Geospatial analysis in R: Part 1

Michael Otterstatter
BCCDC Biostats Session

Sept 6, 2019

Session Overview

- Background and concepts
  - geospatial analysis in public health
  - spatial data
  - coordinate reference systems and projection systems
  - Spatial data in R

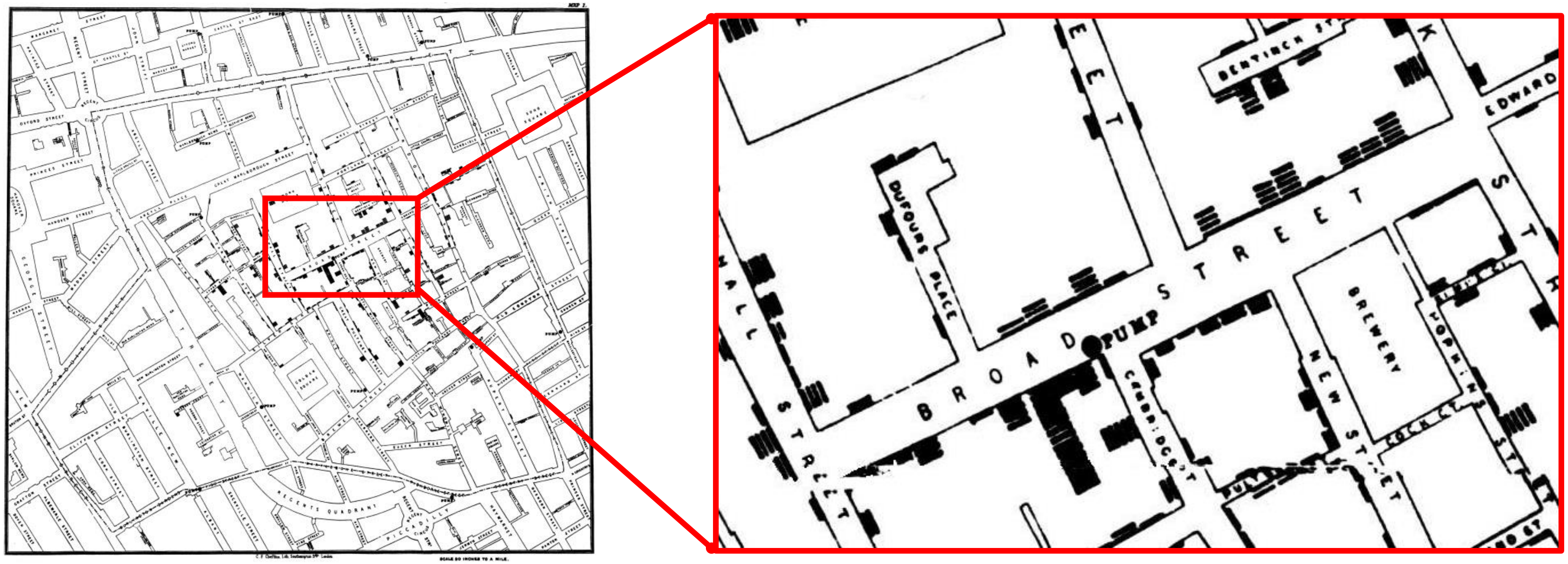
Background and concepts

Geospatial analysis in public health

- A foundation of epidemiology and public health is the identification and analysis of disease patterns in populations in terms of *person*, *place* and *time*, in order to:
  - track diseases
  - detect changes and outbreaks
  - evaluate interventions and services
  - prevent and control disease
- The key element of *place* identifies where a health event of interest occurred
  - Geospatial analysis is the set of methods and tools we use to describe and understand patterns in the location of health events

Geospatial analysis in public health

- The mapping of health events has a long history in epidemiology, going back at least as far as John Snow's mapping of cholera cases in London work during the 1850s



https://upload.wikimedia.org/wikipedia/commons/2/27/Snow-cholera-map-1.jpg

Geospatial analysis in public health

- Modern applications in public health include,
    - evaluate differences in health services and health-related rates/risk across different geographic areas
    - quantify uncertainty in spatial measures/maps
    - identify “clusters” of disease or other health events of interest
    - assess the significance of factors potentially related to disease occurrence (e.g., exposures, services, etc.)

