Combining Spatial Data

- Have two (or more) spatial data sets, want to combine them into one
- Sometimes easy
- Areal data:
  - population by county in one data set
  - sales tax revenue by county ($\propto$ sales) in a second
  - health information (e.g., % BMI $>$ 30) by county in a third
- Same regions in all three, link by county id
- Implementation:
- made easier because areal data is just a data frame
- or the @data slot in a SpatialPolygonsDataFrame object
  - **IF** all three datasets in same order
    - So row 5 in all three is the same county
    - Just cbind() the columns
    - No check to avoid accidentally merging Story county in one data set with Sac county in another
- Geostatistical data: same idea with data in a SpatialPointsDataFrame

Combining Spatial Data

- More robust implementation: use dplyr functions
- dplyr library has lots of tools for data manipulation
  - the `*_join()` functions merge data sets by matching rows by a specified variable
  - 4 different versions:
    `left_join()`, `right_join()`, `inner_join()` and `full_join()`
  - All join two data sets, so with three would use twice:
    A+B to AB, then AB + C to ABC
  - Same result when every row in one data set matches a row in the other
  - Differ in how rows in one data set but not the other are treated

Combining Spatial Data

- Part of the "tidyverse": collection of libraries for manipulating "tidy" data sets
  - Row = observation, column = variables
- One issue: these functions create tibbles
  - extension of data frames, nice print function
  - But, spatial packages are expecting data frames, not tibbles
  - don't know how well (or even if) they deal with tibbles
  - Some may, others may not
  - My solution: flip the tibble back to a data frame with as.data.frame()
  - Save that in the @data slot of a SpatialPointsDataFrame or SpatialPolygonsDataFrame

Combining Spatial Data

- sp and spatstat provide lots of data manipulation functions
- e.g., can:
  - count # points within a polygon
  - calculate averages within a polygon
  - create a spatstat image from values in polygons (a little tedious)
  - lapply() or sapply() to apply a function to each component of a list
    - lapply() returns another list; sapply() returns a vector or matrix
- maptools library:
  - many additional data conversion functions
  - especially between sp and spatstat
- sf library: spatial features
  - the new version of sp
  - simplifies sp data structures
  - and adds more GIS-like operations
  - Not yet supported by all spatial analysis libraries

Combining spatial data

- Sometimes not so easy, some examples
- Voting rate (% eligible voters) by congressional district over last 70 years
  - District boundaries have changed (many times)
- Merging data at different spatial resolutions
  - e.g., two types of satellite imagery
    One at 20m resolution, other at 1km resolution
  - Or combining state-level and county-level areal data
- Or combining different sorts of data
  - Remote sensed data and point data
  - Remote sensed: spatially extensive coverage, grid cells
  - Point data: at specific locations
- My notes heavily based on Gotway and Young, 2002, J. Am. Stat. Assoc. 97:632-648

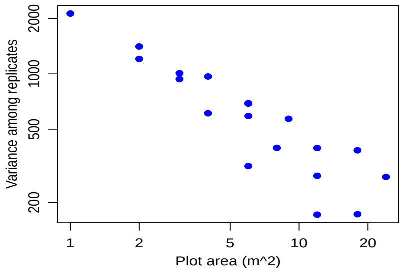
Combining Spatial Data

- These notes:
  - concepts and issues
  - outlines of some solutions
- The most important concept:
  - The "support" of spatial data matters
  - Are the data for point locations, for small area, for larger areas?
  - For many problems, choice of support is arbitrary

"geographical areas chosen for the calculation of crop yields are modifiable units and necessarily so. Since it is impossible (or at any rate agriculturally impractical) to grow wheat and potatoes on the same piece of ground simultaneously we must, to give our investigation any meaning, consider an area containing both wheat and potatoes, and this area is modifiable at choice" (Yule and Kendall, 1950, An Introduction to the Theory of Statistics).

Modifiable Areal Unit Problem

- Different choices of support have different statistical properties.
- Long history, by various names.
- "Fairfield-Smith variance law" (1938):
  - arbitrary choice of plot size changes variance between replicates



Fairfield-Smith

- Proposed model for variance, V_{area}, as function of plot size, A

$$V_{area} = \frac{V_1}{A^b}$$

- V_1 is variance for a unit size plot,
- b is an empirical coefficient, usually less than 1, $\hat{b} = 0.741$
- Related to spatial correlation between nearby small plots
- No single semivariogram model "explains" the law.
- A Matern model is closest
- Plot shape (e.g., square, rectangle) also matters, sometimes

Fairfield-Smith: practical use

- Consider two study designs
  - 1) Treatment applied to plots of size 1, 25 replicates per trt
  - 2) Treatment applied to plots of size 5, 5 replicates per trt
- Same total area for both designs
- Which is better (statistical considerations only)?
  - 1) Estimated $V_1 = 2100$. se mean $= \sqrt{2100/25} = 9.16$
  - 2) Estimated $V_5 = 637$. se mean $= \sqrt{637/5} = 11.28$
- Why the difference?
  - Consider the size 5 plot as 5 adjacent size 1 plots
  - adjacent plots are spatially correlated
  - 5 independent plots less variable (Var = 420) than a single plot of size 5 (Var = 637)

Modifiable Areal Unit Problem

- Two related issues
  - Scale effect = aggregation effect
    contiguous units grouped into larger areal units (data above)
  - Grouping effect = zoning effect
    differences in unit shape at the same or similar scales
- One or both present in a single data set / analysis
- Affects more than just variance
- what is the correlation between % Republican voters and % elderly
  - Data: 99 counties in IA, aggregate into 12 “districts”
  - calculate correlation from those 12 observations
  - can find an aggregation with a correlation of -0.97
  - and another with a correlation of +0.99
  - (Openshaw and Taylor 1979, who coined the name MAUP)

Ecological fallacy

- Mechanism works at the level of individuals
  - Is an older person more likely to vote Republican?
- “ecological” correlation is between groups (e.g. counties)
- ecological fallacy is using correlations among groups to infer associations for individuals
  - using correlations between % elderly in a county and % Republican vote
- conclusions at the two levels are often different
- two possible mechanisms for differences:
  - aggregation bias (grouping of individuals)
  - specification bias (confounding variables have different distributions in groups than in individuals)
- Note similarity to MAUP issues
  - Ecological fallacy now viewed as special case of the MAUP

Combining data: issues and methods

- So how do you appropriately combine incompatible spatial data?
- Different sizes/shapes of units, possible changing boundaries
- e.g.: 2005 data at points and in 3 regions, want to compare to 2010 data in 2 regions
  - need “2005” values for 2010 regions
- Many solutions, just outline them here
- “Non-statistical” solutions, basically weighted averages
  - areal data: weight by proportion of 2005 area in each 2010 region
  - point data: construct Thiessen polygons for each location, weight by area
- “Statistical” solutions, basically predicting on a fine grid using a model, then summing to get estimates for new regions
  - point data: block kriging
  - areal data: model observed areal values as sum of unobserved latent variables
  - point process data: model intensity surface, integrate over 2010 region
- Often using hierarchical models, often with a Bayesian flavor