2000770	Harper	Kansas	20077	HARPER	1880	500527	10565	3607	2187	1264	0	0	48	0.02	0
5	2000790	Harvey	Kansas	20079	HARVEY	1880	337344	43794	37100	15738	13422	0	314	265	0.13	1
6	2000830	Hodgeman	Kansas	20083	HODGEMAN	1880	554371	1458	4782	1995	769	0	392	73	0.00	0
7	2000850	Jackson	Kansas	20085	JACKSON	1880	412838	50614	12517	29771	4947	0	145	729	0.12	1

6. Two important categories that may not be immediately recognizable to you are the first column "GISJOIN" and the 4th column "UNFIPS." Although it isn't apparent from the shapefile shared in the Google Drive, census data often does not arrive tied to county or state shapefiles (its just a spreadsheet). Census data in tabular formats need to be "**joined**" (sometimes called concatenated) to georeferenced administrative polygons (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. "UNFIPS" is another join field, this one assigned by the US federal government and employed in modern census data (since 1970). This attribute table, in other words, give you two options to join data. (We will work more with joins later in the semester)

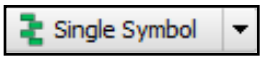
7. What are the other categories? "Name" refers to county (also an alternate spelling under "NHGISNAM"). STATENAM is state name, etc. Year is the year of the ag census and AREA is in km squared. What would you assume the other categorizes mean?

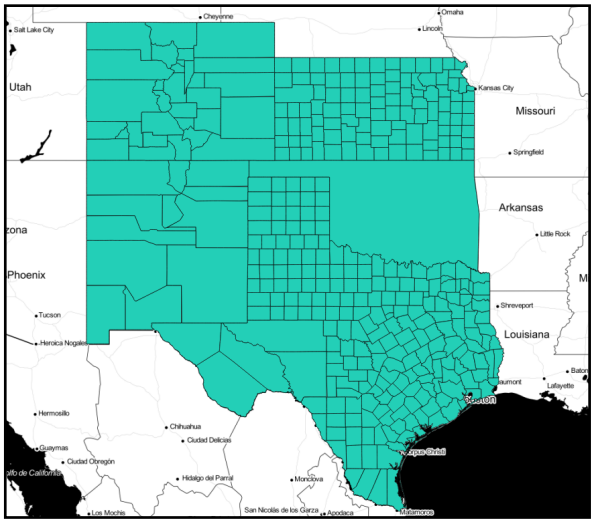
8. Each of these categories lists the area devoted to a given crop in that county in 1880. How can you determine this based on the data in the table? It is important at this point to acknowledge that this is raw data. Most crop data are **NOT** rates or percentages or a portion

of total acreage. In other words, they are not normalized data. There are two exceptions and we will deal with them shortly. How should we visualize raw data like this?

Right click 1880 and open properties. Look to the left hand side of the window. Select "style."

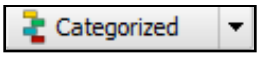
9. The default symbology option for your polygon is "single symbol." You can play around with the symbology, but the changes you make will apply equally to all of your data. Single symbol simply means that all of your data will be "styled" the same way. Go ahead and change the color of your map to see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





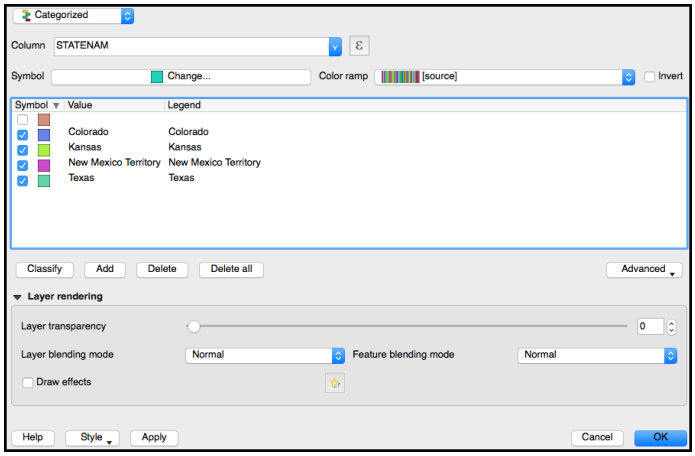
10. In what contexts might it be useful to maintain this type of single symbol styling (hint - think about the maps you’ve produced so far…)

