Chapter 17: Geospatial Data

- Data containing spatial or geographic coordinates like latitude and longitude are a type of *spatial data*.
- For example, the `airports` table in Chapter 15 contained latitude and longitude values.
- But more complicated data files called *shapefiles* have structure that encodes spatial information.
- Certain packages are designed to work with shapefiles and produce visualizations of *geospatial data* that are accurate rather than misleading.

John Snow and the Cholera Data Set

- A very early and historically important use of geospatial statistics dates to 1854, when there was an outbreak of the disease *cholera* in London.
- A physician, John Snow, plotted on a map the locations of the addresses of people who had died of cholera.
- Figure 17.1 shows a context-free plot of these locations, but Figure 17.2 – Snow's actual map – makes these locations understandable by putting them on a map of London streets.
- Snow's map (and the clustering of deaths around a water pump on Broad Street) helped scientists to understand that cholera spread through the drinking supply and was waterborne rather than airborne.
- One can see how this *spatial visualization* would be more immediately convincing to non-statistician policymakers than a statistical model might be.

Figure 17.2

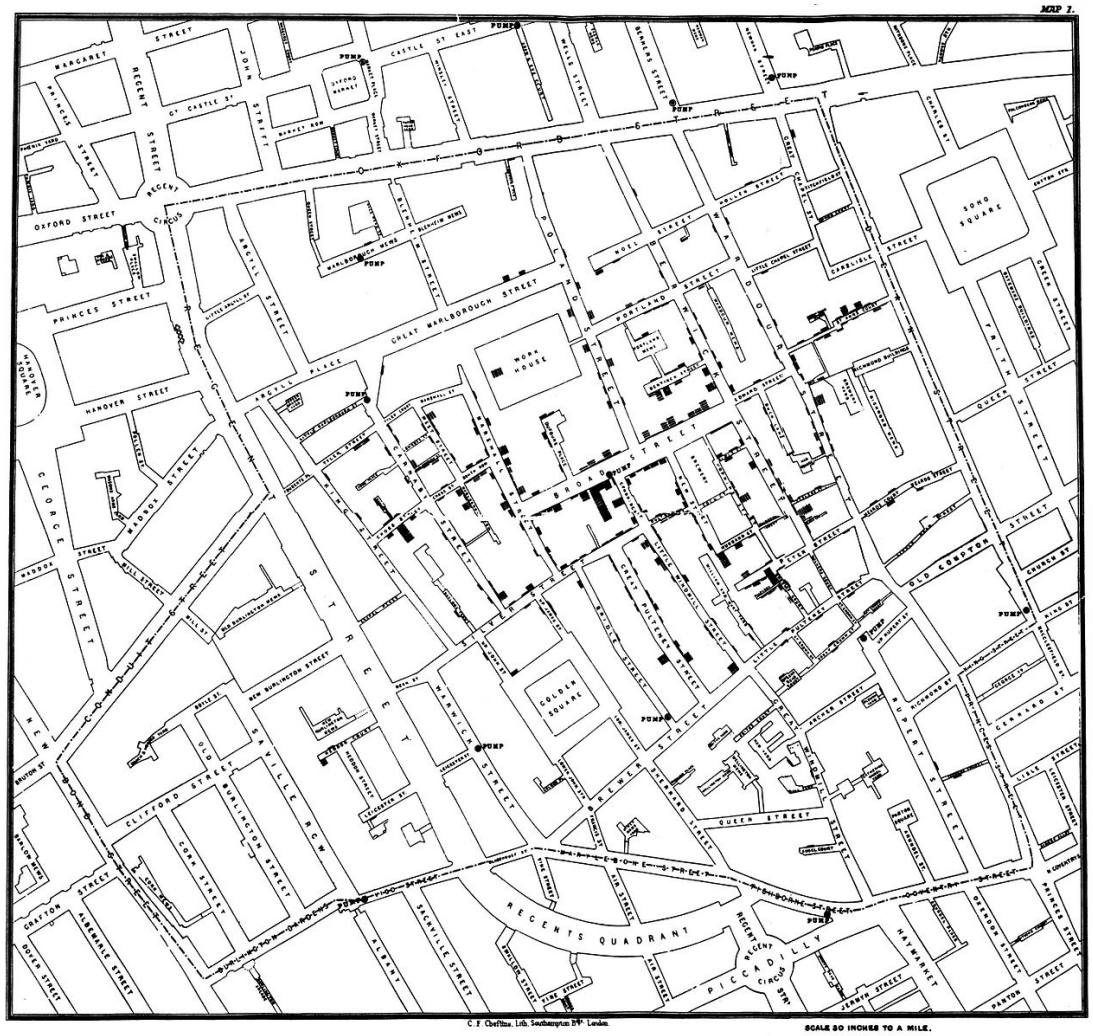

Figure: John Snow's Original Map of Cholera Deaths in London in 1854: Figure 17.2 from MDSR textbook

