Lesson 5: Data for a Map

## INTRODUCTION

This lesson teaches you about the elements related to data for a map. You will learn about the common data types and formats used in map creation. The lesson will help you gain an understanding of normalized data and descriptive statistics. You will also learn about data classification and the various classification requirements and methods used when making a map.

## LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Identify data elements of a map.
2. Compare data classification methods.
3. Create maps using different data classifications.
4. Manage different types of data (vector, raster, and imagery), databases, and metadata to create different types of maps and map scales.

## LEARNING SEQUENCE

|  |  |
| --- | --- |
|  | |
| Required Reading | Read the following:  Data for a Map   * Data for a Mapping Project * Common Data Types/Formats * Descriptive Statistics * Data Classification * Classification Requirements * Classification Methods |
| Assignments | Complete the following assignments:   * Quiz: Data for a Map * Lab: Data Classification |

## INSTRUCTION

## Data for a Mapping Project

### Why is Data Important?

## So why is data important? For any GIS and mapping project you will spend roughly 50% of your time acquiring, converting, projecting, modifying, and managing your data to get prepared for use. The other 50% of your time will be used for actually doing the project. Therefore, you should plan accordingly to spend a significant amount of time simply finding and preparing data for usage.

Data is not always easy to find or in the format in which you need it. In almost all cases you will need to perform some data transformation, selection, generalization, or other data preparation operations before you can use the data for your GIS project or map. Quite often knowing someone will help you get the data quicker. While there are many GIS data clearinghouses and websites on the Internet, they are quite dispersed and sometimes difficult to find and navigate. Additionally some of the data you desire may not be available online so you should use your social network to find data when appropriate.

### Digital Data

## There is digital data available almost everywhere. Even though the data is available in many places sometimes it is still difficult to find. In either case digital data is the least expensive source of spatial data because the reproduction costs are close to zero. Digital data are often developed by governments because the data helps provide basic public services and adds value to the comments.

## There are many available data formats that a GIS can ingest, however, typically when acquiring data from sources, the data is typically going to be in one of these for formats: raster, vector, ASCII, or an interchange format. The raster and vector data types are ready for use in a GIS that any conversion in most cases. The ASCII and interchange formats typically require some data transformation and conversion in order for it to be ready for use in a GIS.

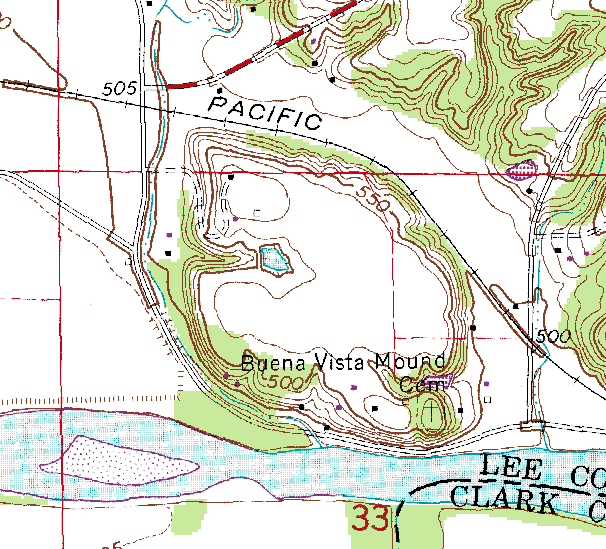
## Common Data Types/Formats

### Digital Raster Graphic (DRG)

A digital raster graphic, more commonly referred to as a DRG, is a geo-referenced raster from scanned USGS maps. The DRGs come in three scales of 1 to 24,000, 1 to 100,000, and 1 to 250,000. The USGS maps that are scanned into DRGs are typically older data but reasonably reliable. Therefore a good use for a DRG is for a quick check for consistency and new data collection. For instance if you receive a road data set and you want to quickly make sure that it is geo-referenced to the correct place you can overlay on top of the DRG to see if they line up.

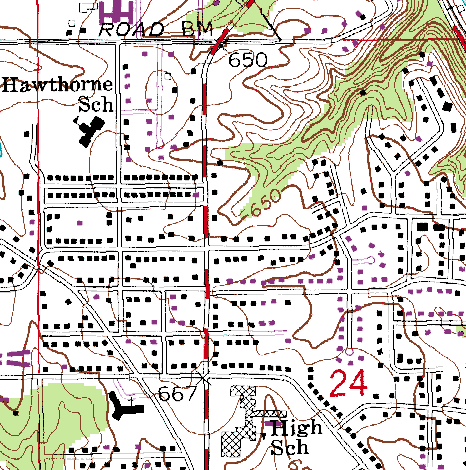
#### General Reference

This is a portion of a USGS map. This is considered a general reference map and displays things such as buildings, contour lines, rivers, political boundaries, roads, railroads, among other features.



#### Populated Area

This is a portion a DRG of a populated area. The squares represent buildings. Notice that some of the squares are black and some of the squares are purple. The black squares and symbols are from the original publishing of the map. The purple squares in figures are from a subsequent update of the map.



### Digital Line Graph (DLG)

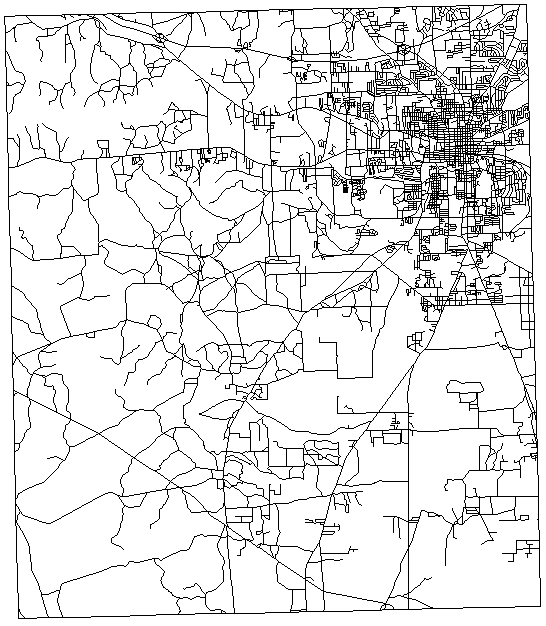
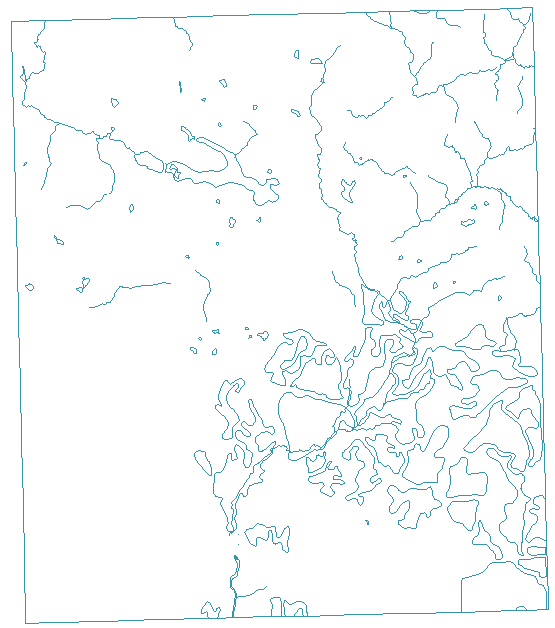
A digital line graph, commonly referred to as a DLG, is the vector representation of most features on the USGS national series of maps which we just saw as DRGs. DLG’s come in three scales: 1 to 24,000, 1 to 100,000, 1 to 2 million. The DLG’s are separated out by theme.

So for example you will have a DLG shapefile for boundaries, a DLG shapefile for roads, hypsography, and hydrography. Like the DRGs, DLG shapefiles are useful for quick check for consistency and data collection. Additionally DLG’s cover the entire United States so they can provide a reasonably good national data set.

