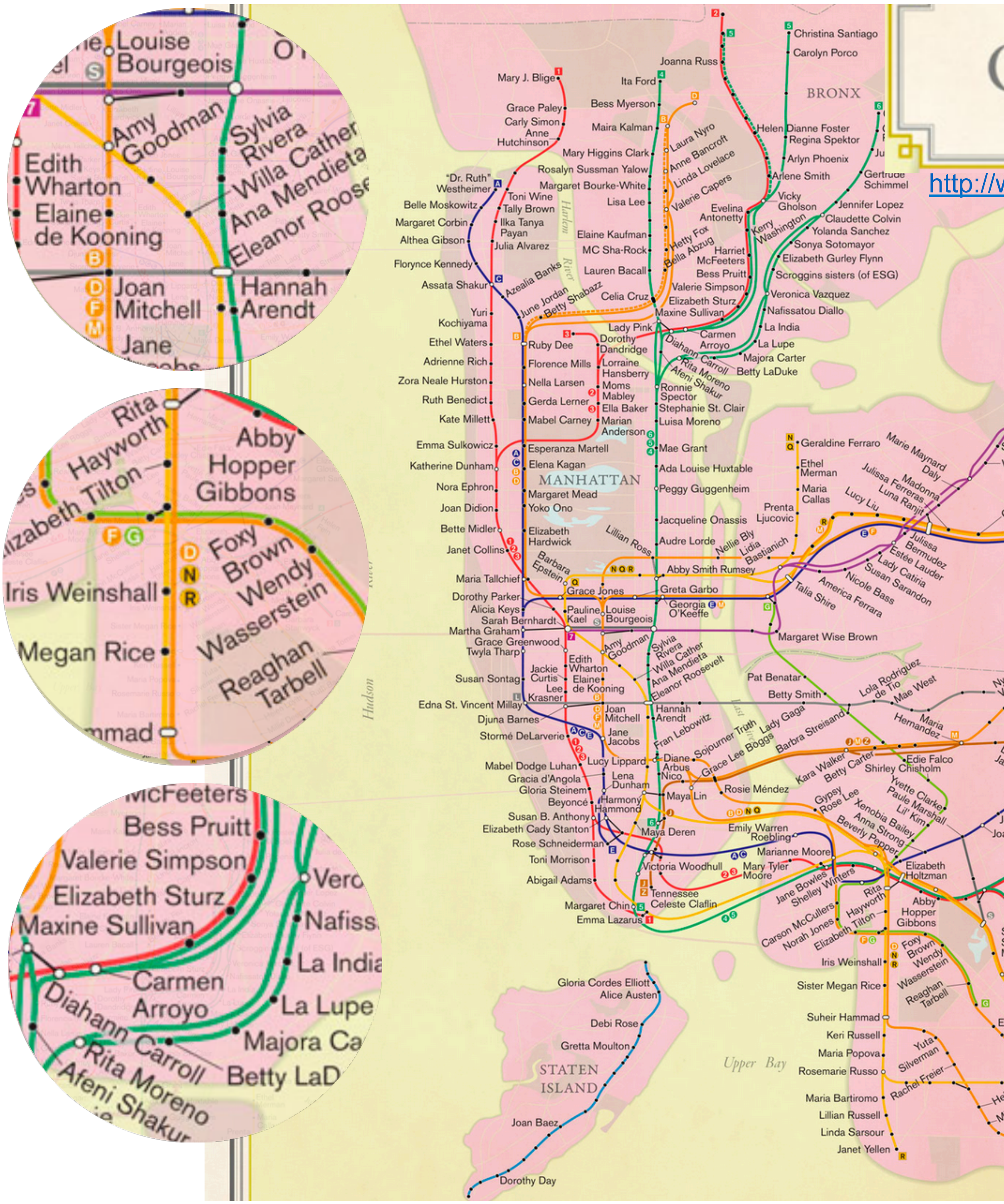
CITY OF WOMEN

http://www.newyorker.com/books/page-turner/city-of-won



What if the New York City subway map paid homage to some of the city's great women? (Hover over the map to magnify.)

CARTOGRAPHY BY MOLLY ROY, FROM "NONSTOP METROPOLIS," BY REBECCA SOLNIT AND JOSHUA JELLY-SCHAPIRO. SUBWAY ROUTE SYMBOLS ® METROPOLITAN TRANSPORTATION AUTHORITY

Geography 360
October 14, 2016

Interpreting data through classification and Mapping data: intensity vs. count

1. *Questions and Demos!*
   - Demoing how to symbolize a layer that doesn't appear to have that option.
2. *Classification (continued.):* Interpretatively simplifying data for clarity and comparison.
3. *Mapping intensity data vs. count data.*
   When do you use choropleth maps?
   When do you proportional symbol maps?

Classification of data

- …is *interpretatively* simplifying data by grouping it into classes.
- You then represent each class distinctively with a visual variable, instead of representing the whole range of original values.

