Data Classification

One way maps make their point is by grouping data. The grouping, or classification, of data produces patterns. Different data classifications produce different patterns.

**qualitative classification**

Beagle  Great Dane  Terrier
Mutt  Dingo
Bulldog  Beagle
Chow  Wolf  Wild Dogs
Hyena  Chow  Feral Dogs
Beagle  Retriever  Domesticated Dogs
Wolf  Canines

**quantitative classification**

<table>
<thead>
<tr>
<th>Urban Population unclassified</th>
<th>Urban Population 5 numeric classes</th>
<th>Urban Population 3 ordinal classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>23,556</td>
<td>23,556</td>
<td>18,000 -</td>
</tr>
<tr>
<td>21,142</td>
<td>21,142</td>
<td>24,000 -</td>
</tr>
<tr>
<td>18,233</td>
<td>18,233</td>
<td>13,000 -</td>
</tr>
<tr>
<td>14,199</td>
<td>14,199</td>
<td>15,000 -</td>
</tr>
<tr>
<td>13,877</td>
<td>13,877</td>
<td>13,000 -</td>
</tr>
<tr>
<td>9,676</td>
<td>9,676</td>
<td>15,000 -</td>
</tr>
<tr>
<td>9,434</td>
<td>9,434</td>
<td>8,000 -</td>
</tr>
<tr>
<td>8,973</td>
<td>8,973</td>
<td>10,000 -</td>
</tr>
<tr>
<td>5,122</td>
<td>5,122</td>
<td>8,000 -</td>
</tr>
<tr>
<td>4,889</td>
<td>4,889</td>
<td>10,000 -</td>
</tr>
<tr>
<td>3,322</td>
<td>3,322</td>
<td>3,000 -</td>
</tr>
<tr>
<td>1,022</td>
<td>1,022</td>
<td>6,000 -</td>
</tr>
<tr>
<td>988</td>
<td>988</td>
<td>2,000 -</td>
</tr>
<tr>
<td>756</td>
<td>756</td>
<td></td>
</tr>
</tbody>
</table>
Data Classification: Qualitative Data

Data classification is shaped by your goals for your map. In general, features in the same class should be more similar than dissimilar; features in different classes should be more dissimilar than similar. Use color hue and shape and texture to symbolize different classes of qualitative data.

qualitative point data

Classification reveals patterns that are difficult to see in unclassified data. Students poll community members about social issues to learn about community politics. The bottom left classification is not very revealing. The bottom right classification reveals more about the political landscape. Include the unclassified data so map viewers can decide if your classification is justified.

Unclassified:

Most Important Social Issue

Poor classification:

Community Social Concerns

Good classification:

Predicted Party Affiliation

- Family Values
- Religious Values
- Legal Values
- Social Welfare Issues

Democrat
Republican

qualitative line data

Roads are often classified in terms of who builds and maintains them (federal, state, local). However, this classification is not the best if your map is for tourists. Your goal for the map (tourism) should shape classification (tourism-based classes of roads).

Unclassified:

Tourist Roads

Poor classification:

Tourist Roads

Good Classification:

- Federal Highway
- State Highway
- Local Road
- Fast Travel
- Historical Sites
- Natural Beauty
qualitative area data

Criteria external to data can serve as the basis of classification. When classifying soils, geology, or land-use data, for example, follow professional or academic classification standards. If standards vary, note which standards you chose on the map. A map created for a city planning department should use the department’s land-use classification. While the map on the left may be OK if you are looking for an apartment or house, the map on the right follows land-use classification conventions.

Poor classification:

- houses
- apartments
- companies

Good classification:

- residential
- industrial
- commercial

Be aware that most qualitative data on maps has been classified. Attempt to determine the criteria upon which the classification decisions were made. Be wary of maps with no logical criteria. Conversely, a map may classify data based upon criteria suitable for one purpose but not necessarily for others.

USGS 1:100,000 Topographic Map

U.S. Geologic Survey (USGS) topographic maps classify vegetation into two categories: vegetation (grey areas on this map) or no vegetation. A user of this map may believe the classification to be based on ecological criteria.