#### Road and Hydrography

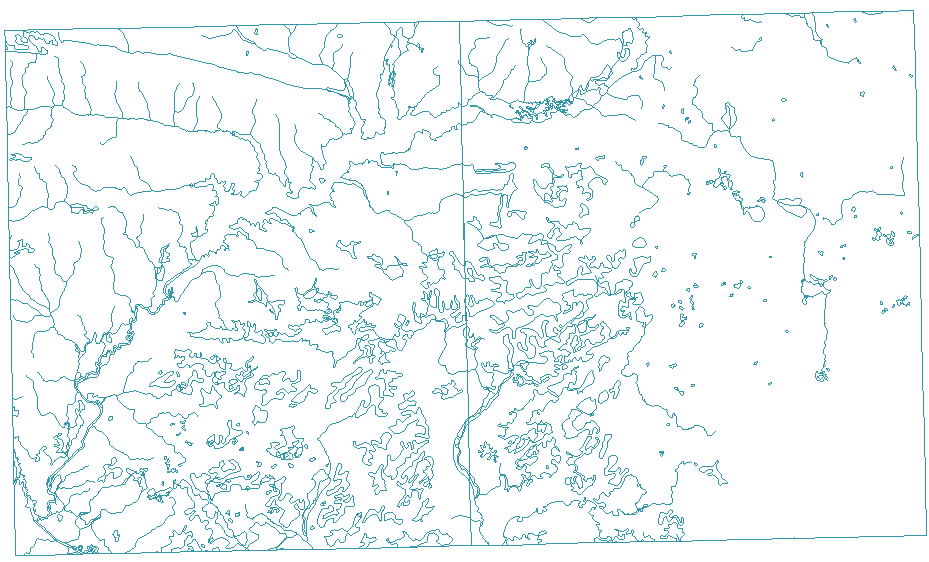
These images are examples of a road and hydrography data set from a DLG.

**Roads Hydrography**

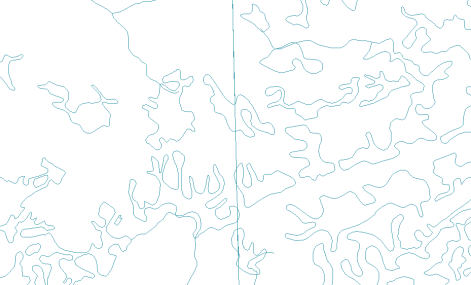
### Edge Mapping

One of the nice things about the DLG data set is that they perform edge matching. As the DLG’s were digitized from DRGs, there are cases where features exist across to different maps. When the DLG’s were created the features were matched up across the two different maps.



### Vertical Line Edge Mapping

In this image the vertical line in the center of the image represents the separation of the two maps. The polygons represent lakes. If you take a look at the lakes that cross the vertical line you can see that the features connect to each other and allow for continuity across multiple DLG data sets.



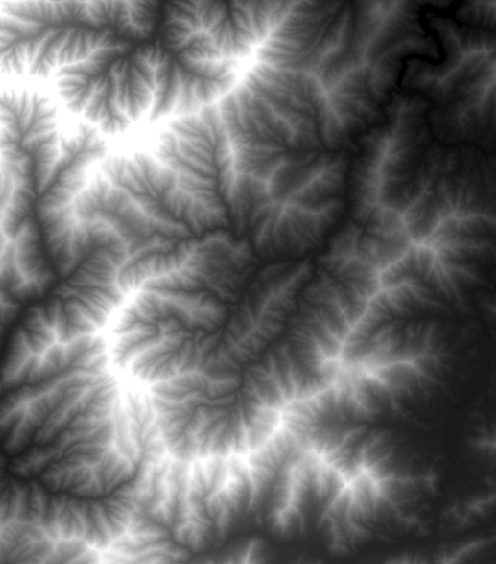
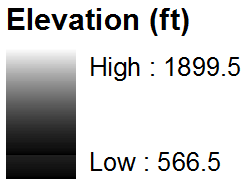
### Digital Elevation Model (DEM)

A digital elevation model, or more commonly referred to as a DEM, is a continuous data set that provides elevation data and a raster format. As raster can sometimes be very large in file size, DEM’s may be initially provided in an ASCII format so that it can be easily compressed down to a smaller size.

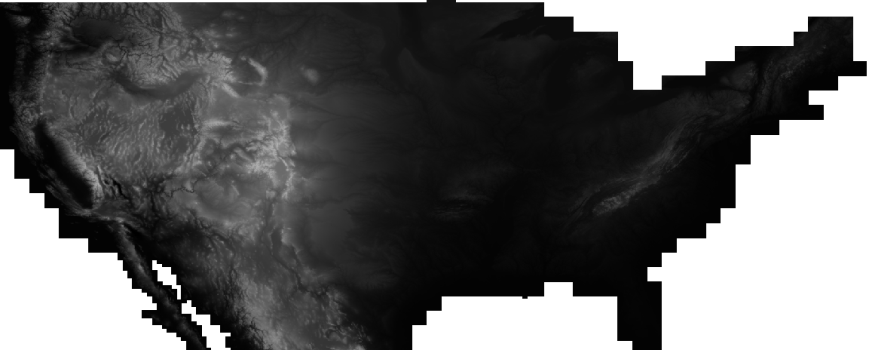
DEMs are provided at various scales depending on who collected the elevation data and for what purpose. DEM’s are extremely useful for analysis and visualization of elevation information. A challenge with DEM data is that there are so many different formats for it, that you may need to do significant amounts of converting before you can place it into your GIS.

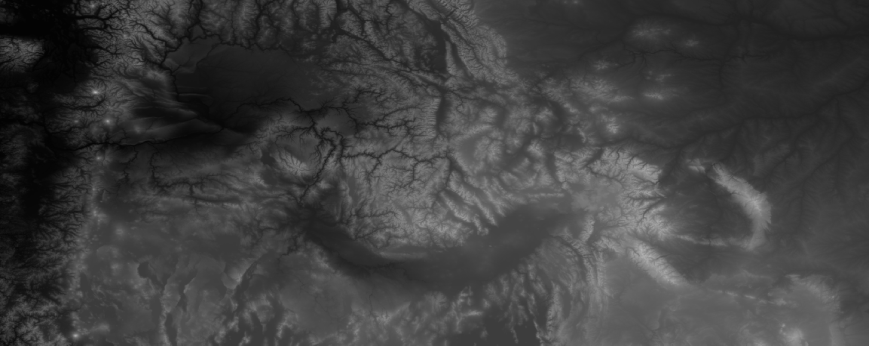
#### Multiple Stream Networks

This image is an example of a DEM that is showing multiple stream networks in a fairly mountainous terrain. As the raster only contains elevation information, DEMs do not have any color information stored inside of them. Therefore, it is up to you to assign a color ramp, or color map to the DEM to represent the different elevations. In a black and white image of a DEM higher elevations are typically represented in lighter colors, and lower elevations are presented with darker colors. For this DEM the highest value is pure white and the lowest value is a dark black. Knowing this we can easily identify where the streams are and the mountain ridges are.



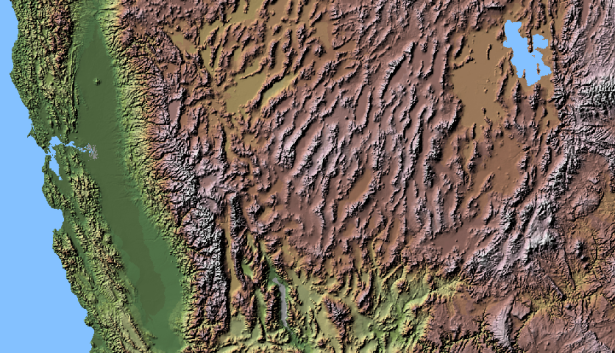
#### Various Coverage Area

DEMs come in many different resolutions and cover different amounts of space. The DEM on the top image covers the entire United States at a very low raster resolution. The DEM on the bottom covers a smaller area of the United States, as a high-resolution and thus has more detail.



#### Color

It is common for DEM’s to be assigned different shades of color to represent different types of terrain at different elevations. Additionally it is common to do shadow modeling based off the DEM’s so that the terrain has a three dimensional look.



