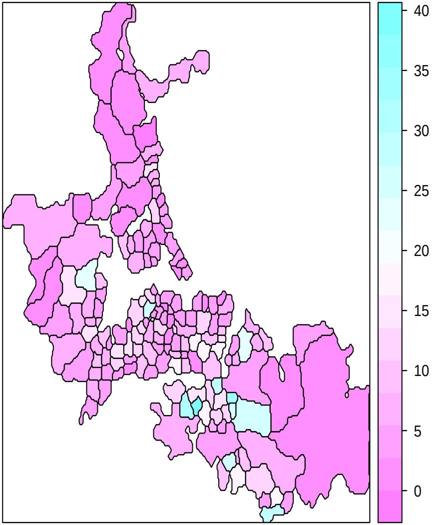
Areal data

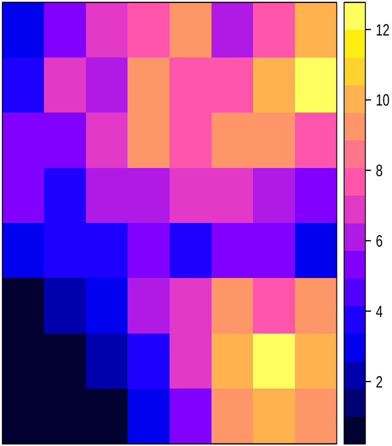
Reminder about types of data

- Geostatistical data:
  - $\boldsymbol{Z(s)}$ exists everyhere, varies continuously
  - Can accommodate sudden changes by using a model for the mean
    - E.g., soil pH, two soil types with different pH
    - model includes mean that depends on soil type
    - error = small scale variation, assumed continuous
    - use Universal kriging to predict/map
- Areal data
  - Spatial data measured and reported by regions
  - Only one value for each region
    - May vary continuously within region
    - But data only available for a region
  - abrupt change at region boundaries are likely
  - Unlike geostat data, "location" is arbitrary within region
- Some examples of areal data in pictures

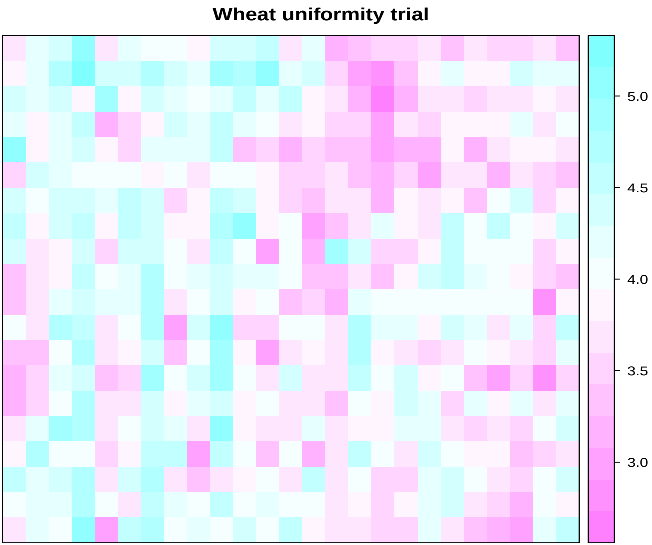
Infant mortality, Auckland NZ districts



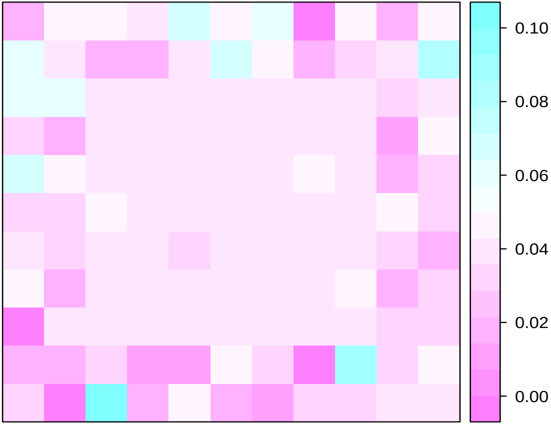
Number of plant species in 20cm x 20 cm patches of alpine tundra