However, vegetative areas are classified based on military criteria: vegetative areas are areas with tree cover at least 6 feet tall which can hide military troops. It is a classification for army guys, not ecologists!
Data Classification: Quantitative Data

Quantitative data classification requires decisions about the number of classes and class boundaries. Quantitative data classification is shaped by external criteria or by the characteristics of the data themselves. Use color value and size to symbolize different classes of quantitative data.

**quantitative point data**

Data classification generates spatial patterns in the data (which may not be evident in the unclassified data). For a map created for a community meeting, the classification on the right (bottom) is better, as it shows what is most important: whether the well water is safe for humans or not.

Unclassified:

**Toxins in Wells (ppm)**

- over .0010
- .0006 to .0009
- under .0005

**Poor classification:**

**Toxins in Wells (ppm)**

- exceeds federal limits (over .0005 ppm); Action: close well
- safe for non-human use (.0000 to .0004 ppm)
- safe for all uses (below .0001 ppm)

**Good classification:**

**Toxins in Wells (ppm)**

**quantitative line data**

A map intended to help guide the restructuring of police patrol routes should classify data, in this case average vehicle speeds, in categories appropriate to the task: increase, maintain, or decrease patrols (suggesting a three-class map).

Unclassified:

**Vehicle Speeds**

- over posted speed limit

**Poor classification:**

**Vehicle Speeds**

**Amount over posted speed**

- over 15 mph
  - Action: increase patrols
- 11 to 14 mph
  - Action: increase patrols
- 5 to 10 mph
  - Action: maintain patrols
- 0 to 4 mph
  - Action: decrease patrols
- below 0 mph
  - Action: decrease patrols

**Good Classification:**

**Vehicle Speeds**

- over 10 mph
  - Action: increase patrols
- 0 to 9 mph
  - Action: maintain patrols
- under posted speed
  - Action: decrease patrols
quantitative area data

You must decide the number of classes. Fewer classes often result in distinct patterns; more classes often result in complex patterns. Which option is best depends on why you are making the map. This map shows the density of mobile homes (dark = higher density).

When choosing the number of classes consider that:

✓ A 2 class map can be suitable for binary (yes/no) data or data with negative and positive values.

✓ 4 to 8 classes usually ensures that map readers can see distinct patterns and match a particular shading on the map to the legend. A good default.

✓ More than 8 classes will produce more complex patterns, but map readers may not be able to match a particular shading on the map to the legend.

✓ Unclassified data (each area has a unique shading corresponding to its unique value) produce the most complex and ungeneralized patterns.

✓ Classify your data using different numbers of classes, and look at how patterns in your data vary. Think about your data and your goals for your map, and make an intelligent decision.
quantitative area data

In addition to determining the number of classes, you must decide where to place boundaries between the classes. Classification schemes set these boundaries. This map shows the density of mobile homes (dark = higher density) unclassified and as a 3-class map using four different classification schemes.

thinking drives classification

Poverty is a contentious issue. Debates rage over defining poverty, why it exists, and how to address it. The U.S. Census Bureau provides official data on poverty in the U.S. and different classifications of Census 2000 poverty data follows.

It is easy to get the percent of people in each county in the U.S. who live in a state of official poverty. But choosing how to map the data is not as easy. Common (and equally valid) data classification schemes – methods for placing boundaries between the classes on a map – are easy to generate but difficult to choose from. Understanding the benefits and problems with each classification scheme is vital, as is clarifying why you are making the map. Together, these guide the thinking behind choosing the most appropriate classification scheme for your data.

graphing data

Selecting a classification scheme without examining your data as a graph is a bad idea. As examples in this section reveal, classification schemes can mask important characteristics of your data and perhaps undermine the goal of your map. A simple histogram can be constructed from your data: the x-axis is your data variable (from low to high) and the y-axis the number of occurrences of each value:

The 2000 U.S. poverty data have a cluster of counties near the lower to mid-end of the graph, with a smaller number of counties skewed out to 57%. You can easily note where a classification scheme places class boundaries, which values are grouped together, and which values are in different groups. If a particular classification scheme seems to violate the basic classification rule (features in the same class should be more similar than dissimilar; features in different classes should be more dissimilar than similar), then consider a different classification scheme. Consider placing the graph on your final map, so map users can see how the data are classified.
unclassified scheme

To create an unclassified scheme assign a unique visual shade to every unique data value. In essence, each unique data value is in its own class. Unclassified schemes produce complex, highly variable, and subtle patterns by minimizing the amount of data generalization and simplification.

