10 Mapping

36-721 Statistical Graphics and Visualization

Jerzy Wieczorek

10/1/15

Last time, etc

- Graphics useful for statistical analysis
- Critique due today in class
- Major League Data Challenge: consider for Project 2?

Today

- Mapping principles, map projections, etc.
- Choosing breaks for color scales: equal intervals, quantiles, Jenks
- Mapping points and filled areas, using `sp` and `ggplot2`
- Adding context maps with `ggmap`
- Some special maps: Linked micromaps, cartograms, heatmaps, statebins/hexbins

Follow along

- Editable code in `10_Mapping_code.R`
- Code with output examples in `10_Mapping_code.html`
- Download 2010 Census Tract level shapefiles (Web interface or FTP site) and unzip before using:
  - `tl_2010_30_tract10.zip` (Montana)
  - `tl_2010_42_tract10.zip` (Pennsylvania)
- Data are a subset of tract-level 2015 Planning Database:
  - `censusTopVars.Rdata` has all tracts in US, but just a recommended subset of variables
  - `censusMT.Rdata` and `censusPA.Rdata` have just these subset variables for all tracts in just those states

Choropleth maps

Choropleths, also known as thematic maps: Color or shading shows statistical data on previously-defined regions (e.g. states, countries), not on regions derived from the data.

- Best used for densities or rates, not totals
- Limit number of color classes; higher data values should be darker
- May mislead: large but low-value areas may stand out more than small but high-value areas
- Mapping uncertainty is difficult; examples by Cornell PAD

Datums and map projections

How to model the Earth: as sphere, ellipsoid, or more accurate shape? Different datums may have different latitude-longitude coordinates, sometimes by a few hundred meters.

How to display 3D Earth's surface on a 2D map? Every projection distorts somehow; must choose whether to preserve angle, area, or distance at expense of others.

- Albers Equal-Area Conic: often used for US and its states; good for choropleths showing data per unit area; advice on choosing parameters
- Web Mercator: used by Google Maps etc.; "north is up everywhere, meridians are equally spaced vertical lines, but areas near the poles are greatly exaggerated"

Map data formats and sources

Shapefile: vector data about your points, lines, or polygons, with associated attributes (name, land area, etc.) Actually not one file but "a collection of files with a common filename prefix, stored in the same directory."

Sources for shapefiles of countries, regions, etc:

- Global: Global Administrative Areas / GADM
- US: Census Bureau TIGER Shapefiles; see also hierarchy of Census geographies

Map data concerns

- **Vintage:** e.g. 2000 Census Tracts differ from 2010 Tracts
- **Generalized** or simplified boundaries: look cleaner at small scale (distant zoom), file size is smaller, and plots faster; but lower resolution and accuracy.
  MapShaper can generalize shapefiles interactively
- **Harmonized** geography: neighboring polygons' shared boundaries are harmonized to match exactly.

Statistical concerns when mapping

If your map regions / level of aggregation are not fixed in advance, beware of Modifiable Areal Unit Problem (MAUP). Different aggregations (e.g. tract vs county vs state) can suggest very different trends in the data.

Beware of Ecological Fallacy: pattern across areas may not be same as pattern for individuals within areas. Example: voter-level vs state-level patterns.

See also Mark Monmonier's *How to Lie with Maps*, the source of the next few slides' figures (highly recommended as an intro to spatial data, its provenance, and its representation)

Statistical concerns when mapping

For comparing two variables, side-by-side maps can mislead.
Use a scatterplot.