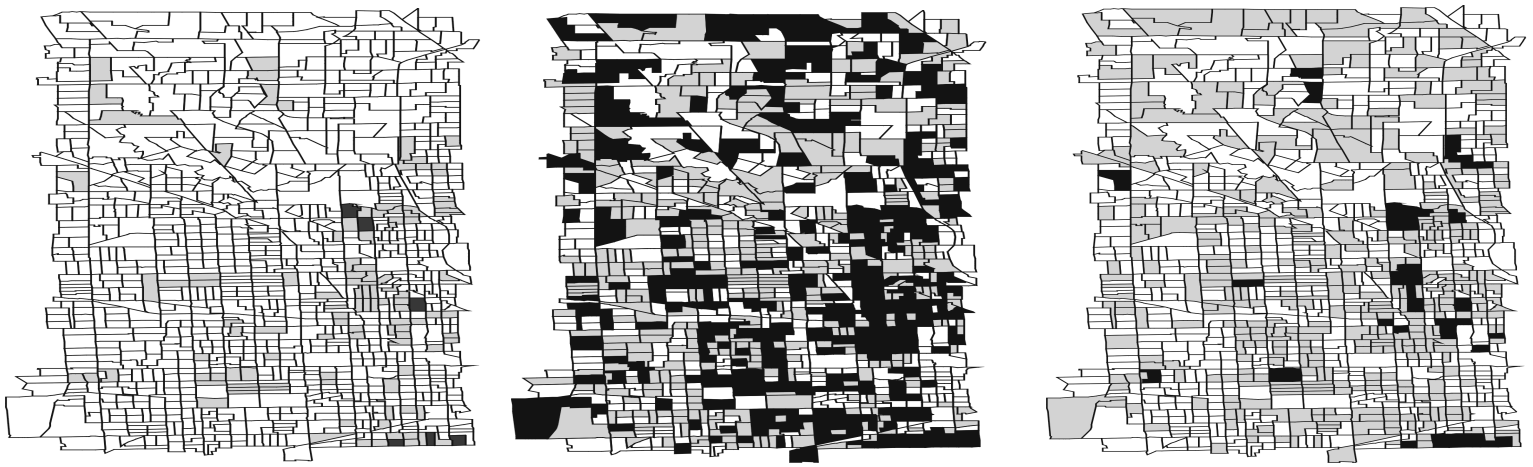
Why classify data?

- For readability.
- To highlight spatial relationships and patterns, especially across different parts of the map.
- To have the visualization better indicate meaningful divisions in the data (e.g., above and below average.)
- To have the visualization better indicate meaningful divisions in the process/phenomena being measured (e.g., above and below sea level.)

A key question to ponder when deciding to classify: By sacrificing detail in this data and map, am I nonetheless able to communicate information more clearly?

Classification methods include:

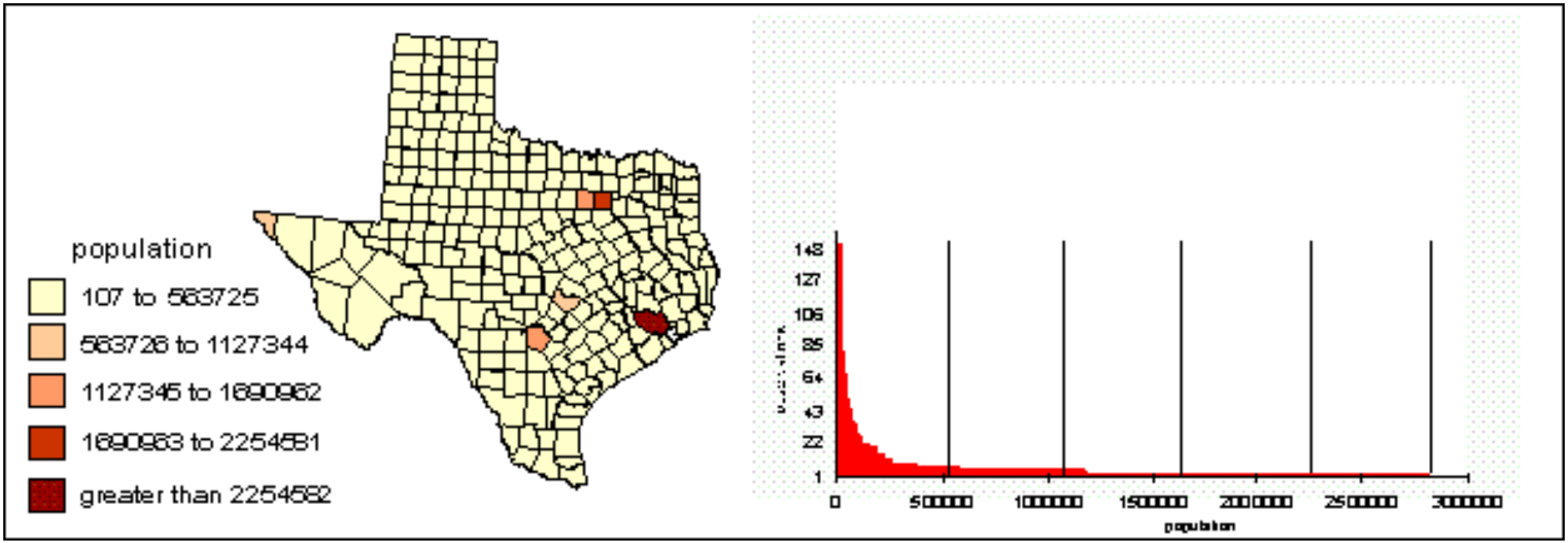
- Equal interval
- Equal area
- Quantiles
- Mean standard deviation
- [Variations on] Natural breaks
- *Note:* You can elect to modify the results of one of the classification methods to reflect knowledge you have about the underlying data or process that generated it (e.g. setting a break to differentiate above/below sea level.)