Spatial data

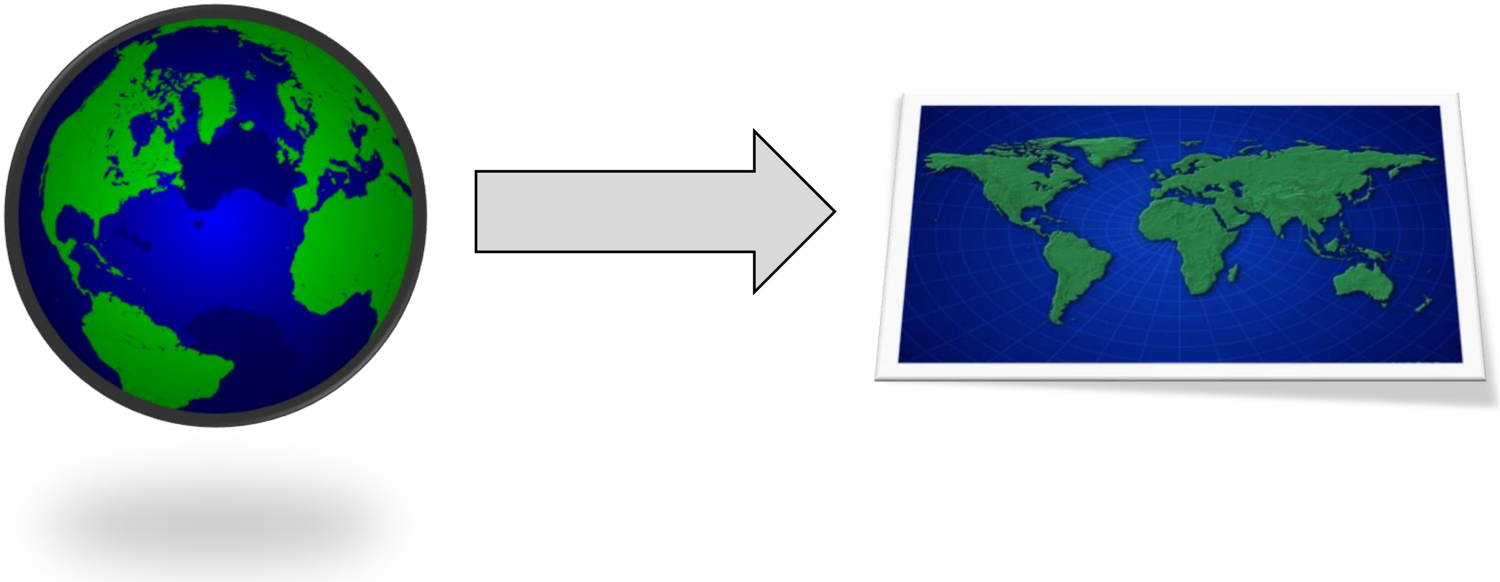
- Spatial data are those data located in two or more dimensions – these additional dimensions provide information to better analyze and understand your data
- The actual location of something is described using a *georeference*, typically a coordinate system

Coordinate reference systems (CRS)

