GIST 3300 / 5300
Geographic Information Systems

Data Classification Methods

Last Time

Qualitative Thematic Data & Maps

Types of Qualitative Data
- Descriptive/categorical/nominal/text
- Names, types, ranks

Types of Qualitative Thematic Maps
- Unique symbol
- Categories
Last Time

Quantitative Thematic Maps & Data

Types of Quantitative Data
- counts
- amounts
- ratios

Types of Quantitative Thematic Maps
- graduated color
- graduated symbol
- proportional symbol
- dot density
- pie chart / bar charts
- isoline maps (not shown)
- colored or shaded rasters

Data Classification Methods for Quantitative Data

After determining which quantitative display to use (graduated colors, graduated or proportional symbols, or dot density), the next step is to determine which method to use to classify the quantitative data and the number of classes needed.

Data Classification Methods
- natural breaks
- quantile
- geometrical interval
- equal interval (defined Interval)
- standard deviation
- and manual

How Many Classes are needed?

Making Classes Easier to Read
What is classification?

Classification - organizes data into different groups or categories.

So features in the same category can be symbolized the same way to separate them from features placed in other groups or categories.

<table>
<thead>
<tr>
<th>POP_2010</th>
<th>Ex. Legend showing 4 population categories. Features within this population range are associated with the appropriate group and assigned the same symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>563,626 - 10,000,000</td>
<td></td>
</tr>
<tr>
<td>10,000,001 - 20,000,000</td>
<td></td>
</tr>
<tr>
<td>20,000,001 - 30,000,000</td>
<td></td>
</tr>
<tr>
<td>30,000,001 - 40,000,000</td>
<td></td>
</tr>
</tbody>
</table>

Data Classification

The way in which you classify your data can have a dramatic effect on the way a map looks – and how other people interpret your results.
Data Classification

- to create quantitative thematic maps, the data must be classified.

- ArcMap has seven different types of data classification methods:
  - natural breaks
  - quantile
  - geometrical interval
  - equal interval (defined interval)
  - standard deviation
  - manual
Data Classification

Access the classification methods from the Classify… Button

Layer Properties Dialog Box

Classify Button

Classification Methods
- manual
- equal interval
- defined interval
- quantile
- natural breaks
- geometrical interval
- standard deviation

Histogram
Data Classification

Classification Histogram  
– relates data distribution to frequency

Example: Texas County Population range by number of counties

Left Skewed Distribution

Data Classification

- choosing a classification method - depends upon the distribution of your data

Ex. County Population  
Ex. Percent Males by County

Ex. Average Family Size by State  
Ex. Percent College Education
Data Classification

**Natural Breaks** (default classification method in ArcMap)

**works:**
- automatic selection of best class breaks (clusters)
- maximized the variation between classes
- while minimizing the variation within a class

**used:**
- can be used for any distribution
- works best for an uneven (clustered) distribution of data

**disadv:**
- 1) class breaks are unique to the dataset and map
- 2) number of classes selected must be appropriate to data
Data Classification

Quantile

works: - places an equal number of features in each class  
- ex. classify lower 48 states into 3 classes, so each class has 16 states

used: - works best for an even distribution or normal distribution of data

<table>
<thead>
<tr>
<th></th>
<th>Even Distribution</th>
<th>Normal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Features</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Range of Values</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

similar range of values in each class  
different range of values in each class

Quantile

disadv: - if the distribution of data is uneven or strongly skewed, very different values can end up in same class

Strongly Skewed Distribution

the last class has a very large range of values
Data Classification

Quantile

**disadv:** if the distribution of data is uneven or strongly skewed (as with most population data), very different values can end up in same class.

**Texas County Population**

Note the class breaks and range of values.

Data Classification – Comparing Natural Breaks & Quantile

**Natural Breaks**

Most counties have low population
A small number have medium or large population

**Quantile**

The class with the greatest population has the same number of counties and the class with the lowest population.
Data Classification

Geometrical Interval

**works:** - compromise between natural breaks and quantile
- originally called "smart quantiles"
- designed for rasters with many equal grid cells values
- became available in ArcGIS version 9.2 (2006)

**used:** - can be used for any distribution of data
- works best for an uneven distribution of data

**disadv:** - number of classes selected should be appropriate to data
Data Classification

Equal Interval and Defined Interval
- both methods classify data in a similar manner

Equal Interval
- user sets the number of classes
- ArcMap calculates the same range of values (interval) for each class

Defined Interval
- user specifies the interval size
- ArcMap determines the number of classes required

