Spatial Statistics

PM569 Lecture 7: Areal Data 1

Meredith Franklin

Division of Biostatistics, University of Southern California

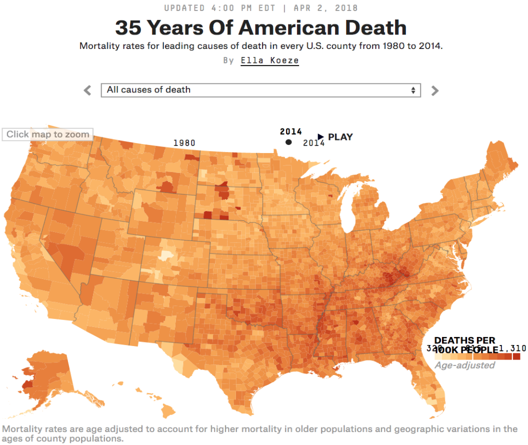
October 11, 2019

Areal Data

- Data referenced at an aggregate level.
- Areal "units" are generally irregular geographic areas, and in spatial analysis we have a collection of areal units. The value we are interested in within one areal unit is presumed to be constant.
- Common areal data are census data (blocks, tracts, counties, states), zip codes.
- We often have shapefiles to deal with, as they are used for collections of polygons.

Areal Data: Example

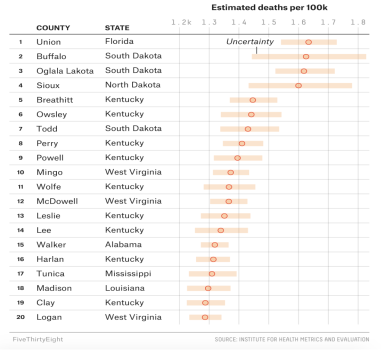
Cause-specific Mortality Rates by County



https://projects.fivethirtyeight.com/mortality-rates-united-states/

Areal Data: Example

- The National Institutes of Health Metrics and Evaluation conducted a spatial analysis of these data (on a national level) `https://jamanetwork.com/journals/jama/fullarticle/2592499`.
- They used conditional autoregressive models (Bayesian).
- Revealed regional and local variations in causes of death.



Counties with highest estimated mortality rates, 2014

Link between geostatistical/point referenced and areal data

- For geostatistical/point referenced data, we use functions of distance to estimate the variogram/covariance that defines spatial relationships.
- Geostatistical prediction involves using fitted covariance functions (kriging), spatial interpolation, or basis spline smoothing.
- For areal data, we use neighbour information to define spatial relationships.

Link between geostatistical/point referenced and areal data

- In general, areal units are irregular (e.g. zip code, county) but methods may also apply to regular grids.
- We care about how areal units connect to each other.
- We will see some analogies between geostatistical data and areal data. Sometimes geostatistical methods are used for areal data prediction, but autoregressive models employing neighbourhood information are more commonly used.
- We will use the R package `spdep()` for areal data analysis.

Areal Data: Inferential issues with areal data

Is there a spatial pattern?

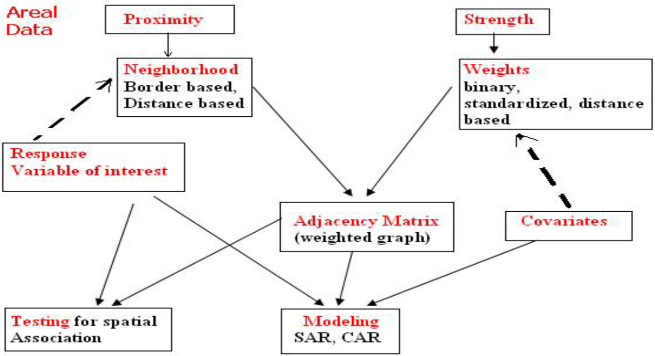
- Spatial pattern suggest that areal observations close to each other have more similar values than those far from each other.
- You might think that there is a pattern through visualization, but this is often subjective.
- Independent measurements will have no pattern, and would look completely random, but there may actually be an underlying pattern.
- If there is a spatial pattern, *how strong is it?*