Wheat yield



Incidence of flu in Iowa: made up data



Areal data

- Areal data often arises by aggregation
  - number of disease cases by county
- but doesn't have to: US states, coded by party of state Governor
- Distinction between Geostat and Areal can be blurred
  - SST data: starts as geostatistical data
  - displayed as single value per spatial grid cell (e.g., 1° x 1° area)
  - To me:
    - relative scale, size of measurement to span of study area
    - are measurements available for all units?

sea surface tem

- Distinction between Areal and Point pattern can be blurred
  - Disease cases: will treat as areal data
  - but point pattern if have individual locations (household address)
  - and OK to assume household address is the "location" (work? shopping?)
- Again, relative scale matters. What is the location of an individual?

Areal Data

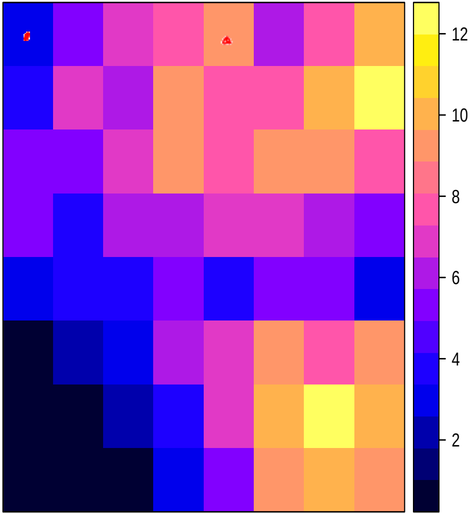
Goals:

- Visualize data
- Describe spatial dependence
- Predict values
  - Less common to predict at new locations
    - because likely to have data on all locations (Auckland districts/quadrats/plots)
  - But predictions at measured locations will smooth the data
    - Assume observations are "noisy", i.e., measurement error at each location,
    - Want to smooth (reduce amount of noise)
    - i.e., predict "true" values at observed locations
    - Like measurement error kriging
- Fit regression models while accounting for spatial dependence

Describing association

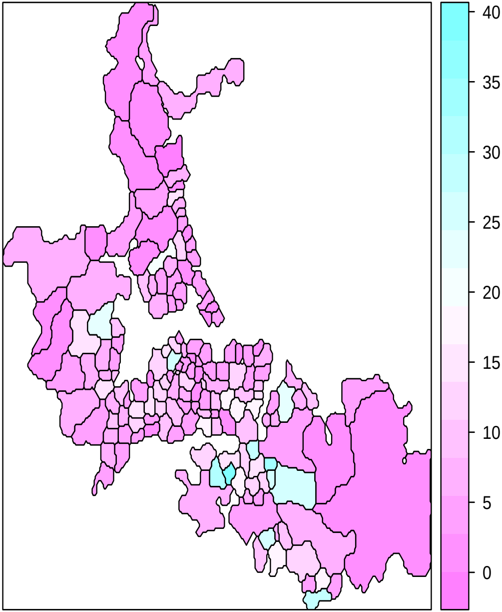
- Spatial association as a function of *what*?
- Consider data on a grid, e.g. species diversity in alpine tundra

Describing association: Spp diversity in alpine tundra



Describing association

- Could describe in terms of distance
  - $1 =$ up/down or across
  - $\sqrt{2} =$ diagonal, and so on
  - $2 =$ up/down or across 2
- But approximate, since distances between areas, not points
- What about irregular regions (e.g. Auckland, on next slide)?
- Use distance between centroids of regions - perhaps
  - but depends on size of region
- Usual solution: pairwise "connectivity":
  - how well connected are two regions?


