Introduction to Applied Spatial Statistics

Spatial data are everywhere!

Practically every field of science produces spatial data

- Small scale: materials scientists study interactions between atoms
- Large scale: astrophysicists study spatial patterns of stars
- Ecology: plants and animals interact in space and time
- Health: “your zip code is more important than your genetic code”
- Economics: industry and policy change regionally
- Environmental science: pollution and weather are local events

Three types of spatial data

1. **Point-referenced data**: observations are made at point locations (e.g., lat/long)
   - Temperature measurements
   - Height of a tree
2. **Areal data**: observations are assigned to areas/regions
   - County-level cancer rates
   - State-level election results
3. **Point-pattern data**: the observations are the spatial location
   - Locations of hurricane landfalls
   - Locations of burglaries

Point-referenced data – Notation

- Let Y_i the response variable for observation $i \in \{1, ..., n\}$
  - Example: air pollution measurement
- The observation is made at spatial location $\mathbf{s}_i$
  - Example: latitude/longitude of the air pollution monitor
- Let X_i be a covariates associated with observation i
  - Spatial: elevation, distance to a highway
  - Non-spatial: time of day, type of measurement device

Point-referenced data – examples

- EPA air pollution data
- Satellite measurements of greenness
- Microbiome data

Point-referenced data – spatial correlation

- Analysis of point-referenced data is often called *geostatistics*
- Data are sampled at n locations, but theoretically they could be sampled at an uncountable number of locations
- Nearby sites are assumed to be correlated
- This is called *spatial correlation*
- Much of geostatistics focuses on estimating this correlation structure

Point-referenced data - objectives (tools)

- Estimate the range of spatial correlation (variogram, maximum likelihood analysis)
- Predict the response at an unmeasured site (Kriging)
- Estimate covariate effects while accounting for spatial correlation (maximum likelihood analysis)

Point-referenced data - advanced topics

- Analysis of non-Gaussian (binary, count) data
- Spatiotemporal methods: spatial data evolve over time
- Multivariate data: more than one type of response
- Design: what is the best set of locations to sample?

Areal data – Notation

- Let Y_i the response variable for observation $i \in \{1, ..., n\}$
  - Example: COVID-19 mortality rate in county i
- Adjacency: $A_{ij} = 1$ if regions i and j are adjacent and $A_{ij} = 0$ otherwise
  - Example: counties that share an edge are adjacent
- Let X_i be a covariates associated with observation i
  - Population density our county i

Areal data – examples

- Bed nets and malaria
- Air pollution and COVID-19
- 2016 Presidential election

Areal data - objectives (tools)

- Test for spatial dependence (Moran's I)
- Estimate the true value in each region (Bayesian methods)
- Estimate covariate effects while accounting for spatial dependence (Bayesian methods)

Point pattern data – Notation

- Let Y_i be the spatial location of observation $i \in \{1, ..., n\}$
  - Example: lat/long of the i^{th} earthquake in 2010
- Let $X(\mathbf{s})$ be a covariates associated with spatial location $\mathbf{s}$
  - Example: distance to a fault line

Point pattern data – examples

- Improvised explosive device explosions

- NBA shot charts

Point pattern data - objectives (tools)

- Test for clustering or repulsion of events (Ripley's K)
- Estimate the spatial intensity of events (kernel smoothing)
- Estimate covariate effects on the intensity (Poisson process methods)