Maps and Spatial Analysis in R

EPIC 2015

They don’t love you like I love you
R, ArcGIS, and Making Maps
R, ArcGIS, and Making Maps

Spatial Analysis for this map: Done in R

Spatial Analysis for this map: Done in R
Maps: why?

• You know how you sometimes read a paper and the figure makes it all make sense?

• Spatial analysis is often well suited to visual communication

• It’s important to remember that maps are just a kind of figure – a way of communicating some underlying relationships determined through spatial analysis
So really: why spatial analysis?

- Surveillance
  - Are the spatial patterns to disease incidence?

- Infectious disease etiology/vector identification
  - Can the spatial pattern tell us something about how the disease is being transmitted?

- Cluster investigations
- Neighborhood influences
- etc.
Spatial Epidemiology

• Geography is a discipline unto itself

• Spatial Epidemiology ought to be its own full semester course (or at least an EPIC course)

• So: this lecture is designed to give you a feel for what you might do rather than introduce everything
So: let’s play around a little

• Install the sp and maps packages

• Load the meuse dataset

  – library(sp)
  – data(meuse)

Meuse river

(The world’s oldest river?)
Getting started

• Use `str` to look at the meuse dataset.
• Where are we going to get spatial data out of this data frame?
  – x and y
Plotting spatial data

• Make a scatterplot of the points. Remember how?

\[ \text{plot(meuse}$x$, meuse$y) \]
Compare with Google Maps

• These don’t really look the same. Why not?

The Google Maps image took me, like, forever to find, because it turns out there’s more than one town named Stein in the Netherlands. Stein => Dutch for Springfield.
Coordinate systems

• All maps need to match date from a ball to a map drawn as plane.
• X and Y are defined within in a coordinate system
Coordinate systems (more)

• Remember this map?
• Its projection is inconsistent
  – The lines on the colored overlay don’t align with underlying features
So, back to the meuse dataset

- Look up documentation for meuse
- `?meuse`

```r
x
  a numeric vector; Easting (m) in Rijksdriehoek (RDH) (Netherlands topographical) map coordinates

y
  a numeric vector; Northing (m) in RDH coordinates
```
Spatially aware R objects

• meuse is a data frame

• Let’s give it coordinates, using a vector or a formula:
  – coordinates(meuse) <- ~x+y
  or
  – coordinates(meuse) <- c("x", "y")

• And a projection:
  – proj4string(meuse) <- CRS("+init=epsg:28992")
Plot with spatially aware object

• Plot as before:
  – plot(meuse)

The shape looks closer now...
Spatially aware R objects

• Okay, now let’s look at the meuse object again:
• \texttt{str(meuse)}

• WTF?!
Aside: why the sp package, which is generally awesome, makes me angry

• Caveat: the following is just my opinion

• Quietly changing the underlying type of an object is evil*

• For example, try this:
  – `data(meuse)`
  – `head(meuse)`
  – `coordinates(meuse) <- c("x","y")`
  – `head(meuse)`

* wherein evil is defined as unnecessarily making it possible to make mistakes without knowing it
SpatialPointsDataFrame

• S4 object
  – ‘slots’
  – head(meuse@data)

• Now try plotting now:
  – plot(meuse)

• spplot also available:
  – spplot(meuse)
  – spplot(meuse, “copper”)
  – Does this look more like what we might want? Maybe…
More about spatially aware objects

• Consider computing distance between two points or area of a triangle. If distances are long enough, the fact that they’re on a sphere might matter.

• Geosphere package is helpful for computing distances

Remember: you can get a PhD in geography. Though we who are getting PhDs in epidemiology probably won’t.
Things you can do with spatially aware objects

• Creating datasets
  – E.g. via spatial merge

• Spatial statistics
  – E.g. cases/zip code

• Spatial interpolation
  – E.g. kriging

• Etc.

• Again, Geography is its own discipline
Geocoding

• In a lot of datasets, you have an address and want to do something spatial with it
  – E.g. you want census data for residential neighborhood
  – Solution: get (lat, long), from the address then do spatial merge with Census data
  – Getting (lat, long) from an address is called geocoding
  – Any guesses as to what reverse geocoding is?
Geocoding in R

• Google ‘geocoding in R’
  – Pick the Stack Overflow hit from 2010
    • (http://stackoverflow.com/questions/3257441/geocoding-in-r-with-google-maps)
    • Confused?
  – Okay, go back and pick the allthingsr.blogspot.com hit
    • (http://allthingsr.blogspot.com/2012/01/geocode-your-data-using-r-json-and.html)
    • Do you want to take this block of code?
  – Okay, now go back and pick the ggmap documentation
    • (http://cran.r-project.org/web/packages/ggmap/ggmap.pdf)
    • This looks like something we’re used to
Geocoding with R using ggmap

- Challenge #1: use ggmap to get the (lat, long) of 722 W 168th St.

  - library(ggmap)
  - geocode('722 W 168th St, New York, NY')
Challenge #2: compute distance

• Compute the distance from the Mailman School of Public Health to the John Snow Pub in London
• Hint: use geosphere library

  library(geosphere)
  msph <- geocode('722 W 168th St, New York, NY')
  ser <- geocode('John Snow Pub, London, UK')
  distm(msph, ser)
Geocoding: a caveat

• 2 Quick challenges

  – Use qmap to get a map of Stein, NL at zoom = 12

  – Now use qmap to get a map of Geleen, NL at zoom = 12.
    • (Does the curve of the river on the left look familiar?)
Spatial Autocorrelation

• Co-variation of properties within geographic space
• Note that perfect mixing is not expected if each cell is independent

![Positive spatial autocorrelation](image1) ![No spatial autocorrelation](image2) ![Negative spatial autocorrelation](image3)
Modifiable Areal Unit Problem

- Different spatial unit definitions can result in different findings

- Usual recommendation: start from theory
Cluster Detection

• Given point data for some phenomenon, are points closer than would be expected by chance?

• Note that we are interested in clustering in the population, not in the sample
  – many sampling strategies (esp. convenience samples) artificially induce clustering.
Nearest Neighbor Index

• Defined as the ratio of the mean distance to the nearest neighbor for all observation to the mean distance that would be expected due to chance alone
  – Values significantly less than 1 indicate clustering
• But: 50 randomly distributed pairs of points would have a very low index, but no real clusters as we'd traditionally think of them
Other cluster detection/analysis methods

- Ripley's K function
- Kulldorff's scan statistic
- Bayesian hierarchical modeling
Two other important spatial techniques

- Land use regression
- Kriging
Land use regression

• Basically a 4-step process
  1. Pick sample locations (usually a grid)
  2. Measure outcome and land uses (e.g. population density, proximity to highways) at each sample point
  3. Regress outcome on land uses, use model to predict in unmeasured locations
  4. Check residuals for spatial autocorrelation
Land Use Regression in R

- Basic LUR ignores spatial covariance, so it does not require custom regression model fitting.
Kriging

• 4-step process
  1. Pick sample points
  2. Sample outcome
  3. Use spatial covariance in outcome to predict at unsampled points
  4. Draw multiple times from posterior distribution of spatial model to get 'conditional realizations' – multiple imputation-like way of including uncertainty in estimates

• Universal Kriging essentially integrates LUR and Kriging
Conditional Realizations

Interpolated Values

Standard Errors

Conditional Realization #1

Conditional Realization #2

Conditional Realization #3

Conditional Realization #4

Conditional Realization #5

Conditional Realization #6

Conditional Realization #7
Kriging in R

- Several packages support kriging
  - GeoR
  - kriging
  - gstat

Kriged using GeoR