### Digital Orthophoto Quadrangle (DOQ)

A Digital Orthophoto quadrangle, commonly referred to as a DOQ, is a digital or film photo saved in a raster format that is been corrected for distortion. DOQs are available for most of the United States at many different epochs in time. DOQs are useful for serving as a background image to your map, identifying control points for geo-rectification or projects, or for using as a base for digitizing new data.

#### True Color DOQ

This is an example of a DOQ that shows a portion of the earth in its true color representation.



#### Infrared DOQ

This is an example of a DOQ that shows a portion of the earth using the infrared portion of the electromagnetic spectrum.



#### Color DOQ Zoom

If we zoom into the true color DOQ, we can see that it starts to become slightly pixelated as we are reaching the limits of its resolution. However we can still see quite a bit of detail on this DOQ such as paint on the street, houses, cars, and other interesting features.



### National Aerial Imagery Program (NAIP)

The National Aerial Imagery Program (NAIP) is a program sponsored by the USDA Farm service agency. Their object is to acquire raster photos during growing season the continental United States of America at the one to two meter resolution. NAIP imagery is useful for digitizing and identifying information about vegetation type or condition.

#### NAIP Image

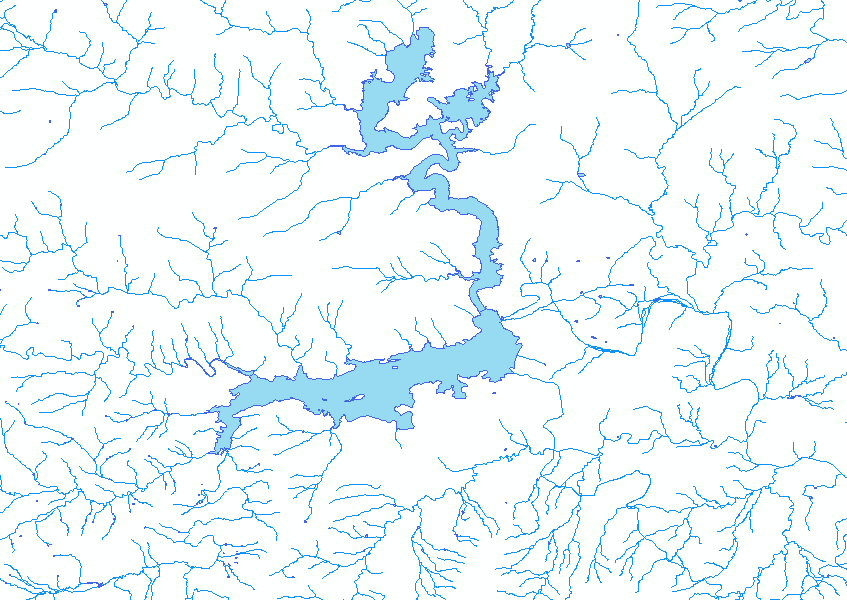
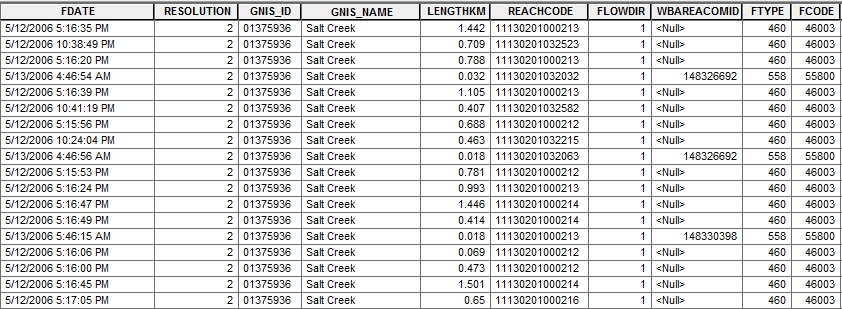
NAIP images are collected using infrared which makes it very easy to identify and differentiate different types of vegetation, and their health.



### Hydrologic Data

For hydrologic data a good source is the National Hydrologic Data Set (NHD). The NHD contains a database of surface waters derived from USGS DLG and EPA reach data. The data is provided at a 1 to 100,000 scale and represents networked topology. The EPA also provides information on watersheds and other data sets. The NHD is provided in a vector format but it might be provided in an exchange format that needs to be converted before use.

This image is an example of a portion of the NHD. The NHD has both line and polygon information and detailed attributes.

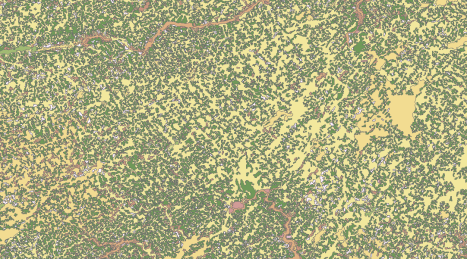


### Digital Soils Data

The natural resource conservation service of the USDA has developed vector data sets of soils data. The digital soils data are provided in three different data sets. NATSGO provides soils data at a small scale. STATSGO provides soils data at a medium scale, and SSURGO provides soils data at a large scale. All of these data sets were developed from soil surveys. The data sets link to very detailed databases that may require some training or reading the instructions to use.

#### Soils Data Example

This example is a snapshot of a data set from the soils data. Each polygon of a different color represents a different type of soil. Each polygon relates to a detailed database that outlines things such as the type of soils and permeability.

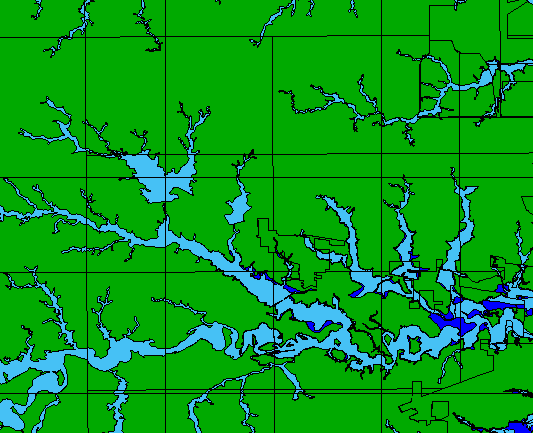


### Digital Floodplain Data

If you are looking for digital floodplain data, FEMA developed floodplain maps and a vector data set that is useful for floodplain determination and for setting insurance rates.

#### Floodplain Data Example

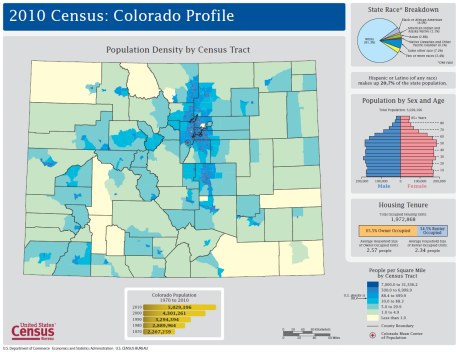
Here is an example from the FEMA floodplain data. The polygons represent different types of floodplains.

C:\Users\Rick\Documents\Cropper Captures\CropperCapture[33].png

### Census Data

The U.S. Census Bureau develops vector Tiger files that contain information collected during the national census. The Tiger files are organized by state, county, tract, block, and group. The Tiger files are in a GIS format that can easily link to census statistical data. This data is useful for a wide number of applications and the census data uses the topology.

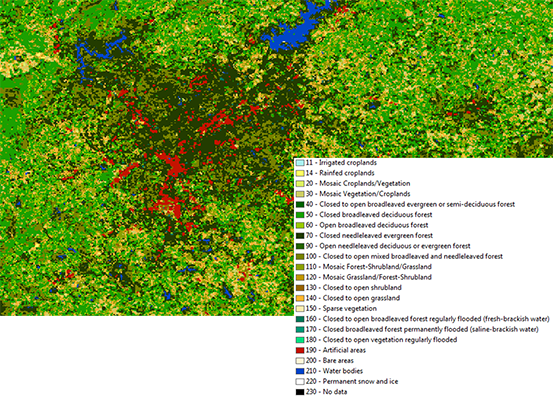
This image was created by the US Census Bureau that shows the 2010 census profile for Colorado. The census data provides information such as total population, population by race, population by sex and age, what types of houses the population lives in, and many, many other useful facets of the population of the United States.