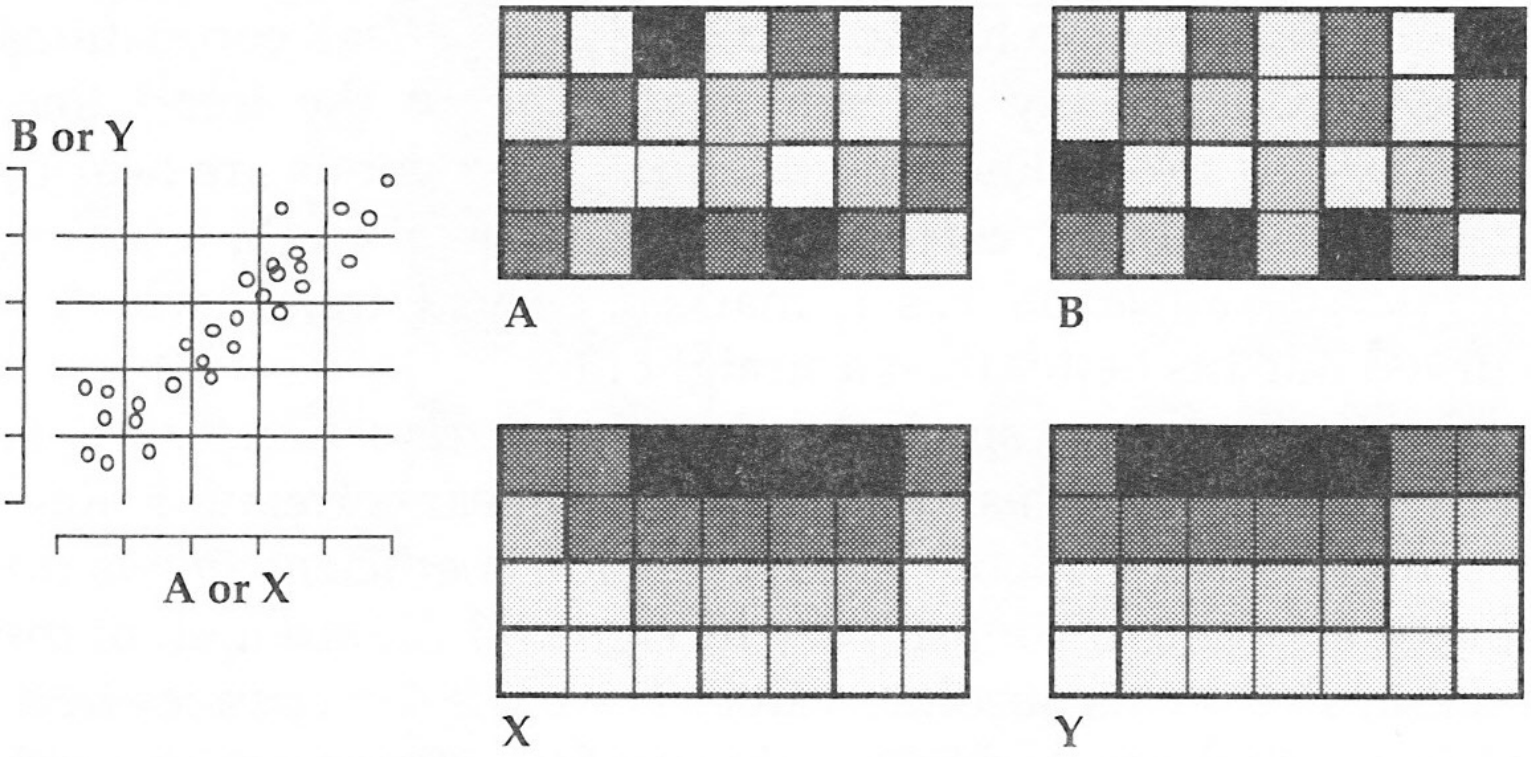

FIGURE 10.17. Two pairs of variables with identical scatterplots, correlation coefficients ($r = .93$), and class breaks, yet distinctly different map patterns.

Monmonier, *How to Lie with Maps*

Data classes for color scales

Choropleth maps often show continuous data using a small number of colors. How to cut the data range into a few discrete classes?

- Equal intervals (each class has same width) with `cut`
- Equal quantiles (each class has same number of points) with `quantile`
- Jenks Natural Breaks (classes are well-separated clusters) with `classIntervals(style = 'jenks')` in `classInt` package