Three perspectives on the same data

Equal interval

- Class breaks are chosen so that the range of values within the classes is constant:
  - i.e. 0-100, 100-200, etc.
- What levels of measurement in the data might you be able to retain/express when using this classification?
  - Up to interval, generally.
  - You can manage ratio if you are careful in how you set the lower/upper bounds of the lowest/highest classes and the symbols chosen for the lowest/highest classes..



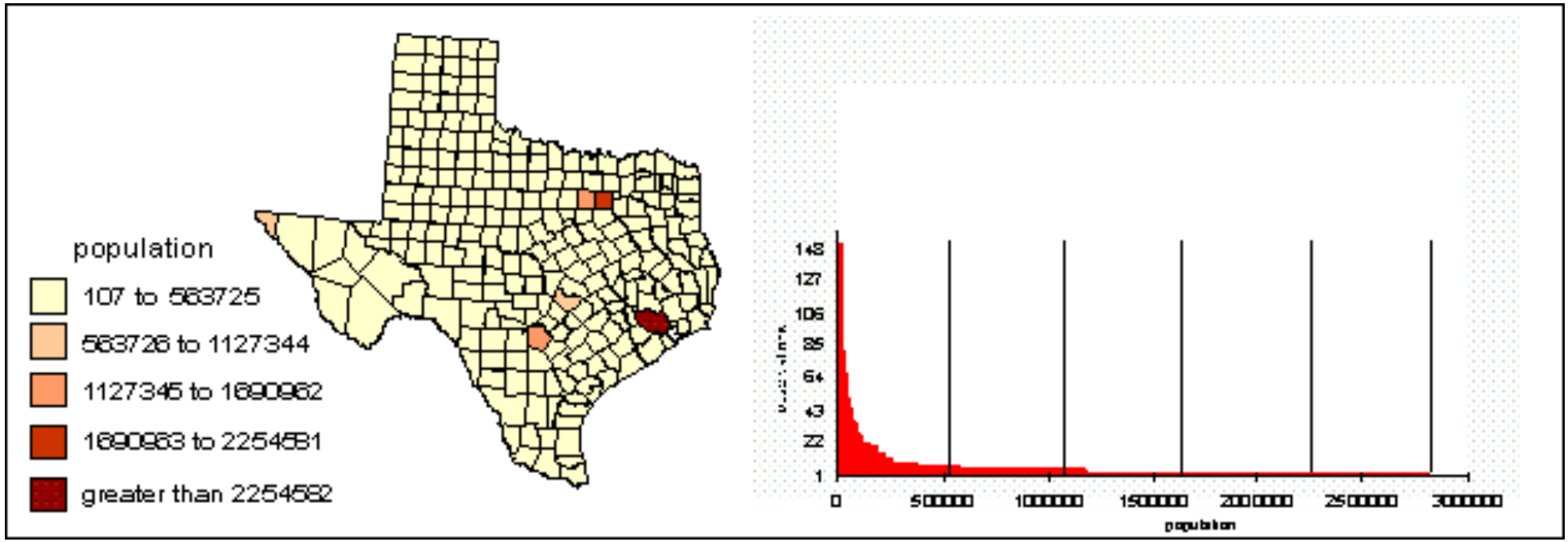
Equal interval

Pros:

- Easy to compute, easy to interpret classes (though sometimes not the map)
- Can be used to make a series of maps comparable

Cons:

- Does not consider data distribution (e.g., can break up clusters)
- Some classes may be empty