### National Land Cover

The national land cover data set is provided in a raster format at a 30 meter resolution that covers the United States. The national land cover data set contains many classes such as where urban areas are, and agricultural areas. The national land cover data set is based on photo interpretation and classification and can be useful for analysis of land-use.

This is an example of a portion of a land cover data set. Each raster cell contains a value that represents a different type of land cover.



### Natural Earth Data

Another useful data set is known as the Natural Earth data. Natural Earth data is a data set that is in the public domain and contains both raster and vector data sets. The Natural Earth data set contains cultural and physical data with worldwide coverage at 1 to 10 million, 1 to 50 million and 1 to 100 million scales. The Natural Earth data provides a very nice generalized data set for small scale mapping.

#### Vector Data

This image is an example of the vector data offered in the Natural Earth data set. It provides information such as physical and political boundaries, water, and man-made features.



#### Raster Data

This is an example of the raster data offered in the Natural Earth data set. This raster data provides attractive elevation rasters with modeled shadows so that it achieves a good three-dimensional look.



### Websites for Data

This is by no means an exhaustive list of websites that are useful for downloading freely available GIS data sets, but are good sources to collect the data that was discussed in the previous section. In addition to these websites each state usually has its own GIS clearinghouse so you should check for one in your state of interest.

* + US Geo Data: [http://geo.data.gov](http://geo.data.gov/)
  + National Atlas: [http://www.nationalatlas.gov](http://www.nationalatlas.gov/)
  + USGS National Map: <http://nationalmap.gov/viewer.html>
  + National Hydro Data: <http://nhd.usgs.gov/data.html>
  + USDA Soil Data Mart: <http://soildatamart.nrcs.usda.gov/>
  + Natural Earth: <http://www.naturalearthdata.com/>
  + National Historical Geographic Information System: <http://nhgis.org/>
  + Census TIGER files from Census Bureau: <http://www.census.gov/geo/www/tiger/index.html>
  + Census Bureau State Data Centers Program: <http://www.census.gov/sdc/>

### Normalization of Data

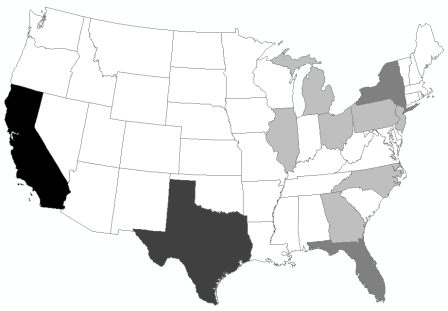
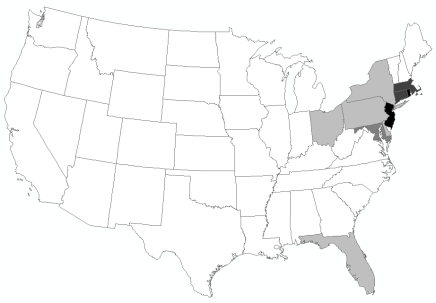
To begin learning about appropriate data use for thematic maps we start with the normalization of data concept. Raw and total values are often useful to maps; however, areas that are larger in size may have more of some value than areas that are smaller in size simply because there is more area. If you want to compare the density of an object between two areas that are different sizes a fair comparison is not possible. Therefore, we normalize data to allow for meaningful comparisons of values by taking size out of the equation. Raw data that is not normalized does not accurately display densities or proportions.

### Raw vs. Normalized Data

### If we compare raw data versus normalized data by looking at the total population of United States we get two completely different looking maps. The map on the left displays total population by state as a choropleth map. The lighter colors are where there is less population and the darker colors are where there is more population. Note that California and Texas are the highest states and are also among the larger states which means that there is more room for people to live so they naturally have higher populations.

### However if we normalize the data by area we can get a view of how dense each state is with respect to population. The map on the right shows the population per square mile which shows a significantly different view than the total population map. Here, the smaller states in the Northeast have less total population but have more people per square mile because the states are significantly smaller than the states out West.

**Total Population Population per Square Mile**

** **

Equal Interval Classification (explained later)

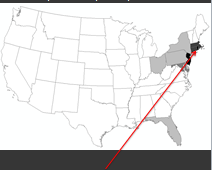
### Density

When discussing density, the idea is that larger enumeration or statistical units often have more of an attribute simply because it covers a larger area. Showing raw numbers per enumeration unit ignores the effect that the size the numeration unit might have on its total values. By accounting for the size of the enumeration unit it allows for a fair comparison between different locations as size is no longer a factor.

#### Enumeration Units

Again we have the two maps of population. In the total population map, large enumeration units lead to larger total population. Small enumeration units with more population per area mean that they are denser.

**Total Population Population per Square Mile**



Large enumeration units -> Small enumeration units with

larger total population. more population per area.

More Dense

#### How Much Exists?

Density reveals how much of an item exists within a given enumeration unit. In other words, you normalized the raw data value by the size of the numeration unit. For example, we would normalize population which is the raw data value by a square mile. This is an areal unit inside the statistical or enumeration unit. Having the size of the numeration units accounted for by normalization reveals a more meaningful pattern in the data.

**Population per Square Mile**

Raw Value Statistical/Enumeration unit

#### Proportions

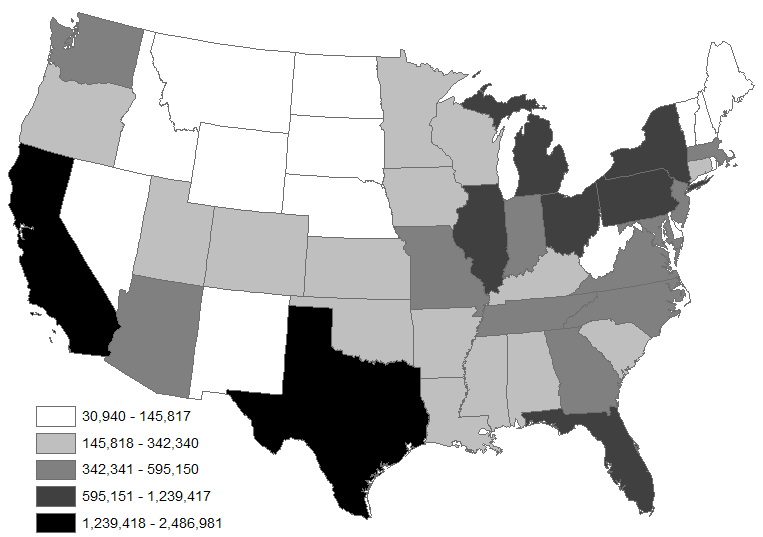
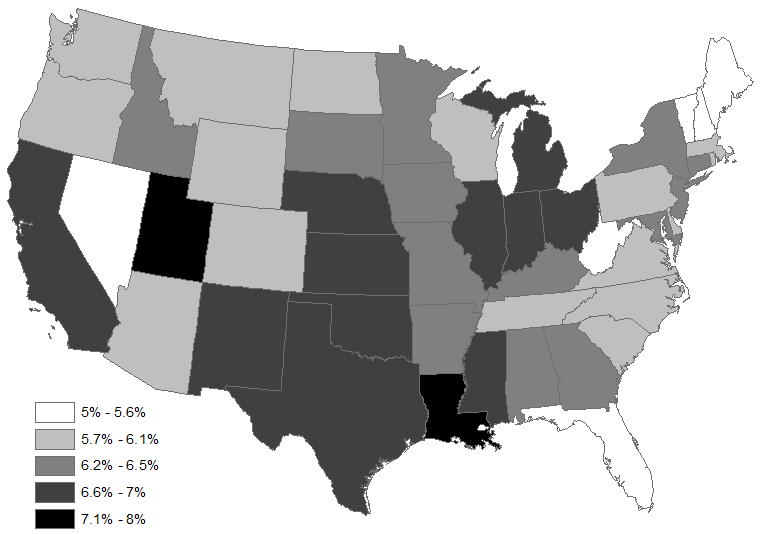
Proportions represent a relationship of a part of the whole. Proportions can be presented as percentages, for example, a population is 59% female. Proportions can also be presented as rates or in per X. For example:37 infections per 10,000 people. Similarly to density, proportions allow meaningful comparisons between enumeration units.