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

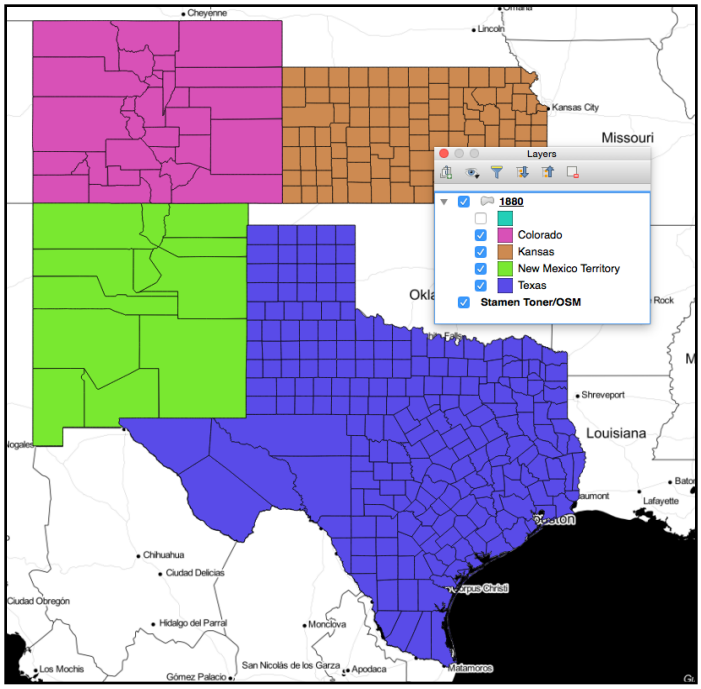


12. Click on the down arrow to the right of"single symbol" and select "categorized."

13. Categorized means that the attribute features in the layer will be shown in different shades of a color based on unique values in an attribute field. The column we are interested in “categorizing” is our STATENAM column. Under “column” select STATENAM. Click the “classify.” This populates the central box with the State name values. You’ll notice that the first box is empty. This is the area in what is now Oklahoma. You can either uncheck the box, leave it be, or type in Oklahoma (for the legend if you make one. Click Ok.

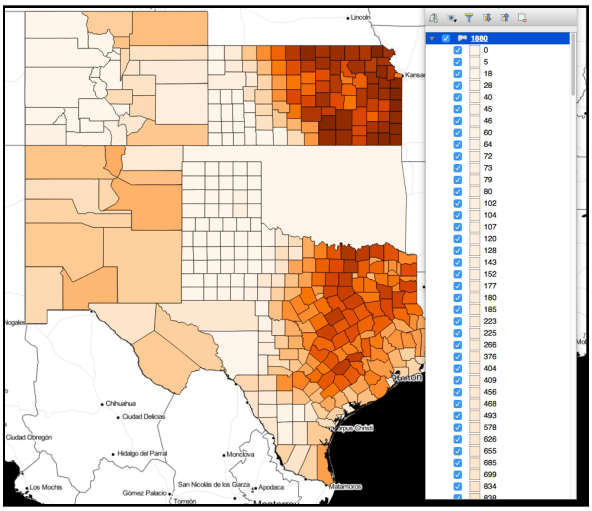


14. Your map now shows different colors for different states. Importantly, if you click the dropdown arrow in the "layers" toolbar, you will note that your new classifications are listed as well. You can use this as your legend until you produce a map.



15. This is a relatively simple, straightforward classification. Are there any other attribute categories that would work well? Why would individual crops be poor choices for these categorized views? Why might PctCorn 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. 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. It gives the false impression that a lot of corn is produced in the western counties (there isn't, they are just BIGGER COUNTIES). 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 unwieldy and the map doesn't give us a real sense of the relative intensity of corn production between counties. What's left?

18. PctCorn2 is likely our best bet. Why? Its percentage data. Large 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. Perhaps the result is less dramatic,

but one could argue that it is much more accurate. After all, only one county in Kansas has 40% of its land area dedicated to corn production, but that county could never be identified in the former map.

18. One doesn't need to rely on an attribute table having a ready made, classified data category like PctCorn2. QGIS will do this for you. Lets explore a third styling option. "Graduated."