40
35
30
25
20
15
10
5
0

Connectivity

- On a grid: (spp diversity picture on next slide)

- two common definitions of connectivity
  - rook's move: a square in the middle connects to the two on either side and the two above/below
  - queen's move: a square in the middle connects to its 8 neighbors (rooks + 4 diagonals)
- squares on edge have 3/5 neighbors; those in the corners have only 2/3

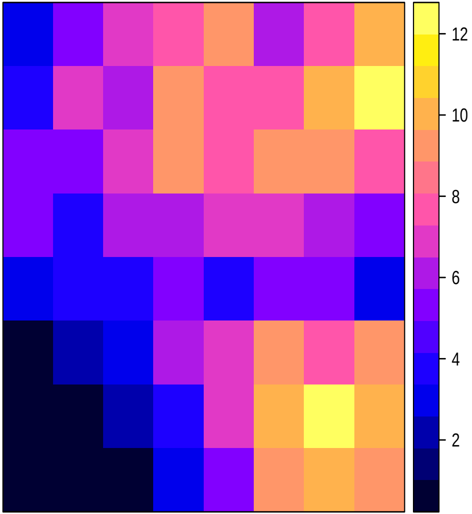
- Connections define the "spatial proximity" matrix, also called "spatial connectivity" matrix
  - # rows = # columns = # regions
  - 1 if a pair of regions connect
  - 0 otherwise
  - always 0 down the diagonals

regions

region

0 1 0
0

Connectivity



Connectivity - options

- Irregular regions: two approaches
  - 1 if two regions share a boundary
    very irregular neighbor count.
    Auckland: from 1 - 10 neighbors, median 4
  - Could also use % boundary shared
    very useful for modeling transport among regions
    economic flows, disease propagation, invasive spp

- Could also use distance. Options include:
  - all regions with centroids within d of target (0/1)
  - find the k nearest neighbors (k smallest distance between centroids)
  - make weight a function of distance, $d^{-\alpha}$
- Can use values as they are
- Or row-standardize
  - so row sum = 1
  - if have 2 N's, each has connectivity 1/2
  - if have 4 N's, each has connectivity 1/4
- All are different views of how region A might influence B

Connectivity

- No standard way.
- If there is something that makes sense for the problem: use it!
- best if connectedness measure informed by subject-matter knowledge
- When no such insight, most common is:
  - shared boundary $= 0/1$, perhaps with row standardization
- choice does have statistical consequences, especially when predicting/smoothing
- Some weight matrices are symmetrical
  - $0/1$ shared boundary
  - $d^{-\alpha}$
- others are not
  - % shared boundary
  - row-standardized matrices

Spatial dependence

- One very common measure, one less common
- Moran's I, dates to 1950

pairs

$$I = \frac{1}{s^2} \frac{\Sigma_{i,j} w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\Sigma_{i,j} w_{ij}}$$

$$s^2 = \frac{1}{N} \Sigma_i (Y_i - \bar{Y})^2, \text{ i.e., mle, not usual unbiased est.}$$

  - w_{ij} is the ij'th element of the spatial weight matrix
- looks like a correlation coefficient between two variables, Y, Z:

$$r = \frac{\Sigma(Y_i - \bar{Y})(Z_i - \bar{Z})}{n\sqrt{\text{Var } Y \text{ Var } Z}}$$

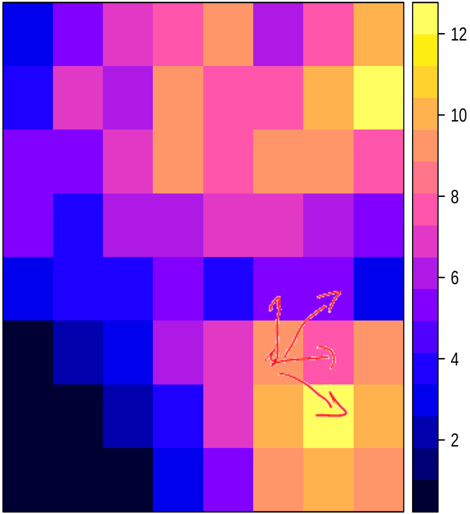
- I ranges from +1 to -1
  - 0: no spatial correlation
  - 1: perfect positive correlation