Areal Data: Analyses

- Response of interest Y_i measured in block or areal unit B_i
- The B_i are supplemented with neighbourhood information (distance between B_i and B_j, area of B_i, boundary/edge connections)
- Areal data analysis involves:
  - Representation of spatial proximity in areal data using weighted graphs
  - Testing for spatial pattern: Global testing using Morans I or Gearys C statistic
  - Testing for spatial pattern: Local testing using local Moran's I or Getis-Ord G^* statistic
  - Modeling spatial pattern for prediction and inference: autoregressive models including Simultaneous Autoregressive (SAR) models and Conditional Autoregressive (CAR) models

Areal Data: Flowchart

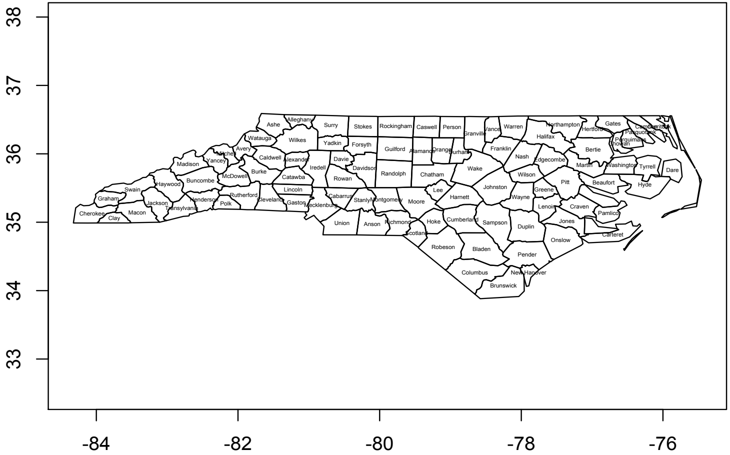
How we analyze areal data



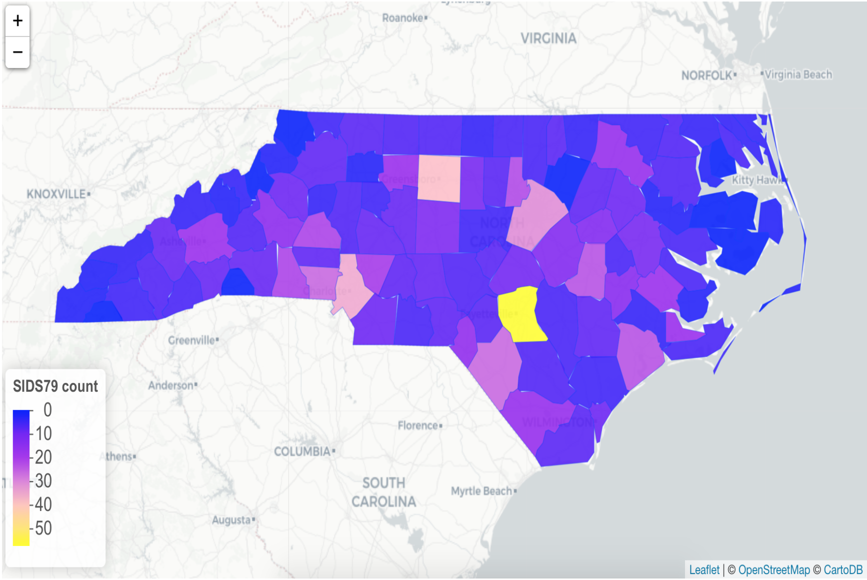
Areal Data Example: SIDS

- Data for 100 counties in North Carolina
- Includes counts of live births and sudden infant deaths for two periods: July 1974-June 1978 and July 1979-June 1984.
- SIDS is defined as sudden death of infant up to 12 months old.
- Risk factors include race, SES, physiologic (respiratory, sleep rate, cardiac function)
- The primary analysis here is not only to see how often SIDS occurs, but where and if there are clusters or spatial patterns.