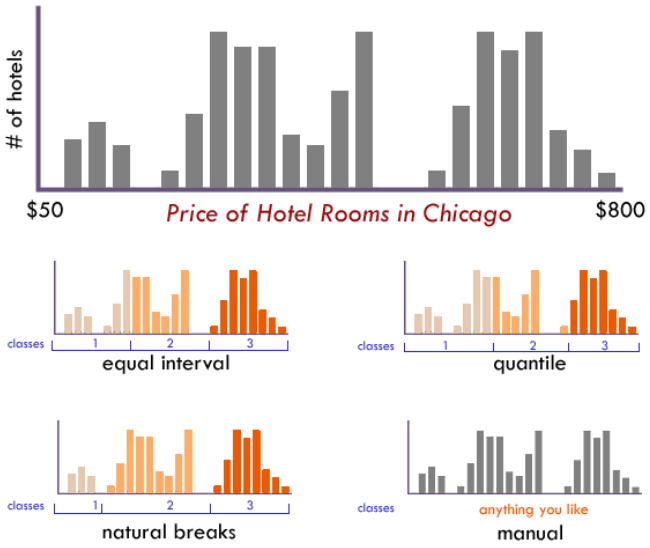
19. Graduated styling is useful when you want to do your own **classification** of data. By classifying, I mean statistically organizing (or grouping) your data. We already know that raw data is not ideal for choropleth mapping, so lets examine our last column, PctCorn. This gives us the amount of acreage dedicated to corn. It is normalized to area, so it is ideal for choropleths.

20. Return to the properties dialogue for 1880 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 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 (right).



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.

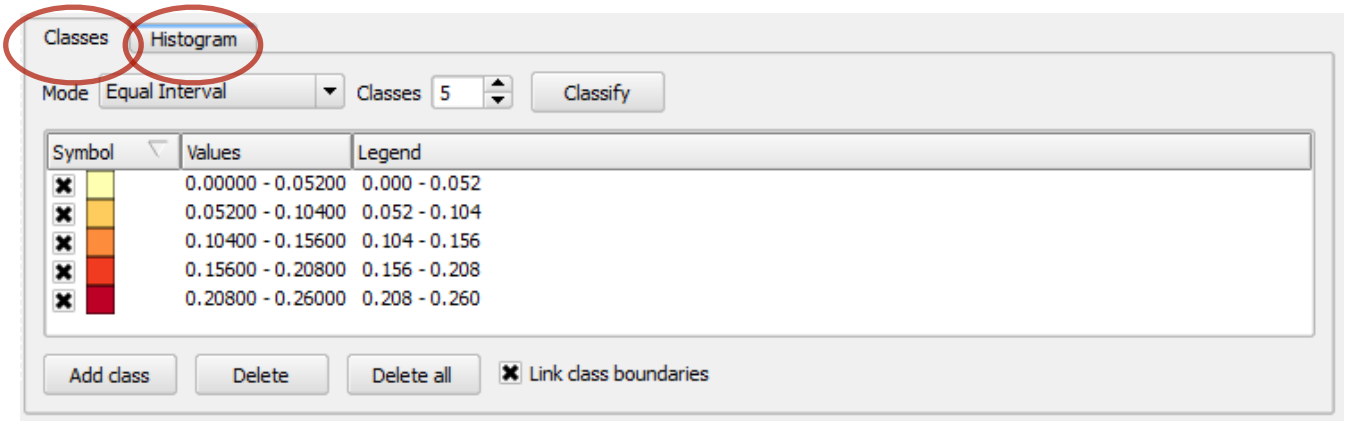
EQUAL INTERVAL divides the data into equal size classes (e.g., 0-10, 10-20, 20-30, etc.) and works best on data that is generally spread across the entire range. **CAUTION:** Avoid equal interval if your data are skewed to one end or if you have one or two really large outlier values. Outliers in that 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