Moran's I

- Tests of correlation = 0:
- 1) If sufficient # regions, $I \sim N$ with mean and variance that can be computed
  - E $\hat{I} = \frac{-1}{N-1}$, variance formula not insightful
  - not clear what is "sufficient", depends on W, but would like 20 or more regions
- 2) permutation: randomly shuffle observed values over the regions, compute I each time

  9999
  - enumerate all permutations: $p = \frac{\#\text{more extreme}}{\#\text{permutations}}$
  - sample (randomization test): $p = \frac{\#\text{more extreme}+1}{\#\text{permutations}+1}$
  - +1 accounts for the observed data (~~already included~~ in all permutations)
- Usually one-sided test, only interested in positive spatial dependence

Moran's I example: spp div data, Queen's move neighbors



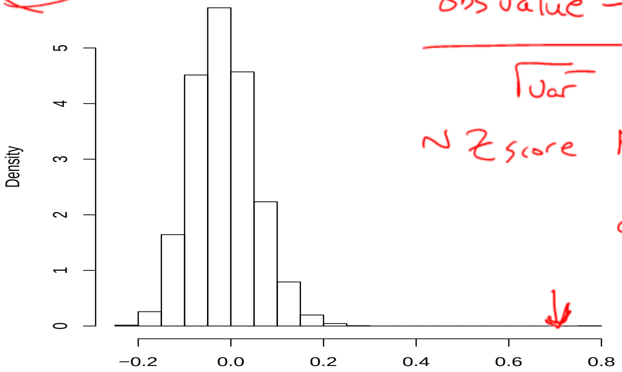
Moran's I example

- $\hat{I} = 0.751$
- randomization test: 0.751 exceeds all 9,999 randomizations, p = 1/10,000 = 0.0001
- normal approximation: E $\hat{I}$ = -0.0157, Var $\hat{I}$ = 0.00451
  $Z = \frac{I - E\ I}{\sqrt{Var\ I}} = 11.4$, p < 0.0001

obs value − Expected / √Var

∼ Z score N(0, 1)

99 p = 0.01



Geary's c

all pairs

- Based on squared differences, not covariance

$$c = \frac{N-1}{2\Sigma_{i,j} w_{ij}} \frac{\Sigma_{i,j} w_{ij}(Y_i - Y_j)^2}{\Sigma_i (Y_i - \bar{Y})^2}$$

- similar in spirit to semivariance in geostats
- denominator scales to ± 1
- usually similar to but not same as Moran's I
- when there is a difference
  - I is a more global indicator, because uses $\bar{Y}$
  - c is more sensitive to differences in local neighborhoods
- Test Ho: no spatial dependence using permutations or normal approximation
- Notice that both I and c sum over all pairs of points.
  - One number for entire region

Local indicators of spatial association

$\sum_{ij} w_{ij} (Y_i - \bar{Y}) (Y_j - \bar{Y})$ areals. r

- rewrite I as:

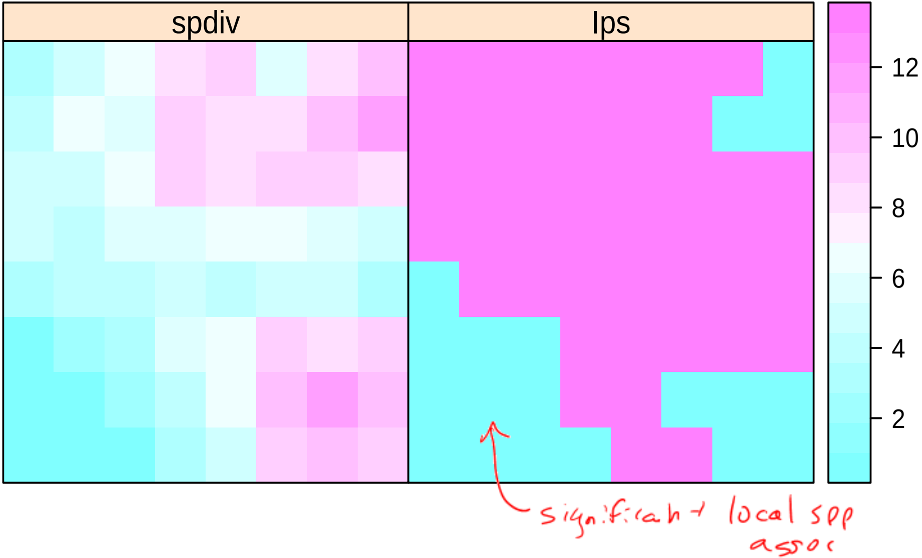
$$I = \frac{1}{s^2 \, \Sigma_{i,j} w_{ij}} \Sigma_i (Y_i - \bar{Y}) \, \Sigma_j w_{ij} (Y_j - \bar{Y})$$

- Calculate second sum separately for each region

$$I_i = \frac{1}{s^2} (Y_i - \bar{Y}) \, \Sigma_j w_{ij} (Y_j - \bar{Y})$$ – local Moran's I

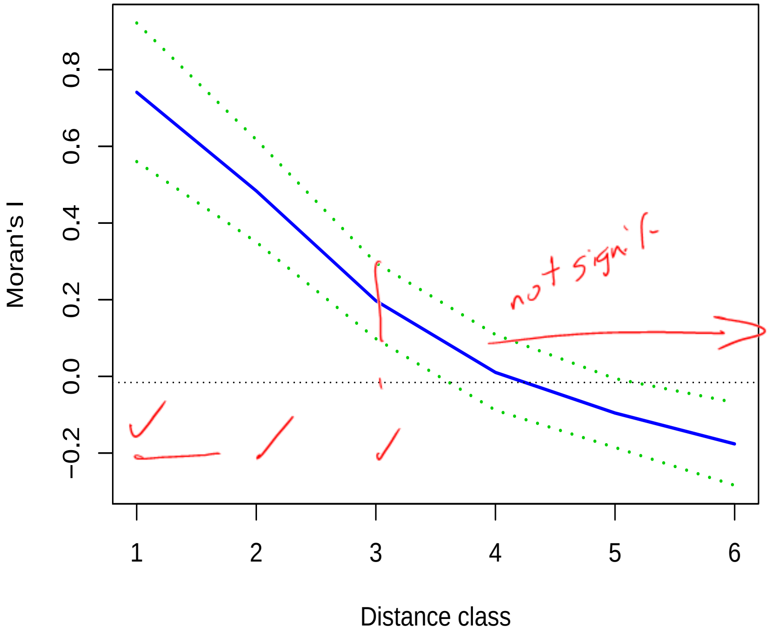
- global statistic is then $I = \Sigma_i I_i / \Sigma_{i,j} w_{ij}$
- illustrate using species diversity data
    - Data and where Moran's I_i marking $Z_i > 2$
    - Blue areas for Ips are marking where that region is significantly similar to neighboring regions

LISA for spp diversity data



Moran correlogram

- Define different weight matrices using different distance classes
  - (0, 1.5) using queen's move neighbors
  - (2, 3) → regions slightly further apart
  - (3, 5), and so on
- Or, use 1st nearest neighbor, 2nd NN, 3rd NN, . . .
- Calculate Moran's I for each weight matrix
- Plot I vs distance