Quantiles

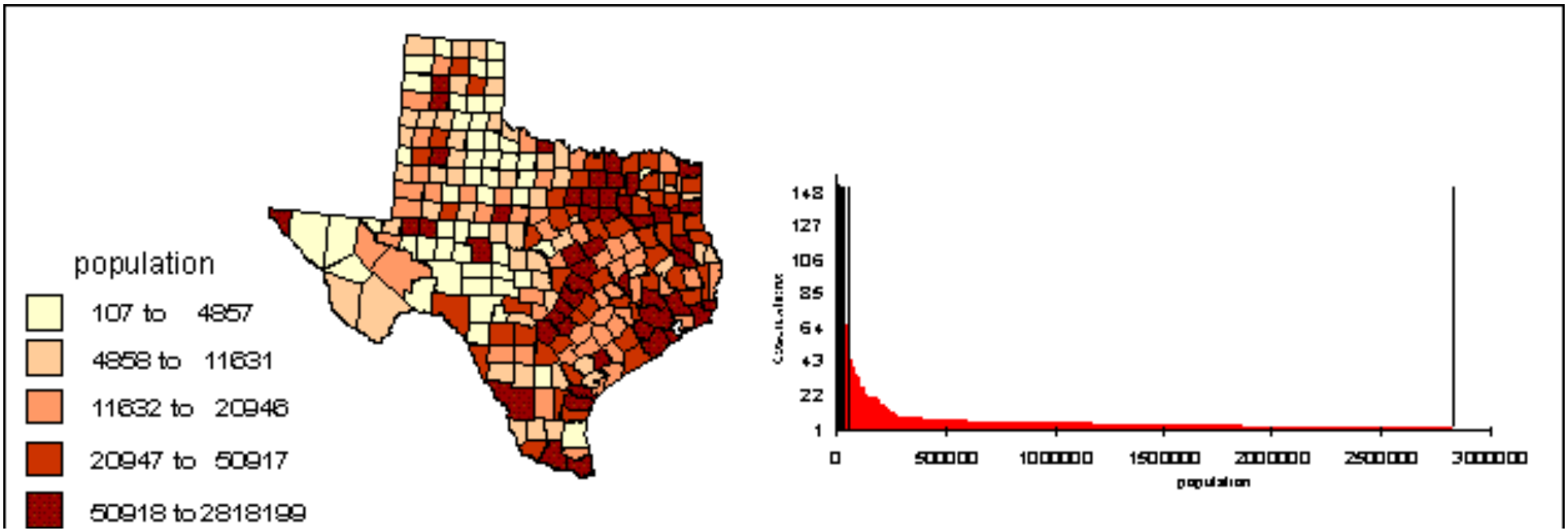
Have equal **numbers** of observations (polygons) in each class.

Pros:

- Easy to compute class limits
- If enumeration units similar in size, each class will cover about the same area (gives a sense of balance)
- Works for ordinal data because it only uses the relative ordering of the data

Cons:

- Does not consider data distribution (e.g., can break up clusters)
- Categories may have odd ranges



Mean standard deviation

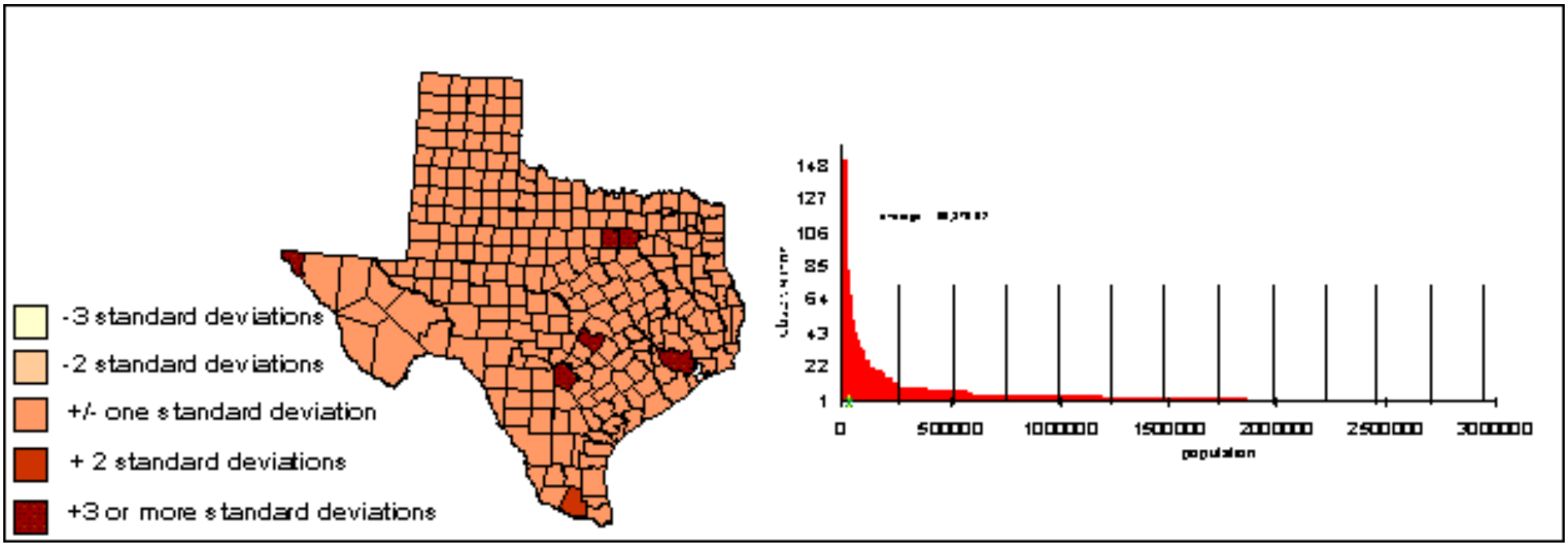
Classes determined by standard deviation (classes are 'sigma' above and below the mean)

Pros:

- Constant class intervals
- Yet does consider the distribution of data, in some senses (e.g., mean can be a useful dividing point for some datasets).