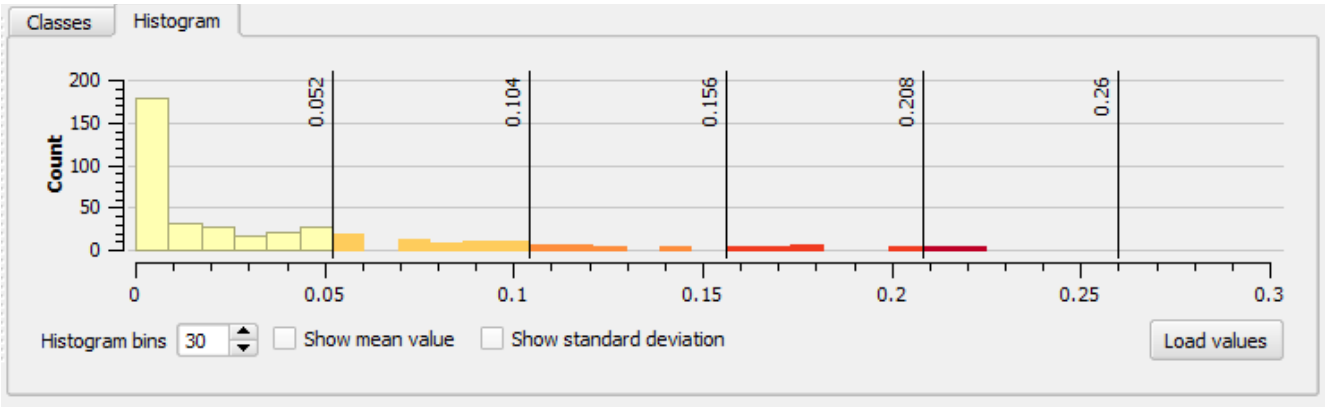
NATURAL BREAKS is a kind of "optimal" classification scheme that finds class breaks that (for a given number of classes) will minimize within-class variance and maximize between-class differences. One drawback of this approach is each dataset generates a unique classification solution, and if you need 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.

21. So how can we use this information in QGIS?
Under column, select PctCorn. Select a suitable color ramp to show gradual change in 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.

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.



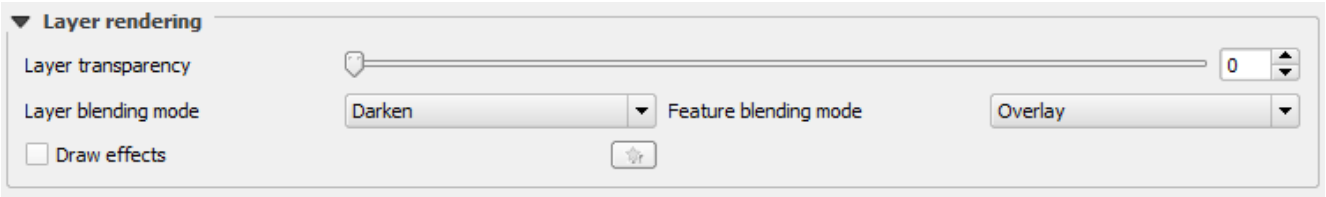
22. Click the histogram tab and click "load values" on the next tab. Equal interval captures much of the first "bin" of information, but it tends to split up the outliers.



23. 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?

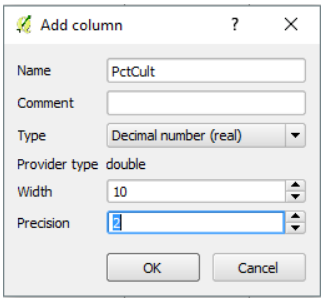
24. Now experiment with Natural Breaks (Jenks), Quantile (number of values in each class are the same), and Pretty Breaks ("pretty" meaning the values are round numbers). You can change the data classification options by clicking the "classes" tab. Of these options, which do you think best displays your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

25. 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.