This map, due in large part to the concentration of counties near the lower end of the range of values, suggests that poverty is not a significant issue in most places, that the number of people living in poverty is somewhat similar across the U.S., and that there are few places with very high poverty.

Quantile scheme

Quantile schemes place the same number of data values in each class. Quantile schemes are attractive in that they always produce distinct map patterns: a quantile classification will never have empty classes, or classes with only a few or too many values. Quantile schemes look great. The problem with quantile schemes is that they often place similar values in different classes or very different values in the same class.

The map suggests that poverty is a significant issue in many counties, and the numerous counties in the top, darkest classes impart a rather ominous view of poverty in the United States.
equal-interval scheme

Equal-interval schemes place boundaries between classes at regular (equal) intervals. Equal-interval schemes are easily interpreted by map readers, and are particularly useful for comparing a series of maps (which necessitates a common classification scheme). Equal-interval schemes do not account for data distribution, and may result in most data values falling into one or two classes, or classes with no values.

The map suggests that poverty is not an issue in most places, as there are few counties in the highest three classes.

natural-breaks scheme

Natural-breaks schemes minimize differences between values within classes and maximize differences between values in different classes. Class boundaries can be determined subjectively, by graphing data and looking for "natural breaks" in the data distribution. Natural breaks are determined by algorithms which seek statistically significant groupings in a set of data. Natural-breaks schemes can serve as a default classification scheme, as they take careful account of the characteristics of the data distribution.

The map makes poverty seem more significant than the equal-interval map does, but is not quite as ominous as the quantile scheme.
unique scheme

Class boundaries sometimes need to be set in accordance with external criteria: immigrant population by state is classified in terms of federal requirements for applying for immigrant-related grants, for example. Unique schemes require an understanding of the broader context of the data.

The map below is made for a study of counties with very high poverty. The researchers are not interested in any county with less than 25% of the population in poverty. The rest of the data are divided into groups which will aid in the analysis of high-poverty counties. While excellent for the study, this map and classification is not good in general, as it suggests poverty is isolated and rare.

Think about what you are seeing: Looking at a number of schemes brings to the foreground geographic facts, both the facts that are variously emphasized and those that are preserved through every variation.

The unique scheme displays counties with the highest rates of poverty. While this can be seen in every scheme, the unique scheme isolates, and so draws attention to, regions of the country where poverty has been a long-standing reality: the coal-mining heart of Appalachia; the old “Cotton Belt,” specifically those counties with the highest proportion of slaves at the time the Civil War began; the counties in the Rio Grande Valley, where the population is overwhelmingly Hispanic; and the reservations of the Navajo, Lakota, Sioux, and other Native Americans. The unique scheme picked out the regions of historically outstanding social injustice.

Unless you looked at a lot of maps, you might not have identified these regions as anything other than those with high levels of poverty. It takes many different kinds of maps to begin to make sense of the world.
more information...


Two excellent guides to graphs are William Cleveland's The Elements of Graphing Data (Wadsworth, 1985) and Stephen Kosslyn's Elements of Graph Design (Freeman, 1994).

In Information Anxiety (Bantam, 1989) Richard Wurman illustrates classification graphically in three pages about the 130 breeds of dogs recognized by the American Kennel Club: well worth taking a look at.

A curious tale about what happens when you don't generalize is Jorge Luis Borges' short story "Funes the Memorious" (in Labyrinths: Selected Stories and Other Writings, New Directions, 1964).

Sources: Bill Bunge's "Continents and Islands of Mankind" is re-created from his Field Notes: Discussion Paper No. 2 (self published, no date). Much of this chapter is based on Dent (1990) and Slocum et al. (2003).