Sudden Infant Deaths in North Carolina



Areal Data Example: SIDS



Areal Data: Proximity

- We represent proximity between areal units (blocks, B_i) using connected graphs
- Adjacency matrix (proximity matrix) is denoted W
- The entries of W are w_{ij} and are called weights
- The w_{ij} connect different values of the process $Y_1, ..., Y_n$, $i = 1, ..., n$ in some fashion
- Generally w_{ii} is set to zero

Areal Data: Proximity

Examples of weights

1. Border based (edge connections): areal units are neighbours if they share a border
   - $w_{ij} = 1$ if i and j share common boundary
2. Distance based: areal units are neighbours if they are within a distance of ϵ of each other
   - $w_{ij} = 1$ if the centroid of i is distance ϵ (ex. 25km) of the centroid of j
   - $w_{ij} = 1$ if j is the nearest neighbour (smallest ϵ) of i
   - $w_{ij} = 1$ if j is one of the k nearest neighbours of i, e.g. the two and three closest areal units j to i are the k=2 and k=3 nearest neighbors of i. This will result in multiple neighbours for each i

Areal Data: Proximity

- Distance can be defined several ways:
  - Euclidean distance (or driving distance or driving time, etc) between centroids (straight line path)
  - Mean driving distance, mean driving time, walking distance, etc. (transit path, not necessarily a straight line)
- The connections between blocks can be examined using a connected graph

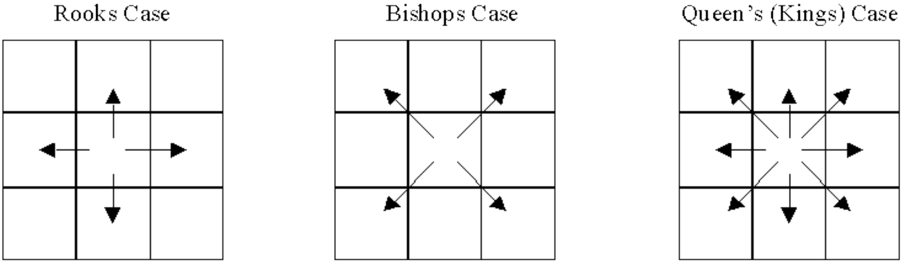
Areal Data: Proximity

Border/Edge based, binary connectivity. Two areal units are neighbours if they share a border

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ share a boundary} \\ 0 & \text{otherwise} \end{cases}$$

Where $w_{ij} = w_{ji}$ (symmetric)

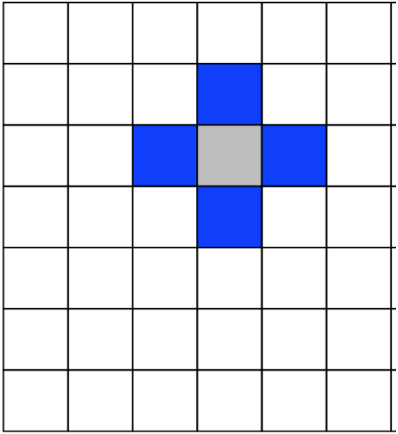
Border/Edge Connectivity



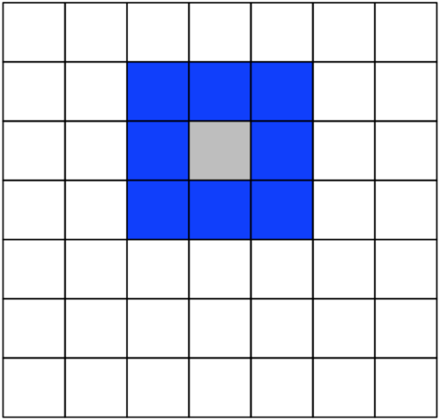
Queen a single shared boundary point means they are neighbours. Rook requires more than a single shared point to constitute neighbours.