* + - Percentages – 0%-100%
      * e.x. 59% Female
    - Rates – N per X
      * e.x. 37 infections per 10,000 people

#### Raw Data

If you look at the raw data of total population under the age of five in the United States shown on the top left map we see that again Texas and California dominate the nation. However if we change the data and enter a proportion so that we show the percent population under five years old per state, we see that Louisiana and Utah top the nation.

**Total** Population under 5 years old **Percent** of Population under 5 years old

## Descriptive Statistics

### What are Descriptive Statistics

So what are descriptive statistics and why should you know about them for mapping? Descriptive statistics are quantitative descriptions of data that provides some basic descriptions of the data set. Cartographers use descriptive statistics to explore the character of the data. Descriptive statistics can describe things such as the central tendency of the data, the dispersion of the data, and the shape of the data.

### Central Tendency

Central tendency of data describes the distribution of the data. There are six measures that we commonly used to describe the central tendency of a data set: maximum, minimum, range, mean, median, and mode.

#### Measurements

The maximum measure of central tendency simply reports the maximum value in a data set. The minimum measure of central tendency returns the minimum value in a data set. The range measure returns the maximum value minus the minimum value of a data set.

#### Mean

The mean measure of central tendency is the average value of the data set. The average value is defined by summing all values of the data set and dividing it by the number of values in the data set. The medium measure of central tendency reports the middle point of the data.

For instance, if we had three observations in our data set we would sort the observations in ascending order and then look at the value at the midway point which would be the second observation of this case. If there is an even number of observations in a data set then the media will be the average of the two most central observations. The mode measure of central tendency reports the most common value found in the data set. If multiple values are tied as the most common value, the data set has multiple mode values.

Mean:

### Dispersion

Dispersion measures the variability of the data set. There are two measures of dispersion: variants and standard deviation.

#### Variance

The variance measure of dispersion takes the sum of the squares of the deviations divided by the number of observations. The units of variance are identical to the original units of measure. So for instance, if the original units of the observations were in feet then they variance will report how much the observations vary on average in feet.

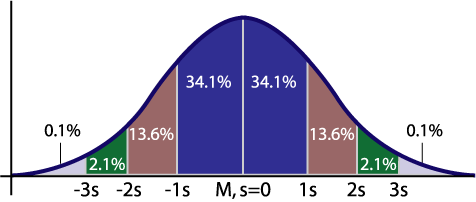
#### Standard Deviation

The standard deviation is similar to the variance except that it takes the square root of the sum of the squares of the variance divided by the number of observations. The standard deviation measures dispersion in a standard way so that two different data sets can be compared. Standard deviation does not report the dispersion in the original units of the observations.

### Normal Distribution

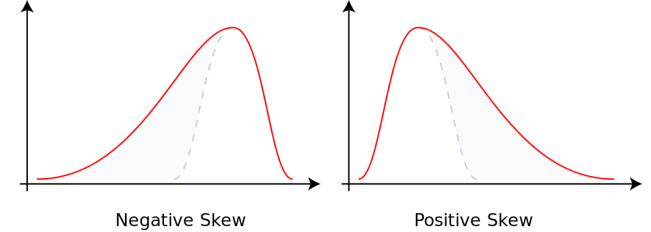
In the image here you are looking at a normal distribution of data which means that most of the examples of the data set are close to the average value while relatively few examples tend to one extreme or the other. The X axis is the value in question and the y-axis is the number of observations for each value on the x-axis. The standard deviation of tells us how tightly all the various observations are clustered around the mean of the data set.

When the examples are pretty tightly bunched together the bell shaped curve is steep and the standard deviation is small. When the examples are spread apart the Bell curve is relatively flat which tells you that you have a reasonably large standard deviation. One standard deviation away from the mean in either direction on the horizontal axis accounts for about 68% of the observations in the data set. Two standard deviations away from the mean which are the four areas closest to the center account for about 95% of the observations the data set. Three standard deviations account for about 99% of all the observations of the curve.



### Skewness

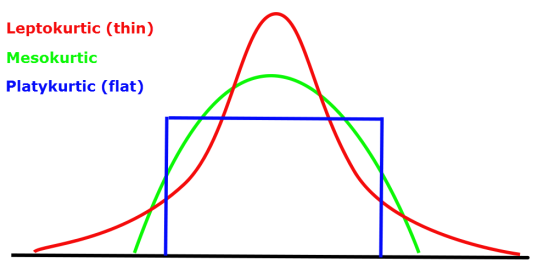
The skewness measure tells us whether the peak of a distribution is to one side of the mean or the other. If a data set has a negative skew then the peak of the distribution is above the average. If a data set has a positive skew the peak of the data set is below the average.



### Kurtosis

The final measure of central tendency is Kurtosis. Kurtosis describes the flatness or peakedness of a distribution. For normal distribution, the Kurtosis is equal to the value of 3.0. A normally distributed data set having the Kurtosis value of 3.0 is a mesokurtic distribution. A value above 3.0 is a leptokurtic distribution. A value less than 3.0 is a platykurtic distribution.

This image is a visualization of the three types of kurtosis. The green line is the normal distribution with a value of 3.0 and is mesokurtic. The red line with a strong peak is the leptokurtic distribution an d has a value greater than 3.0. The blue line which has a flat top is the platykurtic distribution which has a value less than 3.0.



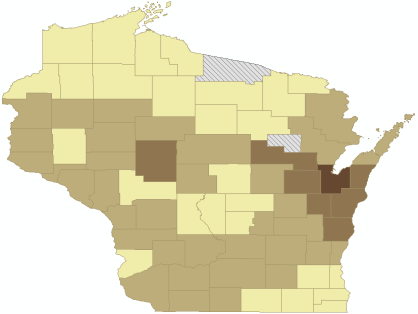
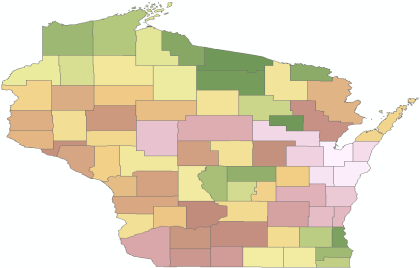
## Data Classification

### What is Data Classification

Data classification categorizes objects based on a set of conditions into separate bins or classes. Classification may add to or modify attribute data for each geographic object. For example a classification could add a new code such as large or small. Classification could also recode an attribute such as changing urban to dense. One attribute can yield many different maps depending on which classification method is chosen. Different classification methods will have a direct effect on how the map and data are perceived by the map user, therefore much care must be taken when choosing a data classification method.

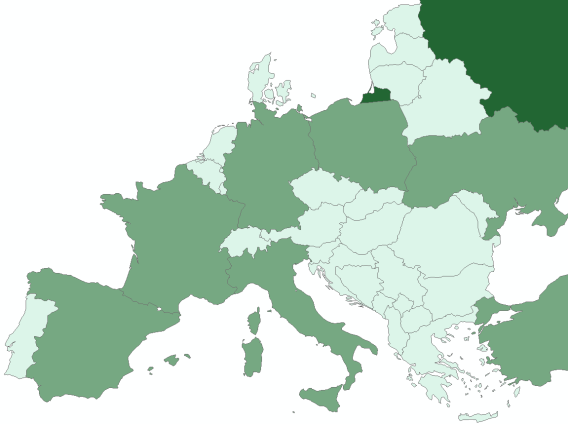
### Why Classify Data?

The primary reason to classify data is to simplify the data for visual display. For instance, the map on the right assigns each value a unique color which in this case means that each county has its own unique color. It is very difficult to look for patterns in this case when each county has its own unique color. The map on the left uses the same data but uses a data classification method to categorize the counties to make it easier to visualize the data. With the classified data the lighter values represent less of an item and the darker values represent more of an item, spatial patterns of distribution begin to emerge.