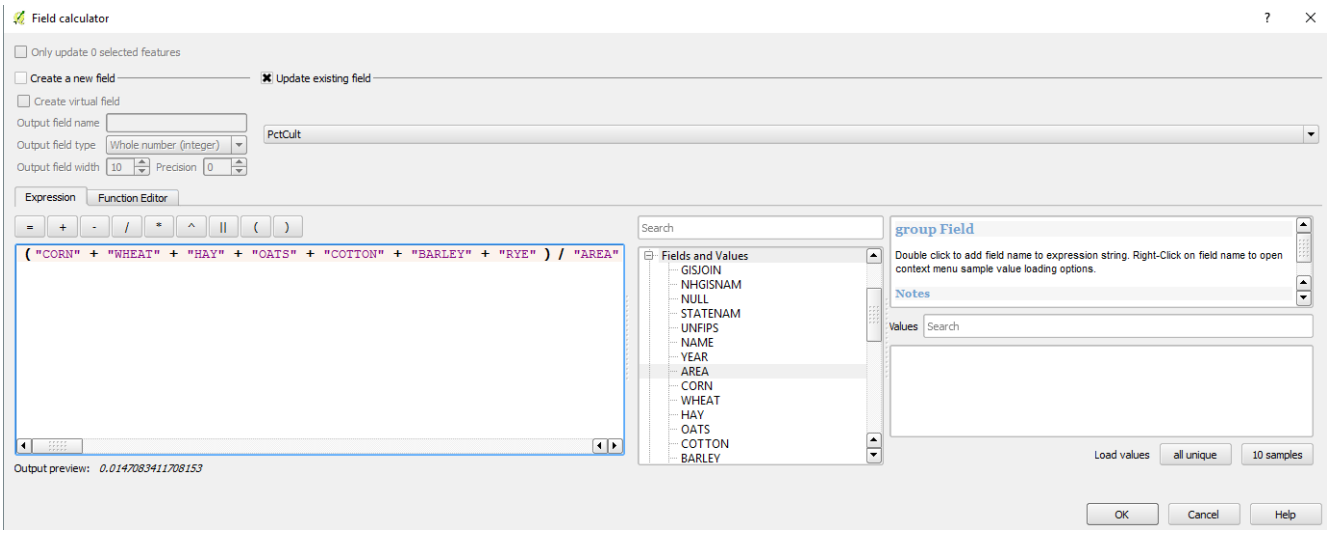
26. Now that we've explored the potential for using census data to produce choropleth maps according to various classifications, let's return to Cunfer's primary concern in his *On the Great Plains:* the amount of land that is (or isn't) cultivated. If you look at the attribute table for your 1880 layer again, there is no column that lists the percentage of land in cultivation or percentage of land that was unplowed (its inverse). We DO have a column for total acreage and we do have columns that lists acreage of individual crops. We need to create a new column that normalizes total cultivated acreage to total area. (i.e., the sum of all crop acreage = total cultivated acreage/area). Its important to note that we are making two big assumptions here. 1. that there were no other land uses that would have "plowed up" land area (others crops/roads/cities) 2. that our units are the same (Area = km2 and Corn = hectares or mi2, this wouldn't work). We could check this by returning to the metadata.

27. We want to make a column that will allow us to map the percentage of cultivated land. To do this, we will have to edit our attribute table. Find the pencil "edit" tool in the upper left of the attribute table window. Next, select "new field" on the same toolbar. Here, you need to give your new column a name "PctCult" (percent cultivated), 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.



28. Next we need to calculate the percentage of cultivated acreage. To do this, click the field calculator icon on the upper toolbar.