Areal Data

Border/Edge Connectivity



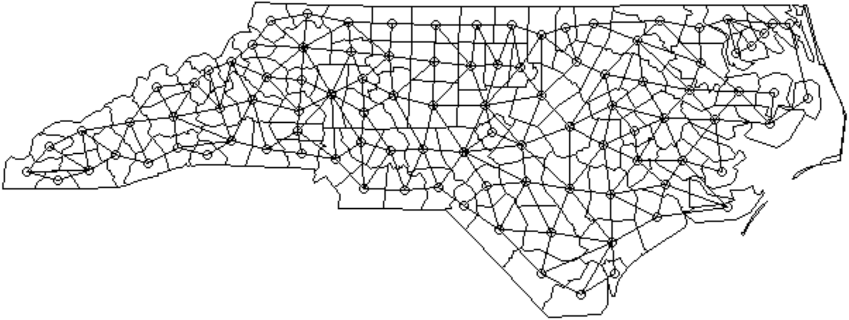
Rook



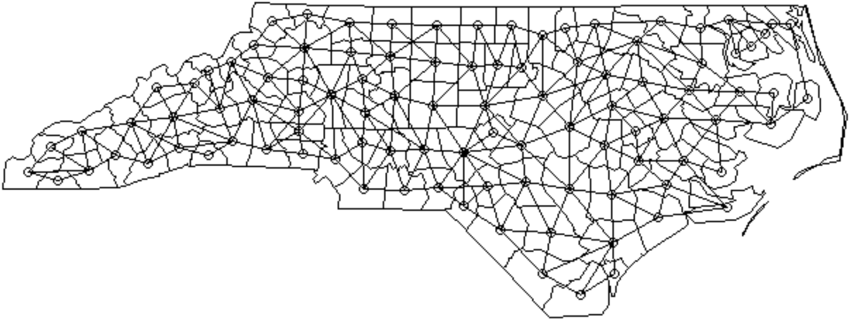
Queen

Areal Data

Border/Edge Connectivity: Rook

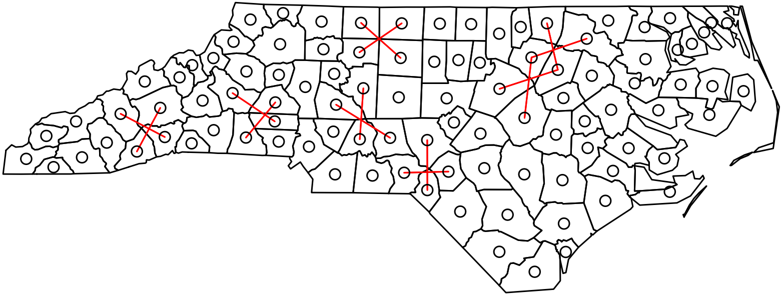


Border/Edge Connectivity: Queen



Areal Data

Border/Edge Connectivity: Difference Rook-Queen



Areal Data: Proximity

Fractional borders

$$w_{ij} = \begin{cases} \frac{l_{ij}}{l_i} & \text{if regions } i \text{ and } j \text{ share a border} \\ 0 & \text{otherwise} \end{cases}$$

Where l_{ij} is the length of the common border between regions i and j, and l_i is the perimeter of region i.

Areal Data: Proximity

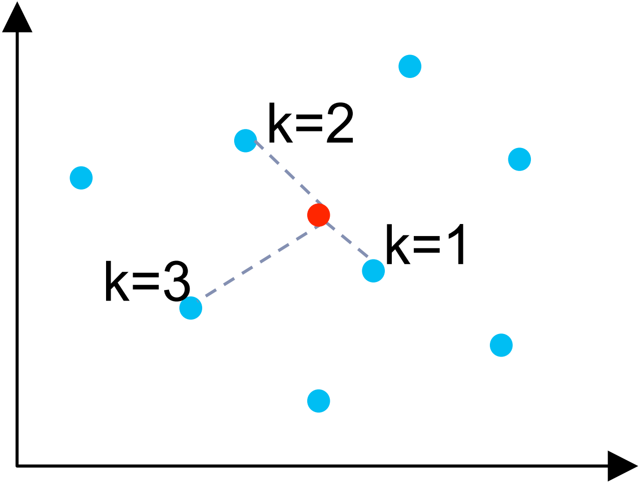
Neighbour Based

$$w_{ij} = \begin{cases} 1 & \text{if centroid of } j \text{ is a } k \text{ nearest neighbour of } i \\ 0 & \text{otherwise} \end{cases}$$