Cons:

- Most appropriate for data that follow a normal distribution (thus, data whose histogram have a single, symmetric peak)
- Readers must also understand basic statistics.



[Manual] Natural breaks

- You use the histogram to find natural groupings of data
- You often are choosing the breaks to minimize differences between data values placed in same class.

- Preserves clusters & natural groups
- Classification will fit data distribution
- Decisions about breaks can be subjective and may vary from person to person

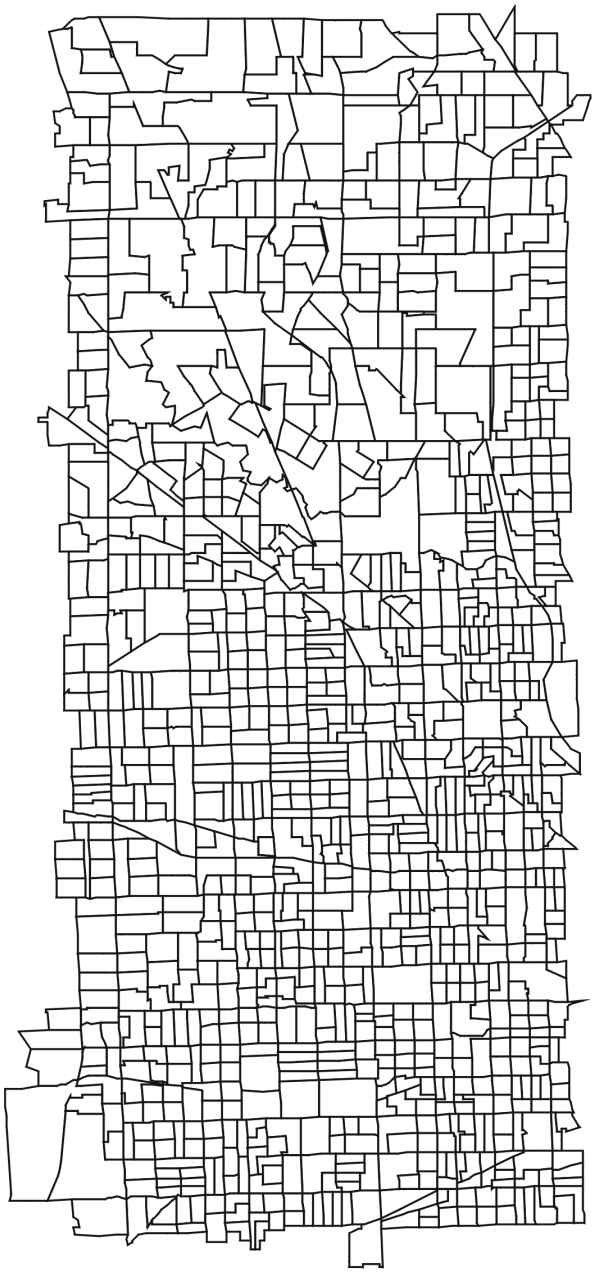
[Algorithmic] Optimized Natural breaks

- Considers variance in the data
- Selects breaks that minimize total variance within classes and maximizes distance between classes

- A certain kind of “optimal”…
- But… can be complicated to meaningfully interpret.

In choosing and applying a classification technique, think about:

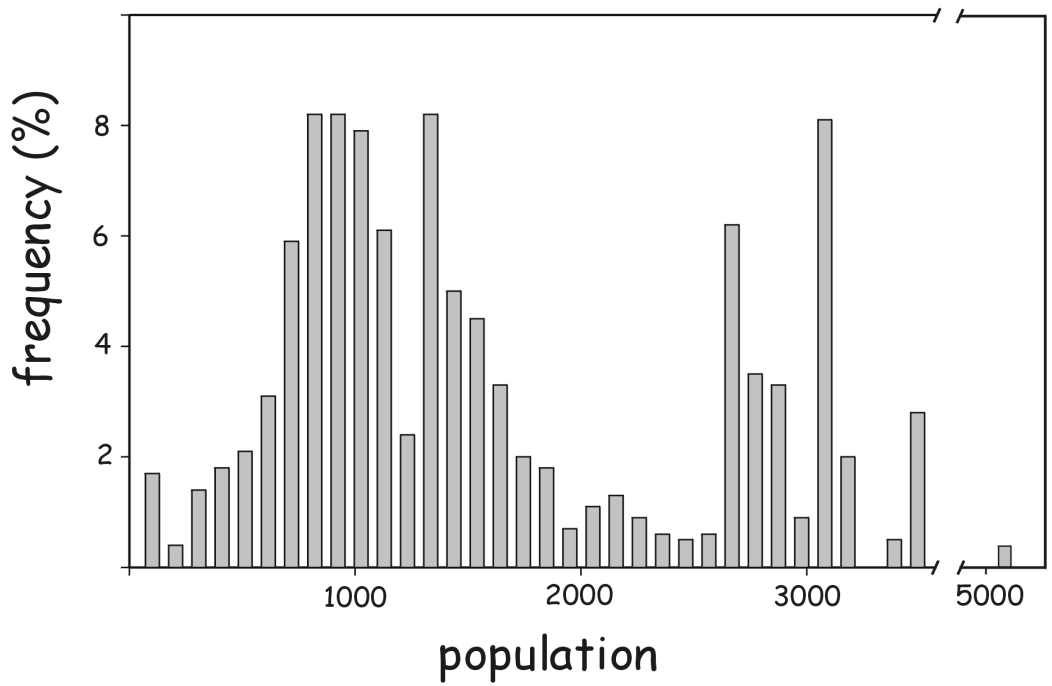
- Data type (level of measurement) and distribution.
- Is the {method, resulting legend, resulting map} easy to understand (for you and for the readers)?
- How many number of classes do you want to use?
- Is there more than one map that you want to compare?
  - For most of these methods, you can pool/merge the data from several datasets into a temporary dataset, calculate the breaks, and then manually apply them in the same way to all the maps.
- Do you want to modify the results of one of the classification methods to *reflect knowledge you have about either the underlying data or phenomenon/process that generated it?*
  - e.g., by setting a break to differentiate above/below sea level