### Example of a Classification

This image is an example of a classification showing countries with high, medium, and low values of population.



### Video Example

This video demonstrates how to create a classification and ArcGIS based on total population for European countries.

Watch "Example Classification in ArcGIS 10.1" (1:15)

### Classification Goals

There are three goals for data classification with regards to cartography. The first goal is to simplify the visualization so that spatial patterns of distribution can become more easily viewable by the map reader. The second goal is to group similar observations together. The third goal is to show the difference between the groups.

## Classification Requirements

### Maintain Character

The main consideration when classifying data is that you should maintain the character of the original data even after classification. The class of data should still represent the trend of the data. Also, do not choose colors or classes that intentionally mislead the map reader. Your goal in classifying data is to simplify the data for visualization, not to modify the character of the data.

### Full Range of Data

The next classification requirement is that you should always encompass the full range of data. Do not exclude data because it is convenient or inconvenient. You will want to exclude “no data” values from the classification. The “no data” data values should still be visualized but not as part of the classification continua. Instead the “no data” values should be visualized differently than the other classes. When looking for “no data” values in a data set you should read the metadata for the data set to see how they represent the no data values. Common no data values are: no data, -99, -9999, null, ND, NA.

### No Overlapping or Vacancies

When classifying the data you should never have overlapping or vacant classes. For example, if we look at this legend for the choropleth map, two classes claim to contain the value of 30. This should never happen as a single value can only exist in a single class. To fix this, you should change the range of one of the two classes.

The class that holds 17 to 30 could be changed to hold the values 17 to 29. You should never have vacant classes, which means a class that has no observations inside of it. The only exception is if you’re making a series of maps covering a wide time range and you want to use a single legend for all of the maps.



Overlapping classes

Which class does ‘30’ belong in?

### Balance

The next classification requirement is that you should strike a balance of the number of classes for your map and data. As a number of classes increase visual interpretation becomes more difficult for the map reader. Therefore you should choose a number of classes that is equal to interpret but still shows adequate complexity in the data.

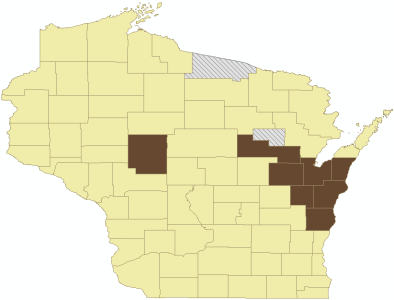
### Choosing Number of Classes

When considering the following when determining the number of classes to choose. If you choose too many classes it requires the map reader to remember too much when viewing your map and that may require them to make frequent trips to the legend. It may also make the differentiation of class colors difficult for the map reader.

If you choose to view classes it oversimplifies the data possibly hiding important patterns. Additionally each class may group dissimilar items together which is in direct opposition of one of the main goals of classification. Typically in cartography three to seven classes are preferred and five is probably the most common and optimal for most thematic maps. Feel free to use more or less classes as you see fit. Keep in mind the issues with choosing many, or few classes.

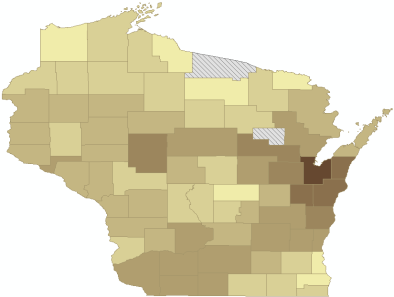
#### Two Classes Example

On this map there are only two classes which make for a fairly boring map and it is probably hiding more complex patterns. Notice the two counties that are not shaded the color but instead have the hachures. These two counties are symbolized differently from all the other counties because those two counties did not have any data reported.

****

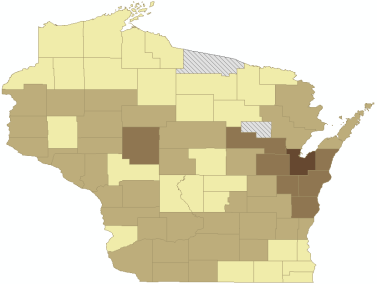
#### Too Many Classes Example

On this map there are way too many classes which make it difficult to tell some shades apart and will require multiple trips to the legend to remember which shade of color corresponds to which value range.



#### Four Classes Example

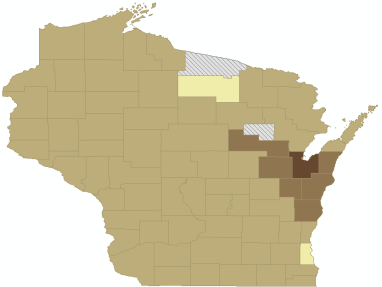
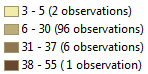
This map has a good number of classes. This map has four classes and still shows useful patterns and complexity.



### Divide Data

The next classification requirement is that you should always divide data into reasonably equal groups. Classes do not always need to be equal in size; however, classes should not have very large discrepancies and observations and has a goal of classification as a group like observations.

In this example classes one, three, and four have two, six, and one observations respectively. The second class has 96 observations, and encompasses a disproportionately large portion of the data observations thus making for a very boring, and unbalanced map.

### Logical Mathematical Relationships

When classifying data you should have logical mathematical relationships if possible. This means that if the data lends itself to common mathematical relationship you should use it. If the data set has a normal distribution perhaps you should choose the standard deviation data classification method.

Academic Grades

100-90 = “A”; 89-80 = “B”; …

### Choose Colors Wisely

You should choose colors wisely when applying them to classes. There are two categories of color choices that are appropriate for classifying quantitative data: sequential, and divergent.

Sequential



Divergent



#### Sequential Color

Use a sequential color ramp to represent categories with data that increase incrementally from a low to high value. To accomplish this you may choose a single queue, or color, with incremental changes in its value or saturation, but not both. Colors that are more saturated represent higher data values and colors that are less saturated represent lower data values.

#### Divergent Color

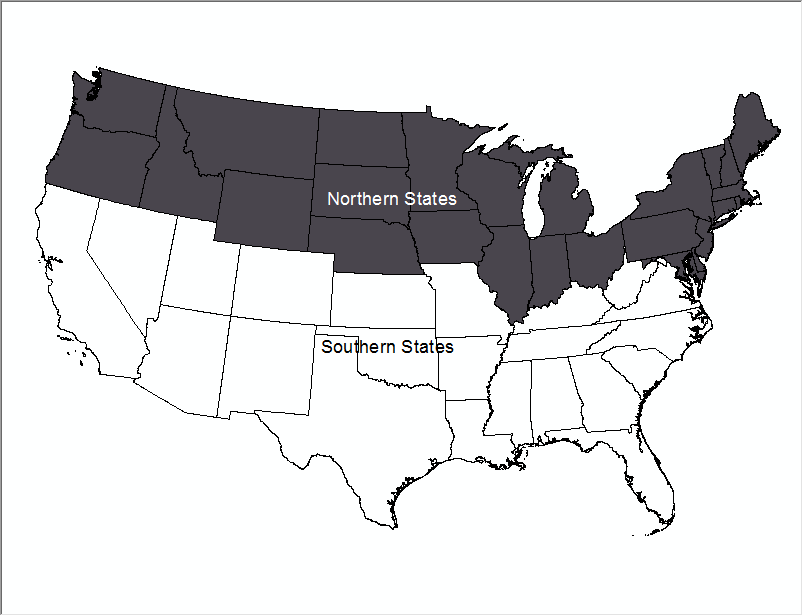
Use a divergent color ramp to represent categories with data that increase in opposite directions from some neutral point. Complementary hues should be used in saturations should increase towards each extreme. Colors that are more saturated represent more extreme data values and colors are less saturated represent less extreme data values that are closer to the neutral point. If a central value does exist it should be neutral gray or equal low saturation amounts.

## Classification Methods

### Binary Classification

Binary classification is when objects are placed into two classes. The two classes can be the value of 0 and 1, true and false, or any other dichotomy that you would use. Typically a binary classification is used to store the results of complex operations when the operation returns either a yes or no answer.

In this example the states were classified into two binary classes: one class for the northern states and one class for the southern states.