Moran's I
0.8
0.6
0.4
0.2
0.0
−0.2
1
2
3
4
5
6
Distance class
not signif

Join count statistics

- Moran's I and Geary's c are for continuous observations
  - c: "similar" because $(Y_i - Y_j)^2$ is small
- What about categorical data
  - E.g., US states, record whether governor is Republican or Democrat
- Is there spatial correlation?
  - i.e., If your state is Republican, are neighboring states more likely to be Republican?
- Usual approach is the Black-Black (BB) join count statistic

$$BB = \frac{1}{2}\Sigma_{i,j} w_{ij} I_i I_j$$ – Indicator variables

- I_i is 1 if the "event" happens in region i / 0 if doesn't
  If w_{ij} is 1 if neighbors, 0 otherwise, BB is the number of pairs where both region and neighbor are events

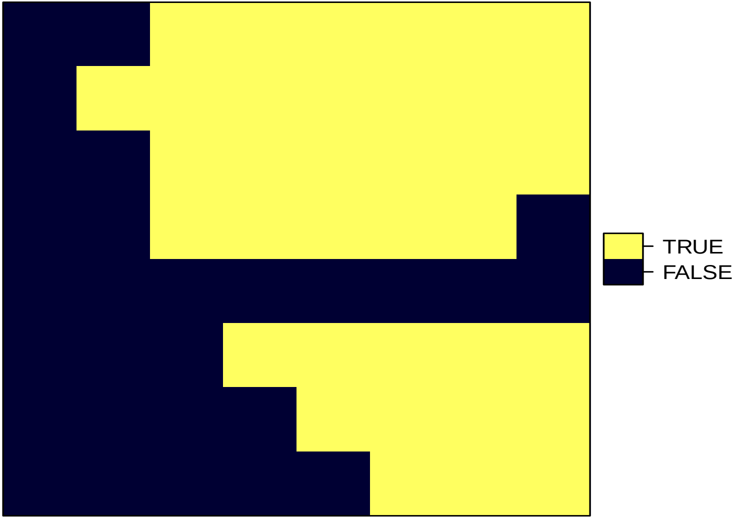
Join count statistics

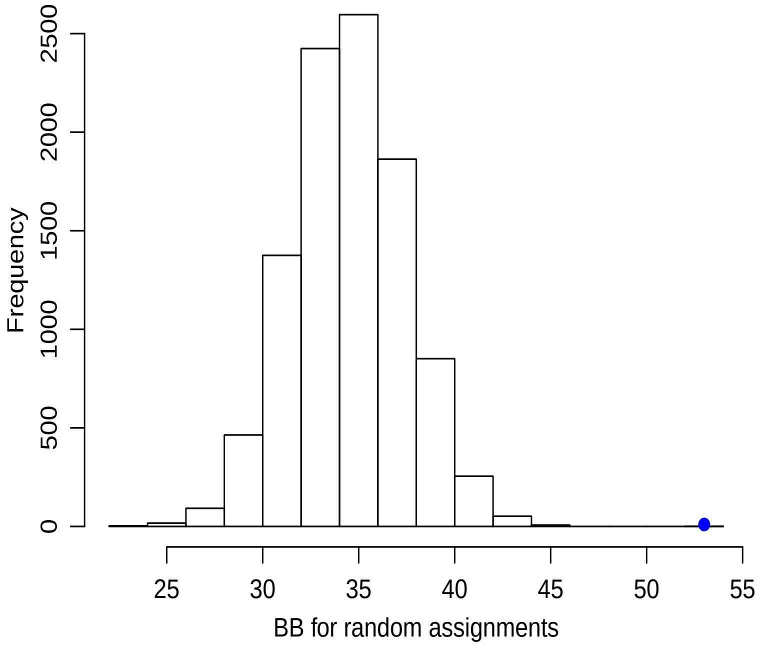
- Compare BB to expected value if no spatial dependence
- Either a Z-test (assume normal-distribution)

$$Z = \frac{BB - \mathrm{E}\ BB}{\sqrt{\mathrm{Var}\ BB}}$$

– when no dependence

- or permutation:
  - keep same number of "event" and non-"event" regions,
  - keep neighbor structure,
  - assign "event" / non-"event" randomly to regions.
  - compute BB for each randomization
- Example: Species diversity plot, define "rich" as 6 or more species
- Rich-rich, rooks neighbors: $BB = 53$, E $BB = 35$, Var $BB = 8.495$
- Normal approximation: $Z = (53 - 35)/\sqrt{8.495} = 6.176, p < 0.0001$
- Permutation: 53 is larger than any of 9999 randomizations, p = 0.0001