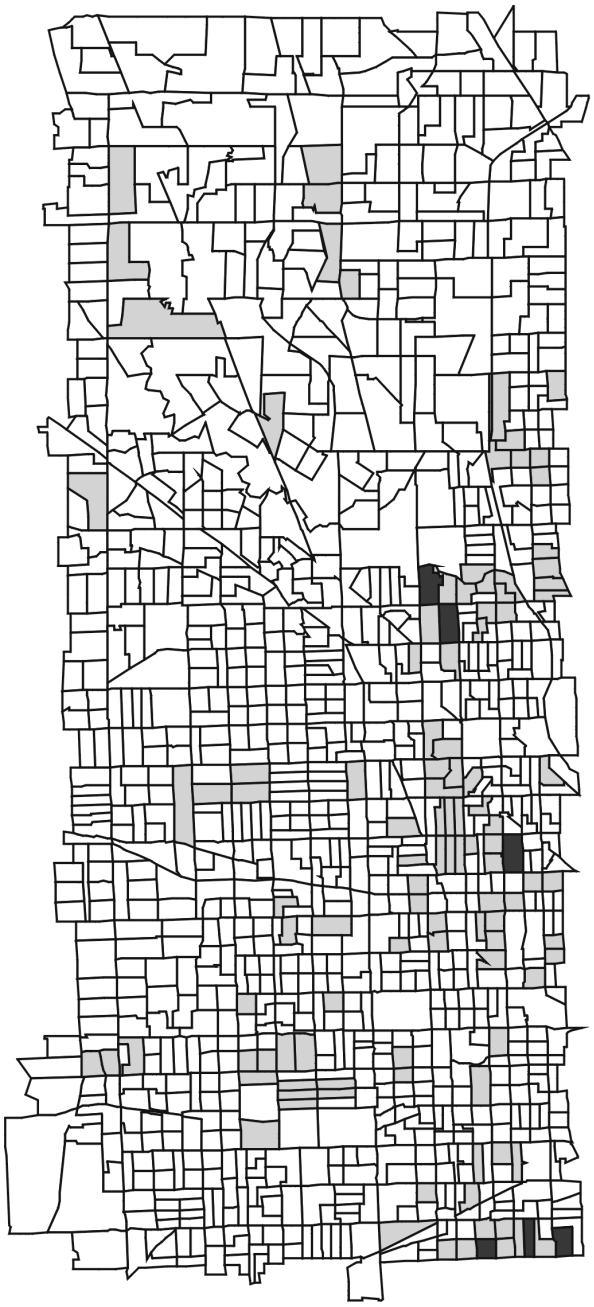


Neighborhoods

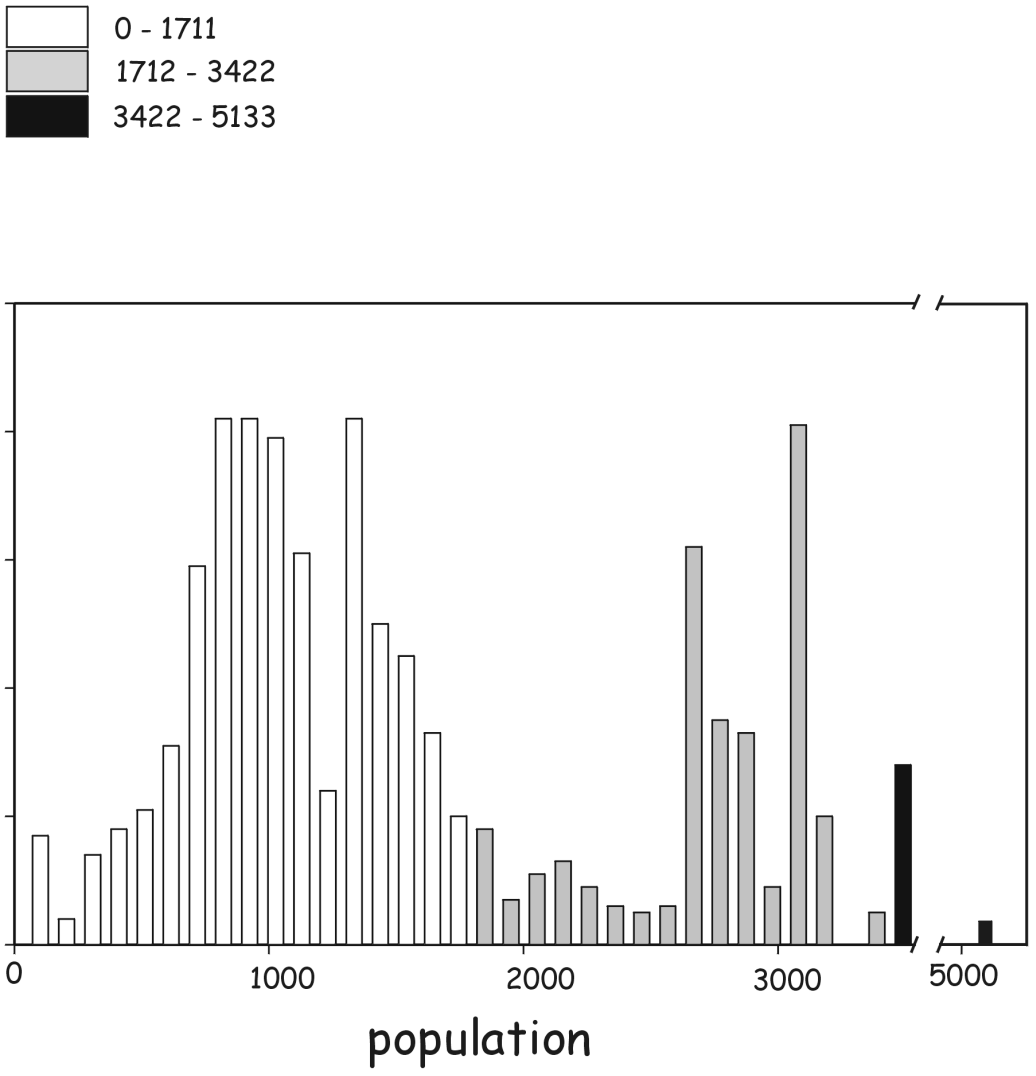
- 1074 polygons
- population for neighborhoods
  ranges from 0 to 5133 (3 outliers > 3300)

Bar graph shows frequency of neighborhood population, e.g., there are 84 neighborhoods with a population between 3000 and 3100