Data Classification

Equal Interval
- user sets the number of classes (ex 5)
- class ranges typically defined by rounded numbers
  ex. 0%-20%, 20%-40%, 40%-60%, 60%-80%, 80%-100%
- ArcMap calculates the range of values
- each class has the same range (interval)

Number of classes: 4
Interval size: 9,736,209 people

Number of classes: 8
Interval size: 5,149,917 people
Data Classification

**Defined Interval**
- user sets the interval size (ex. 10,000,000)
- ArcMap calculates the number of classes

<table>
<thead>
<tr>
<th>Interval size: 10,000,000 people</th>
<th>Interval size: 5,000,000 people</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of classes: 4</td>
<td>Number of classes: 8</td>
</tr>
</tbody>
</table>

**Data Classification**

**Equal Interval and Defined Interval**

**works:** - each class has the same range of values (interval size)

**used:** - works best for an even or normal distribution of data
- defined interval must be used for isolines (contours, isotherms)
  ex. 20 foot contour intervals.

<table>
<thead>
<tr>
<th>Even Distribution</th>
<th>Normal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range of Values</td>
<td>Range of Values</td>
</tr>
<tr>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td>33</td>
<td>52</td>
</tr>
<tr>
<td>31</td>
<td>28</td>
</tr>
</tbody>
</table>

similar number of features in each class
different number of features in each class
Data Classification

Equal Interval and Defined Interval

**disadv:**
1) if distribution is uneven - many features in one or two classes
2) there is the potential for some classes to have little or no data
   - So colors in legend don’t match anything on the map

![Graph showing uneven distribution with number of features and range of values.](image)

Each class has the same value range, but some classes have no data or few features.
Data Classification

Standard Deviation

works: - classes reflect the distance from the mean (average)
- then places class breaks above and below mean at .25, .5 or 1σ

used: - works best for a normal distribution of data
- used to emphasize the extremes (they get their own class)

Disadv: - map legend does not show actual values
- high value outliers can greatly skew the mean
Data Classification

Manual

works: - user defined number of classes and range of class values

used: - often start with natural breaks, then round class values manually

disadv: - class breaks become subjective

- how many classes?

- if fewer than four classes
  - map doesn’t show much spatial variation among the features

- four, five or six classes
  - usually adequate to show spatial variation in features

- if seven or more classes
  - becomes difficult to distinguish between different colors (in a color ramp) representing classes in legend
Data Classification

- how many classes?

Quantile with seven classes
- too many shades

Quantile with seven classes
- too many colors
Data Classification

- making classes easier to read

- round the labels in the legend* (or manually edit labels in TOC)

<table>
<thead>
<tr>
<th>Range</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,572 – 23,456</td>
<td>&lt;= 25,000</td>
</tr>
<tr>
<td>23,457 – 48,292</td>
<td>25,001 – 50,000</td>
</tr>
<tr>
<td>48,293 – 68,842</td>
<td>50,001 – 75,000</td>
</tr>
<tr>
<td>68,843 – 93,223</td>
<td>&gt; 75,000</td>
</tr>
</tbody>
</table>

*Changing the labels doesn't change the classification. Changing the values affects the classification.

- or label values as low, medium and high in the legend

- use if relative values are more important than actual numerical values

<table>
<thead>
<tr>
<th>Range</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02 – 0.08</td>
<td>very low</td>
</tr>
<tr>
<td>0.09 – 0.24</td>
<td>low</td>
</tr>
<tr>
<td>0.25 – 2.34</td>
<td>moderate</td>
</tr>
<tr>
<td>2.35 – 6.95</td>
<td>high</td>
</tr>
<tr>
<td>6.96 – 9.48</td>
<td>very high</td>
</tr>
</tbody>
</table>

Data Classification – using different resolutions (relatively coarser & finer)

2012 Presidential Election Results

regional emphasis using state data with two classes

rural / urban emphasis using county data With two classes
Data Classification

This type of map is called a cartogram which scales areas.

2012 Presidential Election Results

cartogram scaled by electoral college votes

Data Classification

2012 Presidential Election Results

“The purple map”

Counts classified into several classes based on % democrat to republican. Easier to see how counties split the vote. Battle ground areas are more purple than red or blue.
Types of Quantitative Thematic Maps (review from last lecture)

Example: 48states > Layer Properties dialog box > Symbology tab > Quantities
- used to control how you want to display your quantitative data using one of the available renderers or display methods >

Example: US state population data to illustrate the different ways to represent quantitative data
1) graduated color (choropleth map), 2) graduated symbol, 3) proportional symbol, 4) dot density, 5) charts (pie chart, bar chart, stacked chart), 6) isoline maps and colored rasters

[Images of maps illustrating the different types of thematic maps]
**Data Classification**

Definition: Classification organizes data into different groups or categories. Features in the same group can be symbolized to distinguish them from features in other groups.