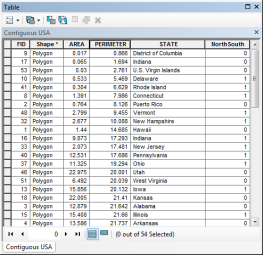
Video: This video demonstrates how to perform the binary classification to achieve the classification of northern and southern states using ArcGIS 10.1.

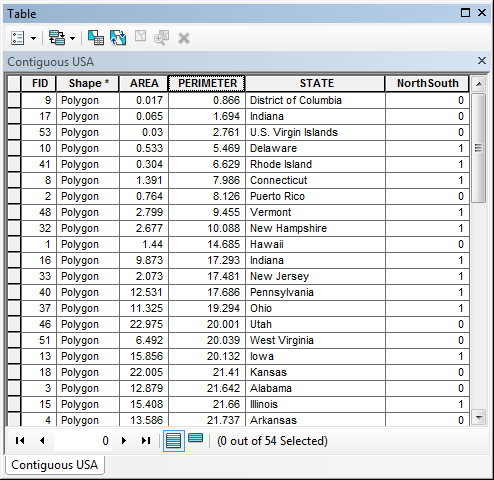
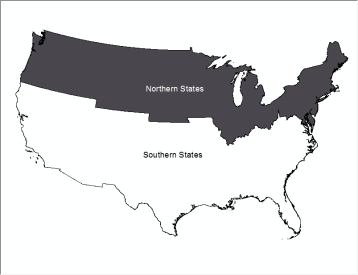
Watch "Example Binary Classification in ArcGIS 10.1" (1:34)

### Dissolve

The next classification method we can use is the dissolve method. The dissolve method combines similar features within a data layer based on a shared attribute. Considering the northern and southern states example again we can dissolve all the states based on whether they are a northern state or a southern state.

The dissolve operation creates new geometry which in this case is polygons. One polygon combines the extents of all of the northern states. The second polygon combines all the extents of the southern states. Take note that the new geometry does not transfer any of the attributes of the dissolved.



### Automatic Classification

This section we explain automatic classification methods. An automatic classification method is where you set up rules for the computer to follow and in the computer executes the data classification.

### Jenk’s Natural Breaks

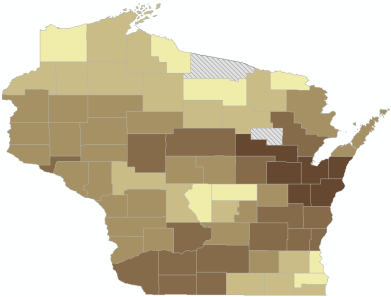
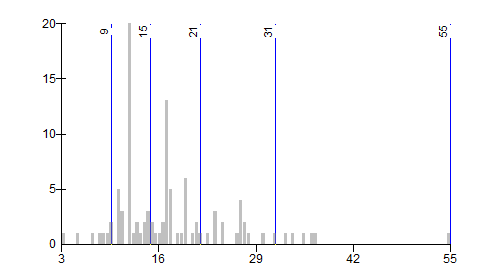
The Jenks natural breaks classification method aims to maximize homogenate classes. It uses breaks in a histogram as class breaks and assumes that group data are alike. Advantages of the Jenks natural breaks automatic classification method is that it considers the distribution of data to minimize in class variance and maximize between class periods. This method produces a classification with high accuracy of maximizing homogenate eating classes. A disadvantage of the Jenks natural breaks is that it is complicated to understand how it works.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Considers distribution of data | Complicated |
| Minimizes in-class variance | Difficult to understand procedure for grouping |
| Maximizes between-class variance |  |
| Produces classification with high accuracy |  |

#### Natural Breaks

This example is the Wisconsin map classified using the natural breaks method. On the right is the histogram of the values where the X axis is the observation value and the y-axis is the number of observations for each value. The vertical blue lines show the extent of each class and are considered to be the class breaks.

Looking at the histogram and the blue lines you can see that natural breaks tend to place the class breaks whether our natural valleys or vacancies in the data set. The map on the left displays the result of the classification visually and shows the five classes varying from a light brown to a dark brown in a sequential color. This map is interesting to look at and displays interesting patterns in the data.

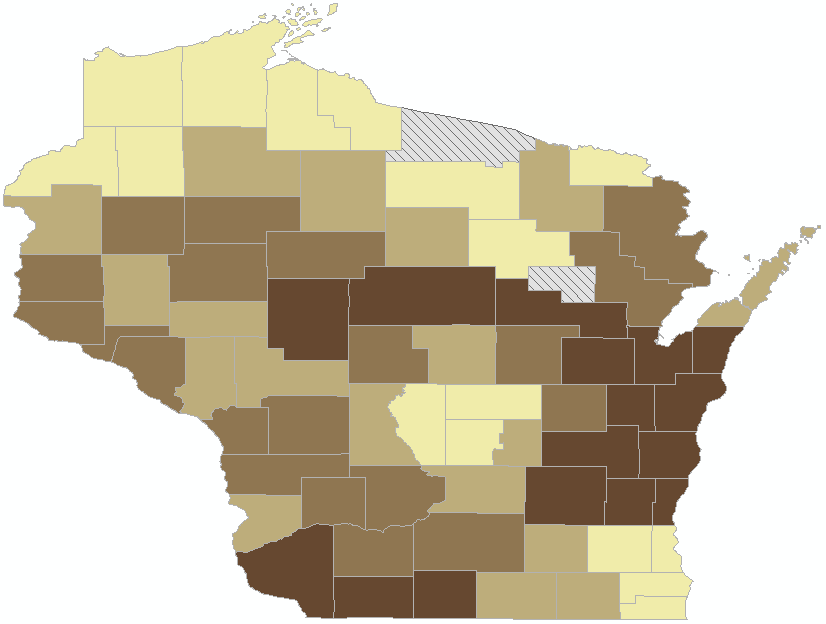
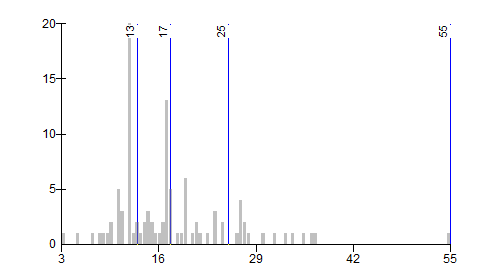
### Nested Means

The nested means automatic classification method creates classes about the arithmetic means of the data set. Additional means can be calculated about the first mean to create additional classes. The advantage of the nested means automatic classification method is that it is easily computed and mathematically intuitive. The disadvantages are that this classification method is limited to 2, 4, 8, or additional powers of two numbers of classes. Additionally, the nested means does not consider the distribution of data, only where the averages of the data fall.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Easily computed | Limited to 2-4-8-…. Classes |
| Mathematically intuitive | Does not consider distribution of data |
|  |  |
|  |  |

#### Nested Means Explanation

The way the nested means works is at first two classes are created 1 above the mean and 1 below the mean. The third and fourth classes are created above and below the means from the first two classes. In this case it creates a reasonably interesting map and each class has about the same number of observations.