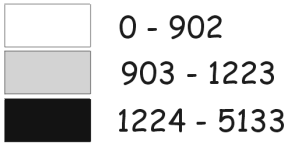


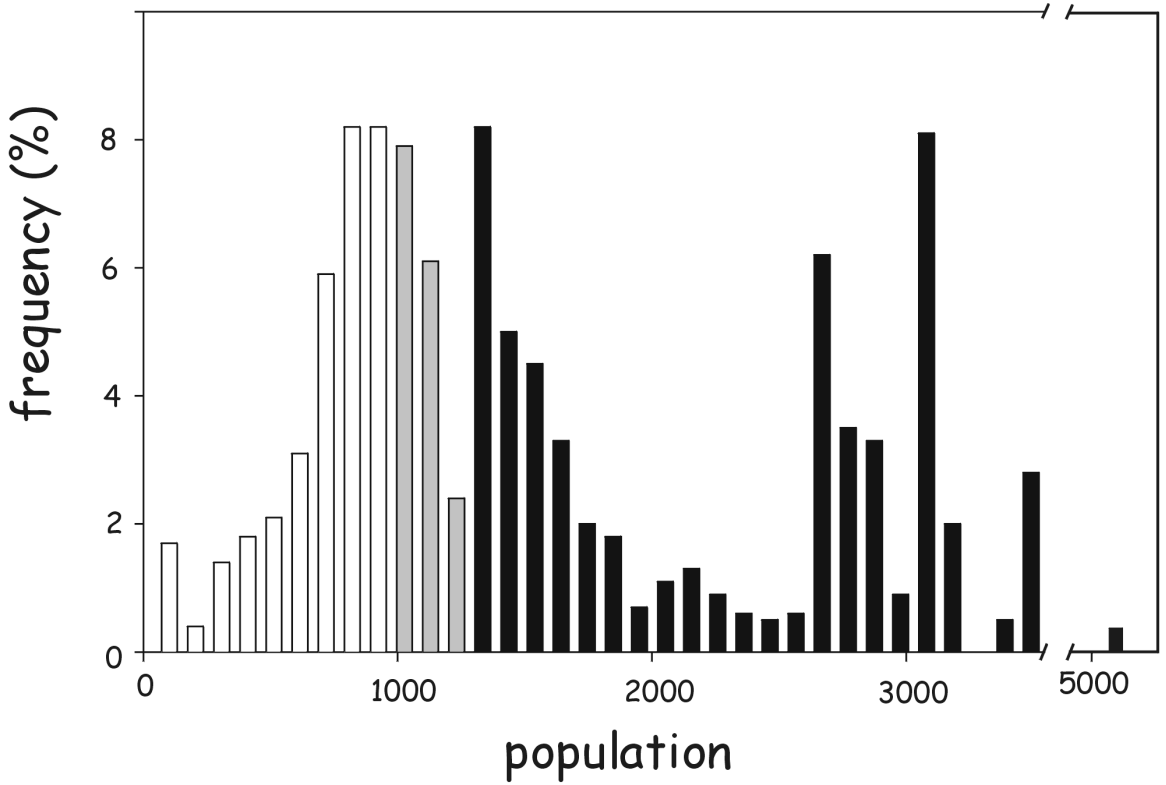
Equal-interval classification



Bolstad 2016

Equal-area classification

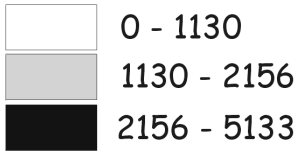


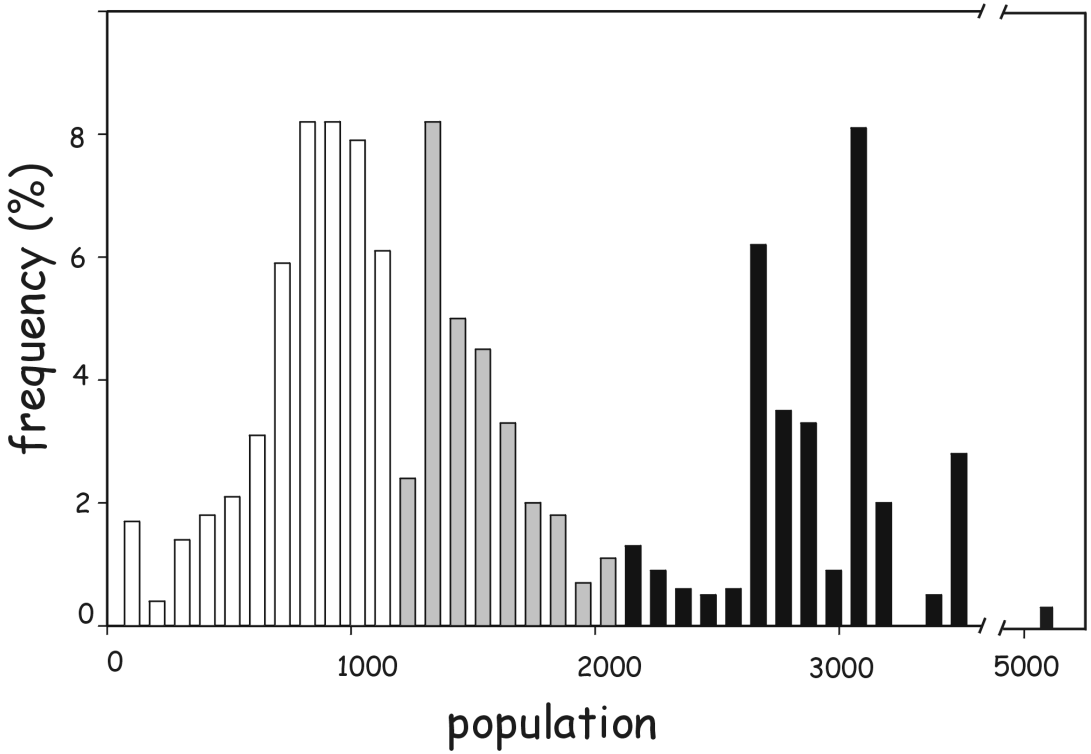


Bolstad 2016

Bolstad 2016

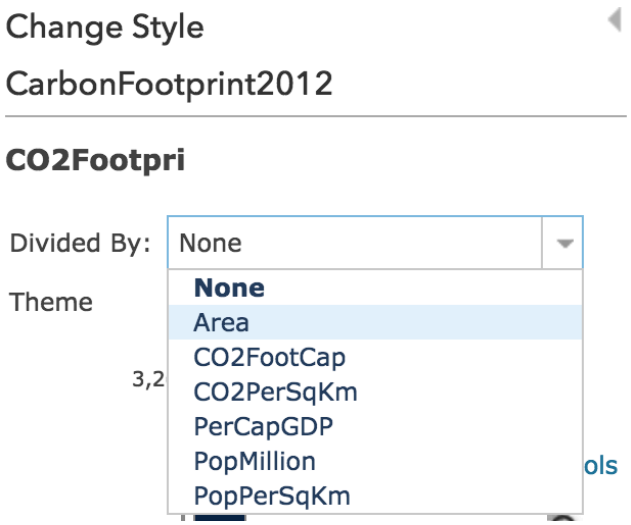
Natural breaks classification





Mapping: Intensities vs. Counts

- *Standardizing* data is dividing it by a quantity.
- Choropleth maps are best employed where *count data* can be *standardized* into being *intensities.*
  - Not people…
    but people per square mile.
  - Not total carbon emissions across a territory…
    but carbon emissions per person.
  - Not total employment in the service industries…
    but service industry employment
    as a percentage of total employment in all sectors.



- If there is a reason to display count data without standardizing them, then consider *proportional symbol maps*.