Where w_{ij} and w_{ji} not necessarily symmetric

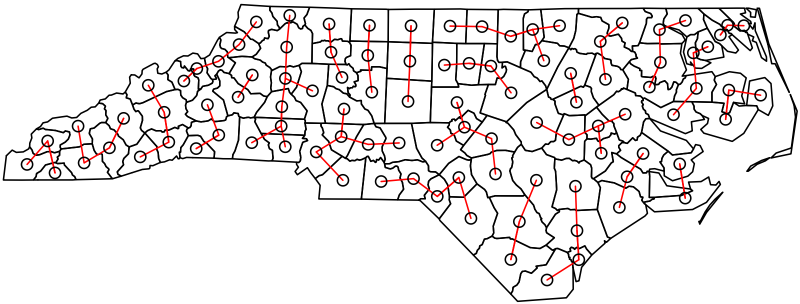
Areal Data: Proximity

k Nearest Neighbours (kNN)

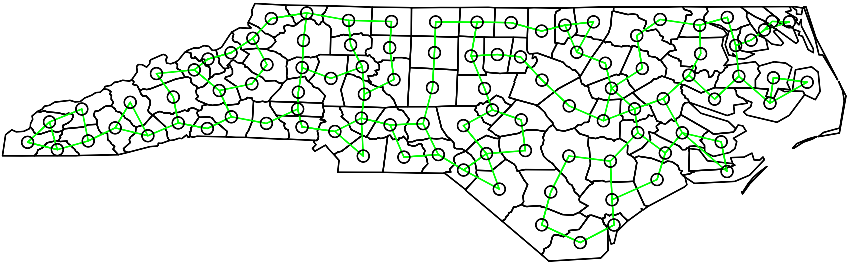


Areal Data: Proximity

Neighbour Based: 1NN

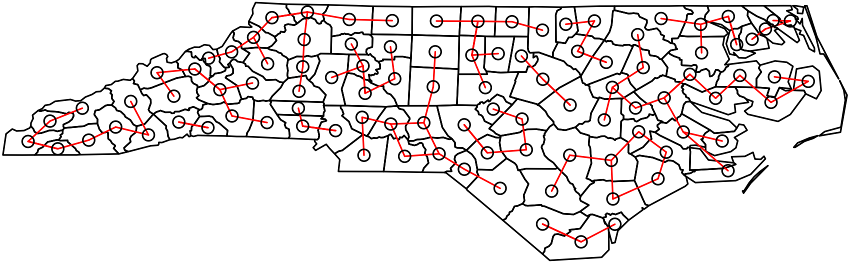


Neighbour Based: 2NN



Areal Data: Proximity

Neighbour Based: Difference Between 1NN and 2NN



Areal Data: Proximity

Distance Based

$$w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

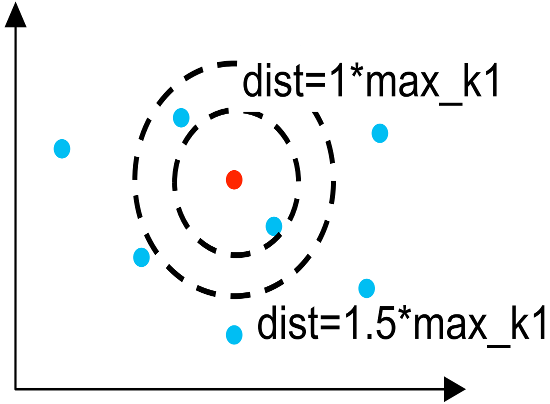
For some specified distance threshold ϵ
Alternatively,

$$w_{ij} = \begin{cases} d_{ij}^{-\rho} & \text{if } \rho > 0 \\ 0 & \text{otherwise} \end{cases}$$

For some power, ρ (recall idw)