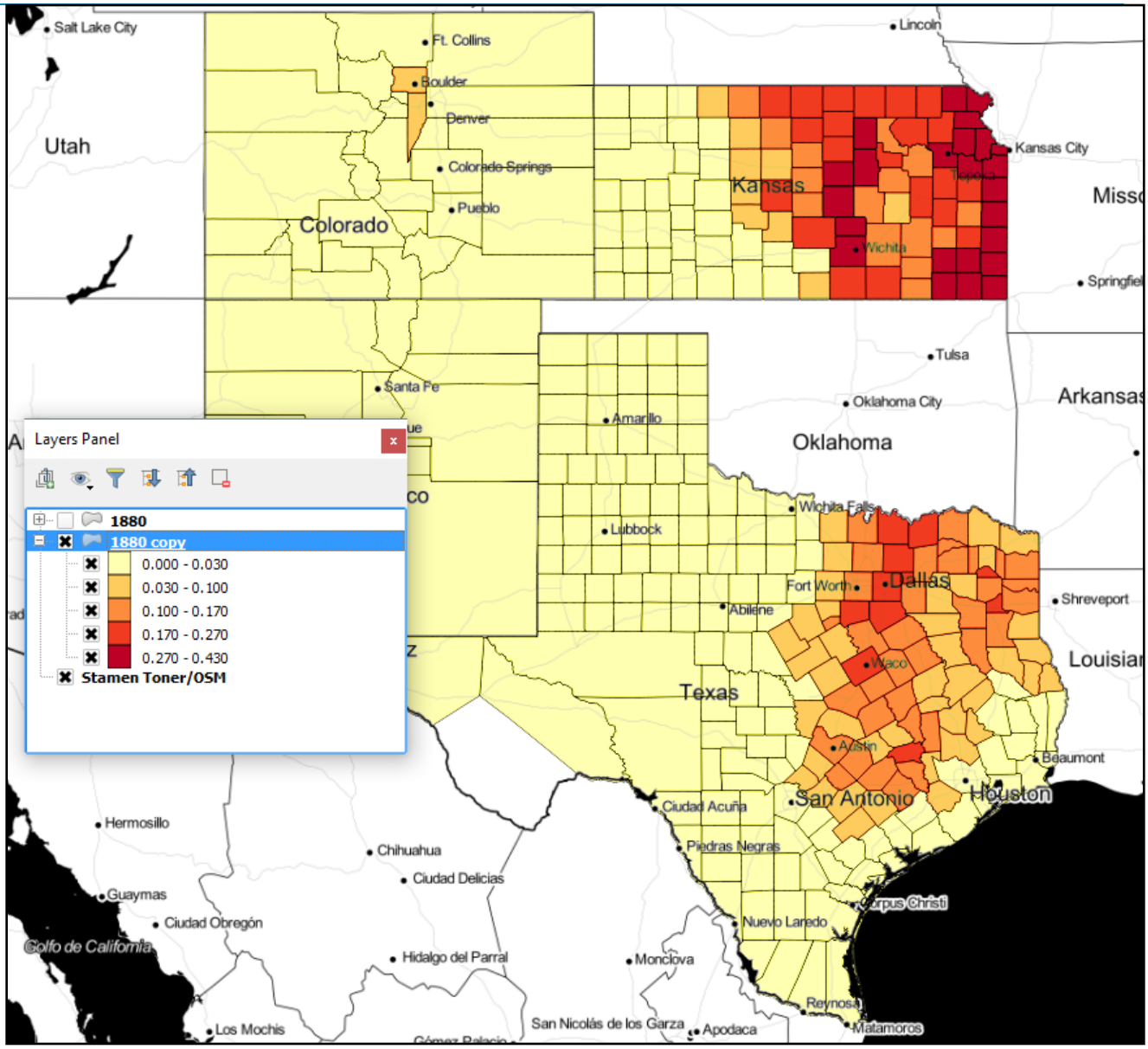
29. The field calculator in QGIS is a very robust multitool, but today we will be focusing on a few key features. First, make sure to uncheck the box "create new field." We have already done this when we made the PctCult column, though alternatively, that could be done in this window. The "update existing field" box should automatically be checked. Next, make sure that "PctCult" 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 adding the values of all of our cultivated acreage together. Click on "Fields and Values and you will see a dropdown menu of all of your field options. Create an equation that adds all of your cultivated acreage together and then divide that number by your AREA field. (see below). Click OK. Save your edits.



30. You now have another column that: 1. Is divided according to enumeration units 2. shows 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 extent of cultivated lands in the southern plains.

31. To show change over time, add a new column to two other years (I'm adding several options). Create a similar choropleth map for both.

Post the three maps to the blog and note any key changes you observe between the years you've identified.


Salt Lake City
Utah
Ft. Collins
Boulder
Denver
Colorado Springs
Pueblo
Colorado
Lincoln
Kansas City
Kansas
Wichita
Tulsa
Oklahoma City
Oklahoma
Arkansas
Santa Fe
Amarillo
Lubbock
Wichita Falls
Fort Worth
Dallas
Abilene
Shreveport
Waco
Texas
Austin
Beaumont
San Antonio
Houston
Ciudad Acuña
Piedras Negras
Nuevo Laredo
Corpus Christi
Reynosa
Matamoros
Hermosillo
Chihuahua
Ciudad Delicias
Guaymas
Ciudad Obregón
Golfo de California
Hidalgo del Parral
Monclova
San Nicolás de los Garza
Apodaca
Los Mochis
Layers Panel
1880
1880 copy
0.000 - 0.030
0.030 - 0.100
0.100 - 0.170
0.170 - 0.270
0.270 - 0.430
Stamen Toner/OSM