**** ****

### Mean and Standard Deviation

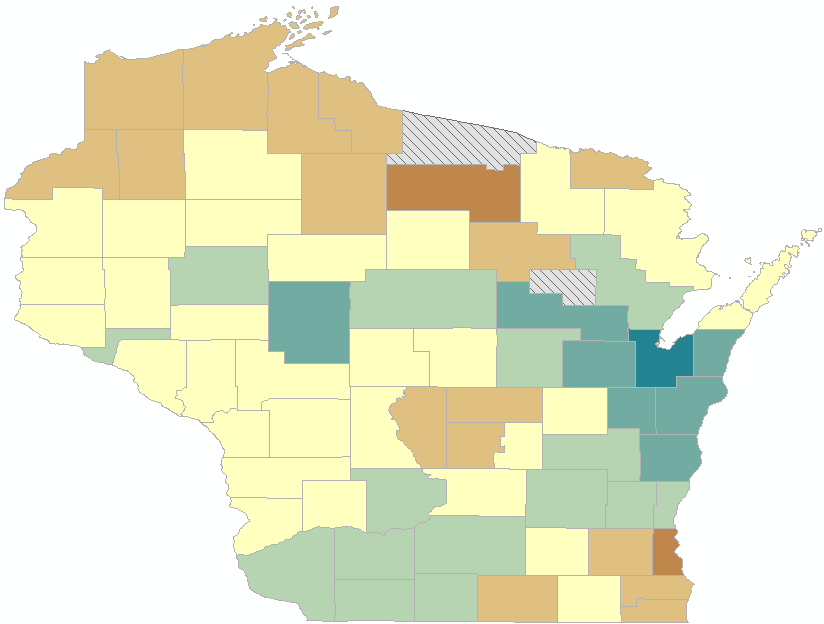
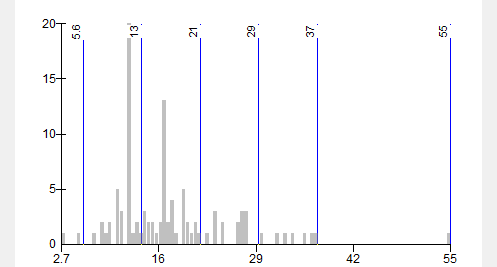
The mean and standard deviation automatic classification method creates classes about the arithmetic mean and standard deviations above and below the mean. For this method since we are looking at divergent behavior you should use a diverging color red.

Advantages of this classification method are that it is good for displaying data with a normal distribution and it considers the distribution of data. It produces constant class intervals, one for each standard deviation above and below the mean. Disadvantages of this method are that most data is not normally distributed and is not a good fit for this method. This method requires that the map user understands statistics to appropriately interpret the map.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Good for data with normal distribution | Most data are not normally distributed |
| Considers distribution of data | Requires understanding of statistics |
| Produces constant class intervals | Not easily understood by map reader |
|  |  |

#### Mean and Standard Deviation Example

In this example a diverging color ramp is used with the colors tending towards brown being observations below the mean and observations towards blue being above the mean. The data for the state has a positive skew, it is not normally distributed, therefore this method is not ideal for this data set but does still tend to display interesting geographic patterns.

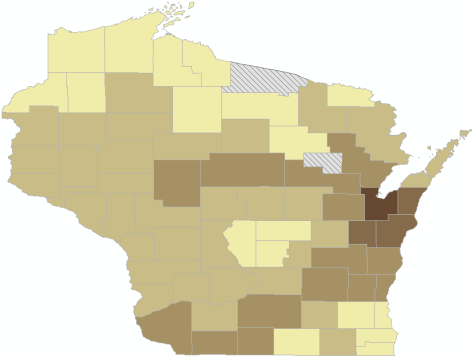
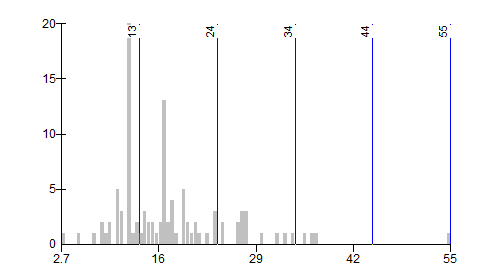
### Equal Interval

The equal interval automatic classification method creates classes with equal ranges. The class range is calculated by taking the maximum value of the data set, subtracting the minimum value from it, and then dividing that by the number of observations the data set. The advantages of the equal interval classification method are that it is easy to understand, simple to compute, and leaves the legend. Disadvantages are that it does not consider the distribution of data at all and it may produce classes with 0 observations which is highly undesirable.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| **Easily understood** | **Does not consider distribution of data** |
| **Simple to compute** | **May produce class with zero observations** |
| **No gaps in legend** |  |
|  |  |

### Equal Interval Example

If we look at the histogram of the data we can see that the class breaks are evenly distributed which is the goal of the equal interval classification method. Looking at the map you see some of the interesting patterns are now hidden since the class ranges are including data that is different from each other. The map is a little boring for this data set.

**** ****

### Equal Frequency

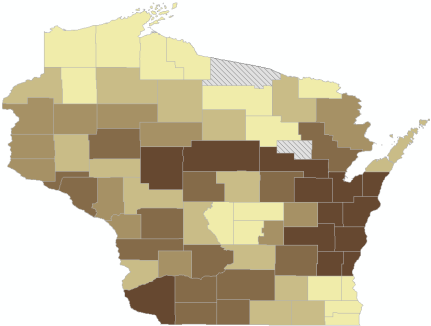
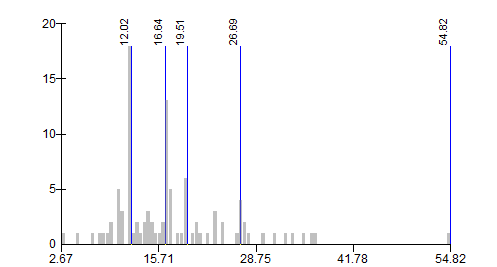
The equal frequency automatic classification method is also known as the Quantile classification method. The goal of the equal frequency method is to distribute observations equally among classes which mean each class will have the same number of observations. If the number of observations does not divide equally into the number of classes then you should place the extra observations across the lower classes which will overload the lower classes.

Advantages of this classification method are that it is easily calculated, applicable to ordinal data, and will not create a class with 0 observations. Disadvantages of the equal frequency automatic classification methods are that it does not consider the distribution of data can create gaps in the legend and observation values inside of a class may not be similar.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Easily calculated | Does not consider distribution of data |
| Applicable to ordinal data | Gaps in legend |
| No classes with zero observations | Observation value may be closer to value in different class |

### Equal Frequency Example

As the equal frequency automatic classification method aims to have the same number of observations in each class there should be about the same amount of each color on the map which leads towards a balanced looking map. A major negative is at a class may contain data values that are dissimilar which goes against the purpose of classifying data.

**** ****

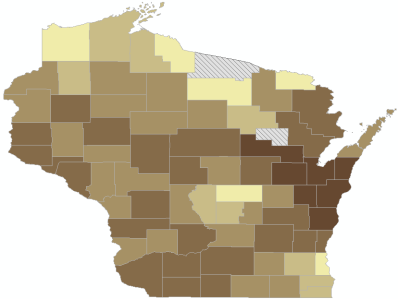
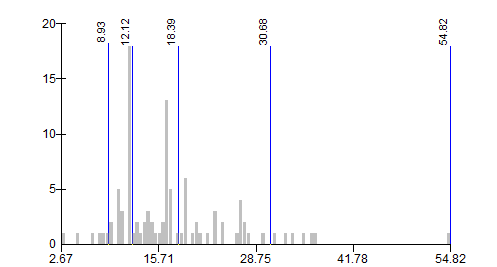
### Arithmetic and Geometric Intervals

The arithmetic and geometric intervals automatic classification methods create class boundaries that change systematically with a mathematical progression. This classification method is useful when a range of observations are significant and the observations follow some sort of mathematical progression that can be followed with the classes. Advantages are that it is good for data with large ranges and the breakpoints are determined by the rate of change the data set. Disadvantages are that it is not appropriate for data with a small range or with linear trends.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| **Good for data with large ranges** | **No appropriate for data with small ranges or linear trends** |
| **Break points determined by rate of change in the data set** |  |
|  |  |

### Arithmetic and Geometric Example

If you look at the histogram you can see that the class ranges slowly increase along the x-axis as this trying to fit the data in the classes that increase at increasing rates. On the map this creates a reasonably interesting map but in this case does seem to over fit the data and make additional patterns exist were they probably should not.

**** ****

### Manual Classification

Finally, the user can define a classification method which is known as manual classification. In this case you can choose your own class boundaries to create your own classification scheme.

## SUMMARY

In this lesson you learned about the elements related to data for a map. You learned the common data types and formats used in map creation. The lesson helped you gain an understanding of normalized data and descriptive statistics. You also learned about data classifications and the various classification requirements and methods used when making maps.

## ASSIGNMENTS

1. Quiz: Data for a Map
2. Lab: Data Classification