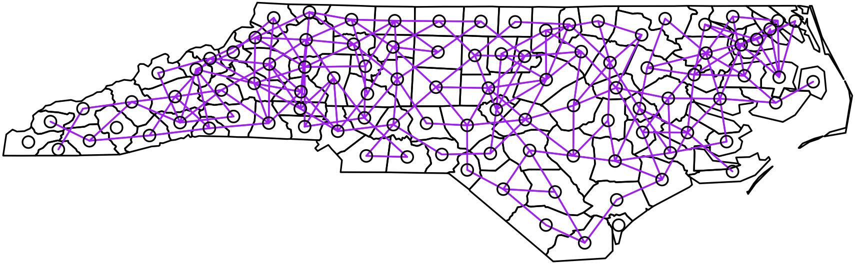
Areal Data: Proximity

Distance based neighbours, ϵ



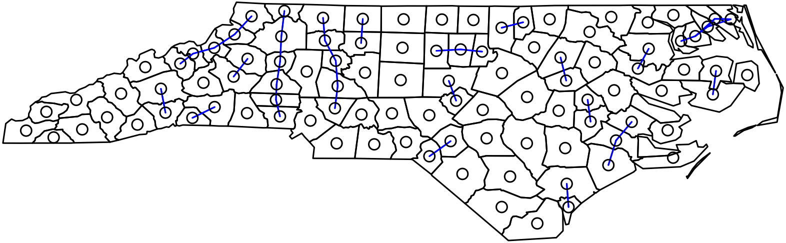
Areal Data: Proximity

Distance based neighbours, ϵ between 1 and 1.5 times maximum kNN distance



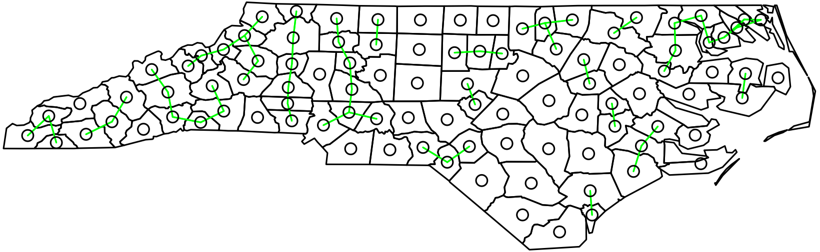
Areal Data: Proximity

Distance based neighbours, $\epsilon = 30$km (i.e. counties connected if 30 km or less apart)



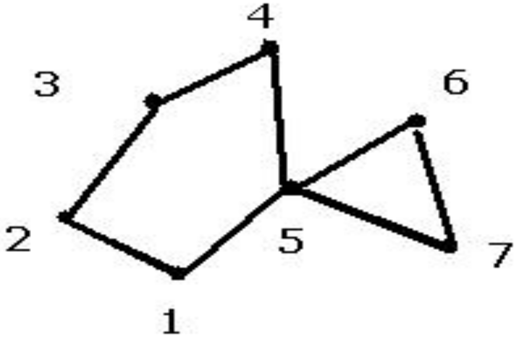
Areal Data: Proximity

Distance based neighbours, ϵ between 10 and 30km



Areal Data: Adjacency

Creating the adjacency matrix from connectivity graphs



Areal Data

Neighbours

1	2	5		
2	1	3		
3	2	4		
4	3	5		
5	1	4	6	7
6	5	7		
7	6	5		

Areal Data: Weights and the Adjacency Matrix

- The adjacency matrix, W is a matrix of neighbours where elements are weights w_{ij}
- Once our list of neighbours (fixed distance or kNN) has been created, we assign spatial weights to each relationship
- Can be binary or variable
- Even when the values are binary 0/1, there is the issue of what to do with no-neighbour observations arises
- Binary weighting will assign a value of 1 to neighboring features and 0 to all other features

Areal Data: Weights and the Adjacency Matrix

Binary weights

- Binary weights vary the influence of observations
- Those with many neighbours are up-weighted compared to those with few

$$\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{array}$$

Areal Data: Weights and the Adjacency Matrix

Row standardization is used to create proportional weights in cases where features have an unequal number of neighbors

- Row-standardized weights increase the influence of links from observations with few neighbours
- Divide each neighbour weight for a feature by the sum of all neighbour weights
- Obs i has 3 neighbours, each has a weight of $1/3$
- Obs j has 2 neighbours, each has a weight of $1/2$
- Use is you want comparable spatial parameters across different data sets with different connectivity structures