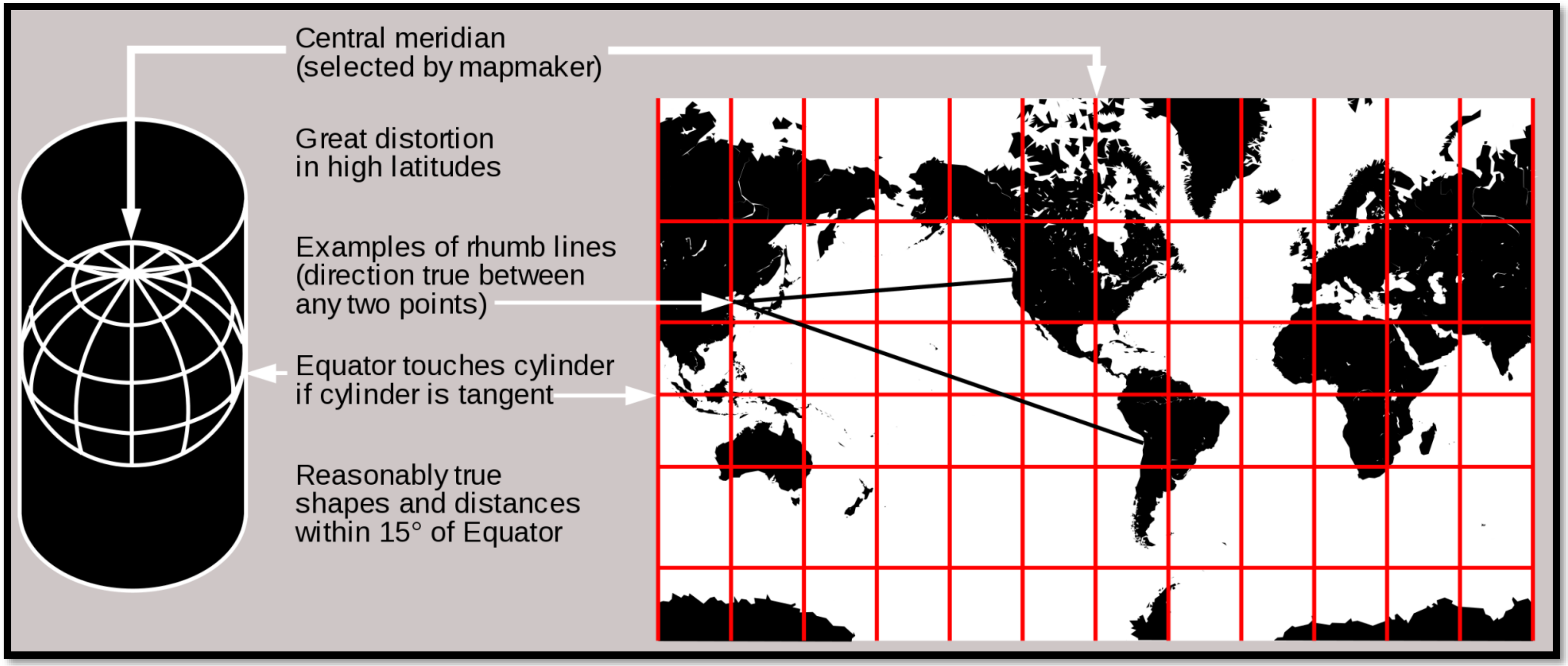
- How do we describe locations (the ‘coordinates’) of our data on the surface of the earth?
- a *coordinate reference system* (CRS) represents locations by coordinate pairs, indicating distance from an origin point
  - *geographic CRS*: based on latitude (degrees north/south of the equator) and longitude (degrees east/west of Prime Meridian) – every point on Earth is defined by a lat-lon coordinate pair, often written in degrees, minutes and seconds (DMS) or decimal degrees (DD)
  - *projected CRS:* based on units (typically meters) from an origin point (a ‘datum’)

Projection systems

- If we simply laid flat the Earth's curved surface, spatial locations would be warped
- Instead, to display spatial information defined by coordinates, we use a transformation (*projection*), converting location on the Earth's surface to a location on a flat two-dimensional map



Projection systems


Central meridian (selected by mapmaker)
Great distortion in high latitudes
Examples of rhumb lines (direction true between any two points)
Equator touches cylinder if cylinder is tangent
Reasonably true shapes and distances within 15° of Equator

- one consequence of projecting is that *scale* (map distance:actual distance) is not the same across a map's surface – reported scales are usually an average

https://en.wikipedia.org/wiki/Map_projection

Projection systems

- Many projection systems exist, with differing strengths/ weakness and degrees of warping
  - *conformal* - preserves spatial shape over small areas, but larger areas are distorted (e.g., ‘Mercator’)
  - *equal area* - preserves area so that relative sizes of regions are correct and comparable (e.g., ‘Albers’ equal-area’)
  - *equidistance* - preserves distance measures

Projection systems

- If the purpose is only a quick map, projection of spatial data is not strictly necessary – simple display of lat-lon is adequate
- However, if analyses involve measures/comparisons of distance, shape or area, a projected coordinate system must be used
- Choice of the projection system depends on what aspects (e.g., distance) must be most accurately preserved

Vectors and rasters

- *Vector data:* representation of geographic features as points, lines, and polygons using coordinate pairs and ordered lists of vertices; attributes may be associated with each vector feature
  - e.g., summarised population characteristics (e.g., health, socio-demography, etc.) by census regions or other distinct boundaries

Vectors and rasters

- *Raster data*: geographic features are represented and stored as uniformly sized pixels; attributes are typically associated with each of these grid cells
  - e.g., satellite imagery and other remote sensing information
- Both vector and raster data can be used together (or converted from one to the other), but vector data usually require much less computational memory and time