TRUE
FALSE



Frequency
0
500
1000
1500
2000
2500
25
30
35
40
45
50
55
BB for random assignments

Combining multiple sites

- Imagine a study of spatial dependence in Iowa that is repeated in Indiana.
- Methods above will describe that spatial dependence in each state
- What if you believed the spatial dependence was similar in the two, and wanted one result
- How can you combine information from both states?
- No problem, but should think about a few issues.

Moran are global statistics
use info entire region.

Combining multiple sites

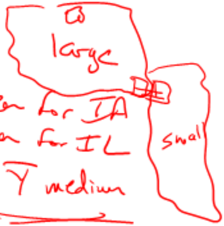
$(Y_i - \bar{Y})(Y_j - \bar{Y})$

$\bar{Y}$ same

IA mean for IA

IL mean for IL

$\bar{Y}$ medium



- 1) One mean for both states, or separate means?
  - the issue is the scale of spatial dependence (within state only or including between state)
  - not an issue if the means are similar, but they are often not
  - Separate means → evaluates pooled within state spatial dependence
  - Separate means is most common
- 2) Include only pairs of regions within states, or pairs crossing states?
  - Same spatial scale issues.
  - Only pairs within states is most common.

Combining multiple sites

IA more counties ⇒ more pairs
IL fewer

- 3) Do the two states contribute equal amounts of information?
  - Here, same within state sampling, same # regions, same weighting scheme (In my story)
  - So, equal amounts of information — same sample sizes.
  - use a simple average of both states (A and B)
  - $\hat{I}$ for both states: $\hat{I} = (\hat{I}_A + \hat{I}_B)/2$
  - E $I = (\text{E } I_A + \text{E } I_B)/2$
  - Var $\hat{I} = (\text{Var } \hat{I}_A + \text{Var } \hat{I}_B)/4$

  - If unequal amounts of information, use a weighted average
  - Many possible ways to weight, depends on study specifics

Combining multiple sites

- Hard part is getting estimate and its variance
- Test Ho: no within state spatial dependence
  - Z score for overall study, or
  - permuting observations within state.
- BTW, same ideas can be used for semivariograms for multiple sites
- Computing trick:
  - artificially separate sites,
  - make sure min distance between IA and IN larger than max within state distance
  - specify max semivariogram distance so that all pairs are within a site.
  - weights each region by number of pairs

Multiple sites: example

- Iowa: $\hat{I} = 0.35$, E $\hat{I} = -0.0159$, Var $\hat{I} = 0.085$
- Indiana: $\hat{I} = 0.42$, E $\hat{I} = -0.0159$, Var $\hat{I} = 0.085$
- Individually:
  - Iowa: $Z = \frac{0.35-(-0.0159)}{\sqrt{0.085}} = 1.25$, p $= 0.10$
  - Indiana: $Z = \frac{0.42-(-0.0159)}{\sqrt{0.085}} = 1.49$, p $= 0.067$
- Together: $\hat{I} = (0.35 + 0.42)/2 = 0.385$, E $\hat{I} = -0.0159$, Var $\hat{I} = (0.085 + 0.085)/4 = 0.0425$
- $Z = \frac{0.385-(-0.0159)}{\sqrt{0.0425}} = 1.94$, p $= 0.026$
- Similar patterns in both areas, aggregate the two $\Rightarrow$ stronger evidence