0	1	0	0
0	0	0.5	0.5
0.5	0.5	0	0
0	0.33	0.33	0.33

Areal Data: Weight matrix

Binary weight matrix

$$\begin{matrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{matrix}$$

Areal Data: Weight matrix

Row standardized weight matrix

$$\begin{array}{ccccccc}
0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 \\
0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\
0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\
0.25 & 0 & 0 & 0.25 & 0 & 0.25 & 0.25 \\
0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 \\
0 & 0 & 0 & 0 & 0.5 & 0.5 & 0
\end{array}$$

Areal Data: Spatial Smoothers

We can use the block values and weight matrices to obtain a smooth value for each region by taking *locally weighted averages*

- If we have a measure of Y_i, such as the SIDS rate in county i, we can get a rough estimate of what it could be predicted as from it's j neighbours
- Essentially, we replace Y_i with $\hat{Y}_i$ where

$$\hat{Y}_i = \frac{1}{\sum_j w_{ij}} \sum_j w_{ij} Y_j$$

- The "new" $\hat{Y}_i$ is a function of it's spatial neighbours j
- This smooths things out because the areal units look more like their neighbours

Areal Data: Spatial Similarity

- We want to summarize similarity between nearby areal units
- Spatial autocorrelation is the the correlation of the same measurement taken at different areal units
- The similarity of values at locations B_i and B_j are weighted by the proximity of i and j
- The weight w_{ij} defines proximity

Areal Data: Spatial Association

Measuring strength of association

- We want to measure how strong observations from nearby areal units are more or less alike than those that are farther apart
- We also want to decide whether the similarity (or dissimilarity) is strong enough that it is not due to chance
- For example:
  - Let Y_i be the response at the ith areal unit, B_i and Y_j be the response at the jth areal unit, B_j
  - Let sim_{ij} be a measure of how similar (or dissimilar) the responses are at areal units B_i and B_j
  - Let w_{ij} be a measure of the spatial proximity between areal units B_i and B_j
- We can define a general statistic by the cross product of the sim_{ij} matrix and w_{ij} matrix

Areal Data: Spatial Association

Example con't:
Y=

0	1	0
0	0	1
1	1	0

define $\text{sim}_{ij} = (Y_i - Y_j)^2$ and

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ share a boundary} \\ 0 & \text{otherwise} \end{cases}$$

Areal Data: Spatial Association

Example con't:
W=

$$\begin{matrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{matrix}$$

Find pairwise similarity from Y and take the cross product to get
C=$\sum_i \sum_j w_{ij} sim_{ij}$

Areal Data: Spatial Association

Example con't

- If C is small that means similarity between neighbours is high and we have positive spatial autocorrelation
- If C is large that means there is little similarity between neighbours

Measuring strength of association

- In general the extent of similarity is represented by the weighted average of similarity between areal units:

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} sim_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}$$

Areal Data: Global Indexes of Spatial Autocorrelation

- The goal of global indexes of spatial autocorrelation is to summarize the degree to which similar observations tend to occur near each other
- Global indexes are summaries over the entire study area, akin to testing clustering rather than a test to detect individual clusters
- Indexes share a common structure: calculate the similarity of values at locations i and j then weight the similarity by the proximity of locations i and j
- High similarities with high weight indicate similar values that are close together; low similarities with high weight indicate dissimilar values that are close together

Areal Data: Global Indexes of Spatial Autocorrelation

Indexes of spatial autocorrelation

- We want to summarize similarity between nearby areal units
- Spatial autocorrelation is the the correlation of the same measurement taken at different areal units
- The similarity of values at locations B_i and B_j are weighted by the proximity of i and j
- The weight w_{ij} defines proximity
- In general the extent of similarity is represented by the weighted average of similarity between areal units: indexes of spatial autocorrelation are built on this basic form:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} sim_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

Moran's I

- Moran's I (1950) follows the basic form for global indexes of spatial autocorrelation with similarity between areal units i and j defined as the product of the respective difference between y_i and y_j with the overall mean
- Similarity $sim_{ij} = (y_i - \bar{y})(y_j - \bar{y})$
- Where $\bar{y} = \sum_{i=1}^{n} y_i/n$
- Divide the basic form by the sample variance to get the Moran's I statistic:
- $I = \frac{1}{s^2}\frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i \sum_j w_{ij}}$
- Where $s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n}$