Points vs areal units

- Spatial information comes in a variety of forms, with space treated as
  - a continuous distribution, giving rise to *point data* with values that can occur at any location in space (e.g., continuous measures of temperature for all points within a region); or,
  - aggregate, or *areal,* units with summary values for each unit (e.g., average temperature in a region) which may be then assigned to a representative point such as the region’s centroid

Sources of spatial data in public health

- **health data** (e.g., outcome data from routine surveillance programs, administrative health databases, observational and experimental research studies), including vital statistics, reportable diseases, disease registries, national and provincial health surveys
- **census data** (e.g., socio-demographic information collected through formal government census), including population sizes by age, sex, region, counts of housing units, employment and income profiles, education, immigration, family sizes
- **environmental data** (e.g., ecological indicators and natural resources routinely monitored by government and other national/provincial organizations), including land use and related characteristics, habitats, agricultural operations and practices, bodies of water and water quality, air quality, climate
- **remote sensing data** (e.g., aerial and satellite imagining data) including, aerial photographs, satellite measures of radiation and sea-surface temperature

Considerations and cautions

- Interpretation of spatial data has certain challenges (modifiable areal unit problem, location uncertainty, ecologic fallacy, etc.) that must be considered carefully
- Different data sources may use different projection systems – these must be reconciled before data sources can be combined for analysis
- specific location information, e.g., street address or corresponding coordinates, can uniquely identify individuals -- care is needed when storing and analyzing such information and it may be necessary to mask locations to protect privacy

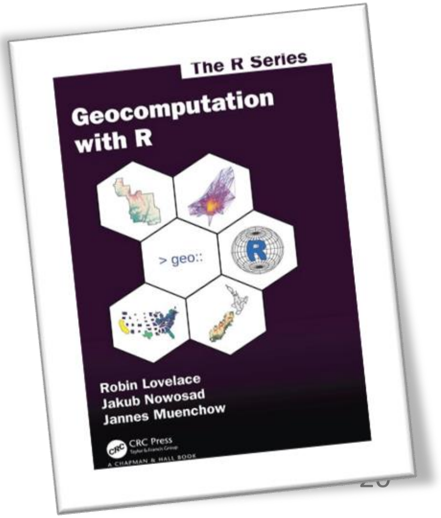
Spatial data in R

Background

- R is a free, open-source statistical software package that can be used for all aspects of data manipulation, analysis and visualisation (https://cran.r-project.org/)
  - R functions (commands) are organized into packages (also called libraries) that can be loaded as needed
  - Most R users interface with the R program via RStudio and use command-line code (scripts)
- Numerous R packages exist for working with spatial data (see: https://cran.r-project.org/web/views/Spatial.html)
  - Here, we focus on the new `sf` ('simple features') package (see: https://CRAN.R-project.org/package=sf)

The `sf` package

- The `sf` package is a collection of tools for working with spatial data in R, where ‘sf’ refers to *simple features*, which are a standard set of geometry types (e.g., point, line, polygon, etc.) supported by the package
- `sf` currently supports tools for vector data only (not raster data)
- use of the sf package is described extensively in the free online book *Geocomputation with R* (https://geocompr.robinlovelace.net/)



The `sf` package

- `sf` data objects appear in R as standard data frames, except with addition of a *geometry* column containing condensed list of all coordinate information for a particular spatial feature

- With a basis in standard data frames, powerful data manipulation tools of R, particularly those in the `tidyverse`, can be used with `sf`

The `sf` package

- For example, we can work with sf data objects in the same way as any (tidy) data object in R, including chaining commands together (piping) with `%>%`, using `filter` to include only certain rows, `select` to include only certain columns, and `mutate` to create new derived variables:

```
world_asia <- world_data %>%
 filter(continent == "Asia") %>%
 select(name, continent) %>%
 mutate(map_col = ifelse(name == “China”, “white”, “grey”))
```

The `sf` package

- `sf` contains numerous functions (commands) for importing, manipulating, transforming and exporting spatial data – all have intuitive names and begin with the prefix `st_`
    - `st_read` reads spatial data
    - `st_point` creates spatial points
    - `st_crs` retrieves the data's coordinate reference system
    - `st_distance` calculates Euclidean distance
    - etc.