- to create quantitative thematic maps - the data must be classified
- and choosing the best classification method often depends upon the distribution of your data

- How to Lie with Maps by Mark Monmonier – a fun book on the use and abuse of maps and how to critically evaluate them.

- to view the distribution of features across a range of data …
  
  - open the Layer Properties pf the dialog box, select the Symbology tab … and then select Quantities
  - then select the Classify button … this will open the Classification dialog box
  - in the Classification dialog box you can view a histogram showing the distribution of your data
  - the histogram shows the number of features across the range of data values

- looking at this histogram for various classification methods we typically can find that …
  - the number of features is **unevenly distributed** across the range of data values
  - or the number of features is **evenly distributed** across the range of data values
  - or the number of features is **normally distributed** across the range of data values
  - or the number of features has a **skewed distribution** across the range of data values
Data Classification Methods

- ArcMap offers seven different methods for classifying quantitative data

1) natural breaks
   - automatic selection of best class breaks (default classification in ArcMap)
   - minimizes the variation within a class - while maximizing the variation between classes
     - works best for an uneven (clustered) distribution of data
     - but the number of classes must be appropriate for the data
     - note that class breaks are unique to that dataset …
       - this makes it difficult to directly compare different maps in a time series

2) quantile
   - each class has an equal number of features
     - works best for an even or normal distribution
     - often produces a good-looking map (same number of features in each class)
       - colors are equally distributed on the map
     - if the distribution is uneven or skewed …
       - you can end up with very different values in the same class

3) geometrical interval
   - a compromise between natural breaks and quantile
     - recently became available in ArcGIS 9.2 - originally called “smart quantiles”
       - works well for almost any distribution of data
       - like quantile, it often produces a good-looking map
       - if the distribution is uneven, number of classes should be appropriate for the data

4) and 5) equal interval and defined interval)
   - each class has same range of values
     a) equal interval – user selects the number of classes – ArcMap calculates the interval
     b) defined interval – user selects the interval – ArcMap calculates the number of classes
       - works best for an even or normal distribution
       - if distribution is uneven - potential for many features in one or two class
       - potential for some classes to have no data (which is not acceptable on a map)
       - equal interval must be used for elevation contours and most other isoline maps

6) standard deviation
   - classes reflect the distance from the mean (emphasis on very high and very low values)
- works best for a normal distribution
- map legend shows standard deviations - not the actual range of values in each class
- there is also the potential for extreme outlier values to skew the mean

7) manual

- user defines the number of classes and range of class values
- often start out with natural breaks or quantile,
  - then round the class break values to make logical class intervals
- when a manual classification is used, class breaks become a subjective decision
  - which might or might not be important

- how many classes?
  - less than four classes - doesn't show much spatial variation between features
  - four, five or six classes - usually adequate to show spatial variation in features
  - seven or more classes - difficult to distinguish between colors (in a color ramp)

- making classes easier to read
  - often the class breaks are a series of apparently nonsensical numbers
    - round the class values in the legend (see manual classification)
  - can also label classes as very low, low, medium, high and very high in the legend
    - can be used if relative values are more important than actual numerical values
  - changing the labels doesn't affect the data.
  - changing the values affects the classification! The classification values can be changed in the:
    Break values area
    Moving the break lines in the histogram
    Changing the number of classes
    Editing the values under the Range area of the symbology will change the classification
    Editing labels under the symbology tab or in the table of contents does not affect the classification

- Example maps were displayed with two classes (Red and Blue representing Republican and Democratic parties) with Presidential Election results to compare State and County election maps. The county map appeared more red, but is misleading in the sense that sparsely populated areas are displayed the same as densely populated areas.

- Example cartogram scales the area of a feature according to the value it represents. So states with more electoral votes appear larger than states with fewer electoral votes. A cartogram gives a more accurate representation of the distribution of electoral votes.

- The Purple map classifies counties into several classes ranging from red to blue with shades of purple in between based on the % democrat to republican vote to show how counties split the vote. So you can see where some of the battle ground areas might be where the map is more purple and less red or blue.

DEMONSTRATION:
Used the Lower 48 States and classified using Natural Breaks, Quantile and Geometrical Interval.
Used the POP_2010 attributes to classify and compare the three different classification methods.