Data classes for color scales

Equal intervals or quantiles

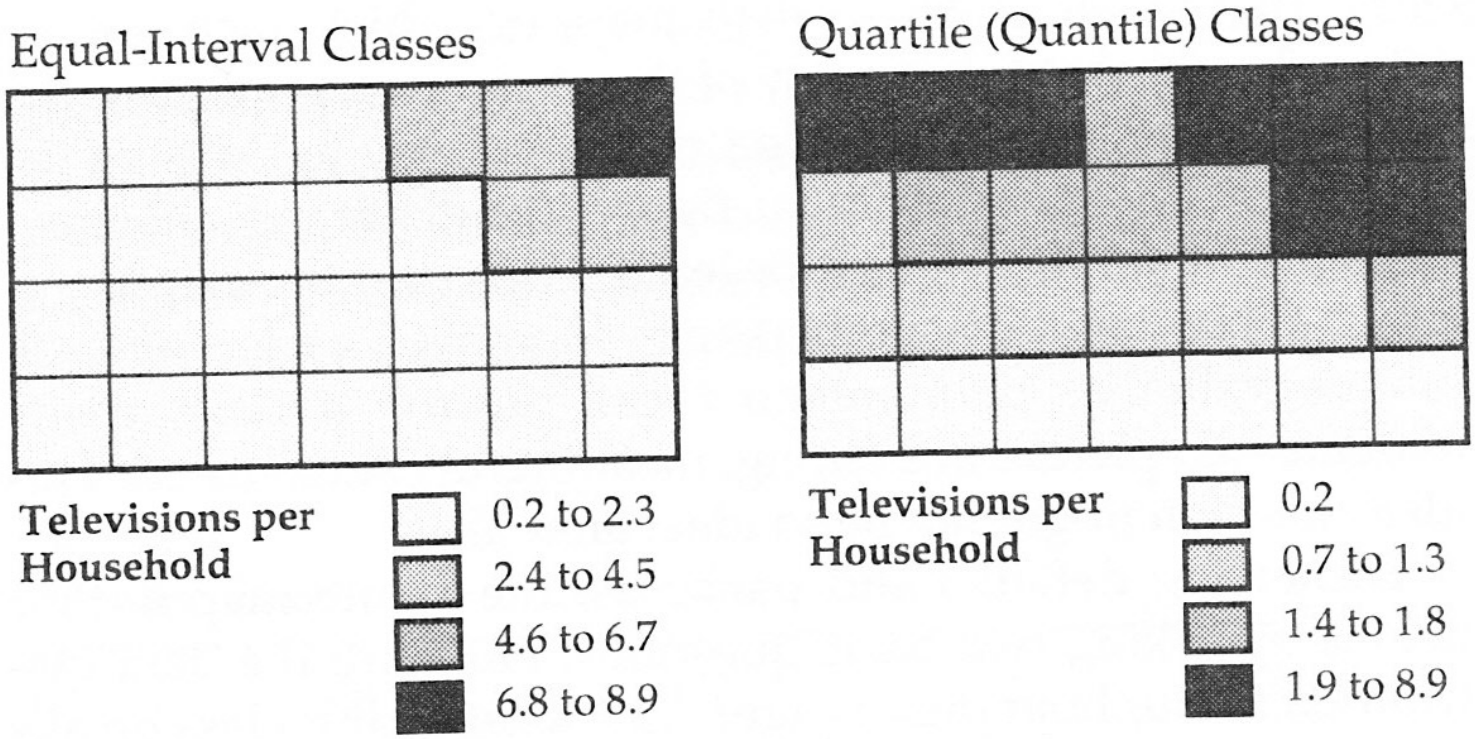

FIGURE 10.7. Two common classing schemes used as "defaults" by choropleth mapping software yield radically different four-category patterns for the data in figure 10.4.

Data classes for color scales

Actual distribution of the data

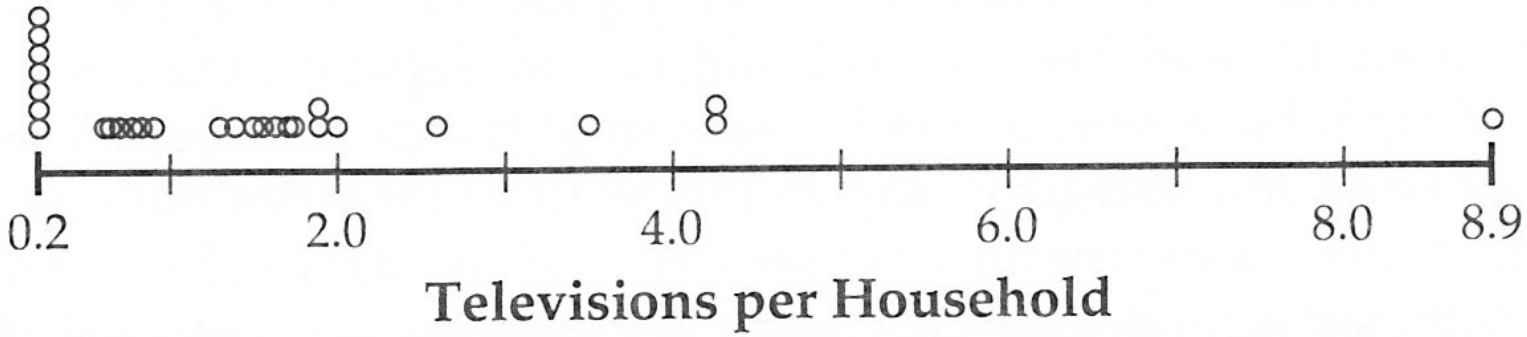

FIGURE 10.9. Number line for the town-level television-ownership rates in figure 10.4.