Areal Data: Global Indexes of Spatial Autocorrelation

Moran's I

- I is a random variable having a distribution defined by the distributions of and interactions between the y_i
- When neighbouring regions have similar values (pattern is clustered), I will be positive
- When neighbouring regions have different values (pattern is regular), I will be negative
- When there is no correlation between neighbouring values: $E(I) = -\frac{1}{n-1}$
- When n$\rightarrow \infty$, E(I) $\rightarrow 0$
- I is asymptotically normally distributed where $\frac{I+\frac{1}{n-1}}{\sqrt{Var(I)}} \sim N(0,1)$

Moran's I

- Moran's I is similar to Pearson's correlation but it is not bounded on [-1,1] because of the spatial weights
- Null hypothesis: NO spatial association, i.e. y_i iid
- Compare the z-score to a standard normal distribution
- The z-score that we compare to the standard normal is $z = \frac{I-E(I)}{\sqrt{Var(I)}}$ where $E(I) = -\frac{1}{n-1}$ and V(I) is a little complicated (shown later)
- In R use `moran.test` in `library(spdep)`. We look at Moran's I standard deviate (which is N(0,1) so comared to z-value) and associated p-value.

Moran's I in R using North Carolina SIDS data

```
# Define neighbours.  Choose k=2 NN
IDs<-row.names(as(nc, "data.frame"))
sids.kn2<-knn2nb(knearneigh(coordinates(nc), k=2, RANN=FALSE),
row.names=IDs)

# Convert to weight matrix (row standardized)
sids.kn2.w<-nb2listw(sids.kn2, style="W")

# Use moran.test for Moran's I
moranSIDS<-moran.test(sids79.rate,sids.kn2.w)
```

Moran's I in R using North Carolina SIDS data

Result:

```
data:  sids79.rate
weights:  sids.kn2.w

Moran I statistic standard deviate = 2.4465, p-value = 0.007213
alternative hypothesis:  greater
Sample Estimates:
Moran I statistic Expectation Variance
0.214682382 -0.010101010 0.008442014
```

The null hypothesis of no spatial correlation is rejected.

Areal Data: Global Indexes of Spatial Autocorrelation

Geary's c

- Geary (1954) devides the contiguity ratio or Geary's c
- Similarity $sim_{ij} = (y_i - y_j)^2$
- If regions i and j have similar values, sim$_{ij}$ wil be small
- $c = \frac{n-1}{2\sum_i (y_i - \bar{y})^2} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\sum_i \sum_j w_{ij}}$
- Like Moran's it is a weighted average, but here it is scaled by a measure of the overall variation around the mean, $\bar{y}$

Areal Data: Global Indexes of Spatial Autocorrelation

Geary's c

- c ranges from 0 to 2 with 0 indicating perfect positive spatial correlation and 2 indicating perfect negative spatial correlation
- c is not a Pearson correlation (related to the Durbin-Watson statistic)
- Low values of Geary's c denote positive autocorrelation and high values indicate negative correlation.
- Expected value, $E(c) = 1$ under spatial independence.
- In R use `geary.test` in `library(spdep)`.

Areal Data: Global Indexes of Spatial Autocorrelation

Geary's c in R using North Carolina SIDS data

```
# Define neighbours.  Choose k=2 NN
IDs<-row.names(as(nc, "data.frame"))
sids.kn2<-knn2nb(knearneigh(coordinates(nc), k=2, RANN=FALSE),
row.names=IDs)

# Convert to weight matrix (row standardized)
sids.kn2.w<-nb2listw(sids.kn2, style="W")

# Use geary.test for Geary's c
gearySIDS<-geary.test(sids79.rate,sids.kn2.w)
```

Geary's c in R using North Carolina SIDS data

Result:

```
data:  sids79.rate
weights:  sids.kn2.w

Geary C statistic standard deviate = 1.6166, p-value = 0.05299
alternative hypothesis:  Expectation greater than statistic
sample estimates:
Geary C statistic Expectation Variance
0.83688116 1.00000000 0.01018165
```

We have a marginal p-value for positive spatial autocorrelation (Geary C closer to 0 (neg spatial) than 2 (pos spatial).