Structure of Shapefiles

- *Shapefiles* are one type of format for storing spatial data – another is the *KML* format, and there are other formats.
- The type of spatial data stored in shapefiles are not simply data frames with rows and columns.
- They have instructions for drawing maps with geographic boundaries.
- We will work with the `sf` package, a recently developed package for handling shapefiles that have the class `sf`.
- Let's examine some shapefiles related to Snow's cholera map.
- After downloading and unzipping a zipped folder, we see numerous files that correspond to two *layers* (sets of shapefiles), here called `Cholera_Deaths` and `Pumps`.
- The `Cholera_Deaths` layer is actually both a shapefile and a data frame in which one column is called `geometry`: This gives some coordinates for the spatial location in the row.

Making Static Maps

- The `geom_sf` function in `ggplot2` will plot geospatial data with some context (basically on top of a coordinate system such as latitude / longitude).
- Even better is the `annotation_map_tile` function in the `ggspatial` package which plots the geospatial data on top of a known map with boundaries, streets, and other features.
- See the Cholera death counts plotted on top of map tiles taken from Open Street Map (`type = "osm"`).
- The initial plotting attempt in this example is somewhat inaccurate because the coordinates of the geospatial data and those of the map tiles are in different units, and the translation produces a discrepancy (hundreds of meters off).

Projections

- When we make a map of a location on the Earth, we are projecting a 3-D spheroid onto a 2-D flat surface.
- Some information will be lost in the projection, but we can try to choose a projection system that will preserve the most important aspects of the map features (shape/angle? area?)
- The Mercator projection was popular centuries ago because it preserves angles, which made it useful for navigation at sea.
- But it distorts the areas of landmasses: Features near the poles appear much larger than they really are, while features near the equator appear smaller than they are.

Other Projections

- The Galls-Peters projection is a system that preserves area.
- The Lambert projection (which preserves angles) and the Albers projection (which minimizes gross distortions) are now commonly used.
- The implications of the choice of projection are more critical the larger the area that is on the map, such as in maps of the whole world or large countries.

Coordinate Reference Systems

- The *Coordinate Reference System* (CRS) is a way to track locations and their projection onto coordinate systems.
- There are three main formats of CRS: *EPSG* (i.e., European Petroleum Survey Group), an integer label; *PROJ.4*, a short text code; and *WKT* (i.e., Well-Known Text), a long and detailed description.
- The `st_crs` function will translate an EPSG code to the other formats, which are more understandable — especially the WKT format.
- The `st_transform` function will project geospatial data to a different CRS, which can be useful.
- See the example of projecting the `CholeraDeaths` data to the same CRS (`epsg:4326`) as the Open Street Maps tiles uses, which allows the deaths location markers to appear in the correct places on our map!

Dynamic Maps with `leaflet`

- The `leaflet` package, briefly described in Chapter 14, uses the `htmlwidgets` platform to produce interactive spatial maps.
- This allows the user to scroll, pan, and zoom on the maps as in the other interactive plots we've seen.
- We can add markers at special locations on the map.
- See the example of the interactive map of the area around the White House in Washington, DC.

Extended Example of Congressional Districts and Election Results

- Section 17.4 describes an extended example of plotting election results for congressional districts in North Carolina.
- It makes use of 2012 election data in the `fec12` package and 2012 congressional district shapefiles that are in a zipped folder that we can download and unzip.
- We can use color (varying shades of red and blue) to indicate where in North Carolina that Republicans and Democrats tend to get more votes.

Common Types of Maps

- Some common types of maps include:
  1. **Choropleth**: a map that colors or shades regions based on the value of a variable
  2. **Proportional symbol**: a map that associates a symbol with each location, but scales its size to reflect the value of a variable
  3. **Dot density**: a map that places dots for each data point, and displays their accumulation

Good Practices When Making Maps

- The scale of the proportional symbol should be based on its area, not its radius.
- When making maps with a color palette, categorical variables should be displayed using a qualitative palette, while numerical variables should be displayed using a sequential or diverging palette.
- The shade of a region on a choropleth should typically be based on a *rate* (like vote proportions) rather than a *raw count* (like vote totals).
- Again, the choice of projection is important, especially when plotting maps of large areas (see example plots of the maps of the whole United States).