Data classes for color scales

Natural, well-separated classes for the same dataset

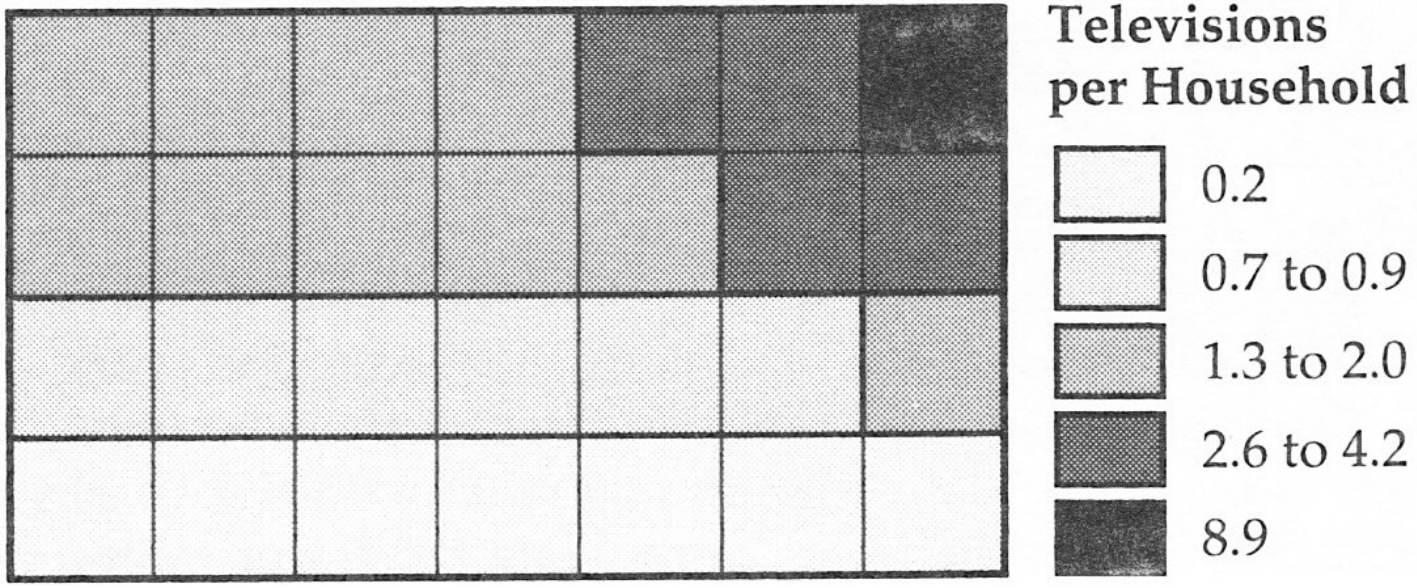

FIGURE 10.10. Choropleth map based on the number line in figure 10.9 and the character of the data.

Map examples with `sp`, `ggplot2`, `ggmap`

Pre-processing for a choropleth map:

- Read in the shapefile, with your map's boundaries and basic attribute table
- Ensure the shapefile's attribute table and your other dataset have a compatible variable, containing a unique ID for each area
- Merge extra variables from other dataset into the attribute table

Map examples with `sp`, `ggplot2`, `ggmap`

Useful packages:

- `sp` and `rgdal` read in shapefiles, project them, and plot them
- `ggplot2` needs `rgdal` or another package to read the shapefile into R first; then `fortify` it into a format that `ggplot` understands; then plot it with `coord_equal` or `coord_map`
- `ggmap` (extends `ggplot2`) downloads and plots basemap layers for context (roads, terrain, etc) from Google Maps, Open Street Maps, etc.

Specifying projections in R

- For `spplot` in `sp`, specify using PROJ.4 parameters; see examples in my code
- For `coord_map` in `ggplot2`, specify using `mapproject` parameters; see examples at docs.ggplot2.org

General mapping in R: tutorials and resources

- sp graphics example figures
- spatial.ly, Introduction to Spatial Data and ggplot2
- Barry Rowlingson, Geospatial Data in R and Beyond
- Bivand et al., *Applied Spatial Data Analysis with R* (Springer, Amazon)
- CRAN, Spatial task view (lists the most widely-used R packages for spatial data)

Linked micromaps

Developed by Dan Carr to show more information (e.g. a time series) for each area

- Example and overview
- Base R code
- R package `micromap` for use with `ggplot2`

Cartograms

Cartogram: a map with areas distorted so that their sizes are proportional to some data variable

- R package `getcartr`: examples, GitHub

Heatmaps

Not geographic maps, but share some concerns with choropleths: choice of data classes for coloring, and how to show uncertainty

- FlowingData, How to Make a Heatmap
- Use and misuse of the term: Cartonerd, When is a heat map not a heat map

Statebins and hexbins

Heatmaps for US states, with approximately-correct spatial arrangement: like a choropleth where larger states don't overshadow smaller ones

- Statebins in R
- Hexbins in R

Exercises

- Redraw the example-code maps for Pennsylvania data
- Use `ggmap` to show different basemaps (like Stamen Maps instead of Google Maps);
  see Kahle and Wickham (2013) paper on `ggmap`
- Try making a linked micromap, cartogram, heatmap, or statebin/hexbin map

For next time

- Tues 10/6: **no class**
- Thurs 10/8: high-dimensional data; install GGobi to follow along
- Sat 10/10: Project 2 due
- Tues 10/13: networks and trees
- Thurs 10/15: TBD
- Sat 10/17: Project 3 due
- Sat 10/24: final resubmissions due