Spatial Statistics
Areal Data Unit 1

PM569 Spatial Statistics

Lecture October 20, 2017

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 (lattices), 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

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: Misrepresentation with maps

Crude birth rates by state based on equal-interval cut points

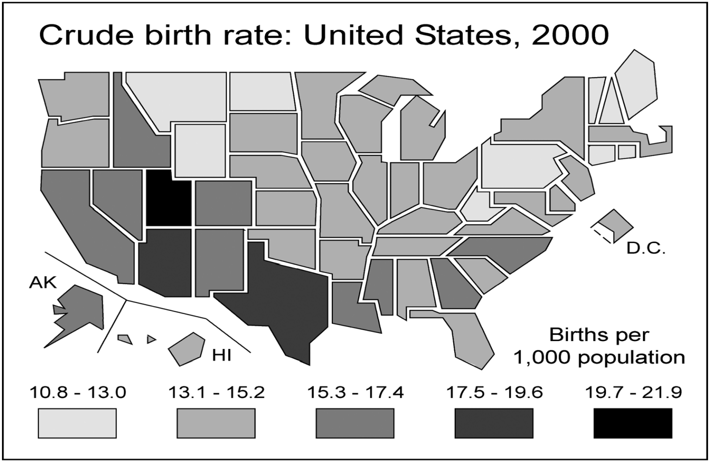


Figure: Monomier, N. Lying with Maps. Statistical Science 2005, 20(3) 215222.

Areal Data: Misrepresentation with maps

Crude birth rates by state based on quantile cut points

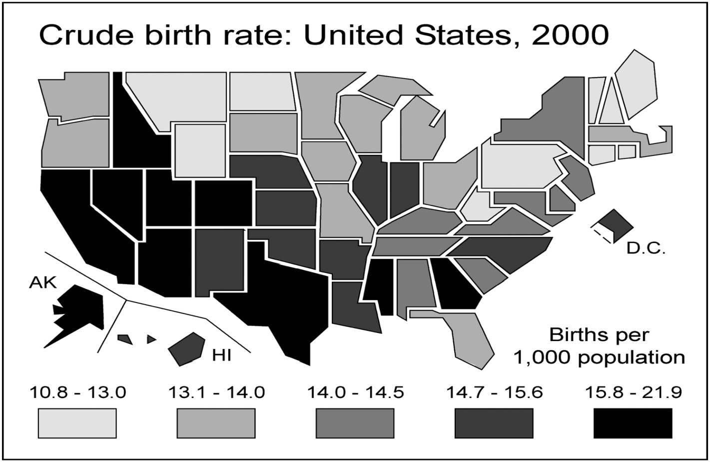
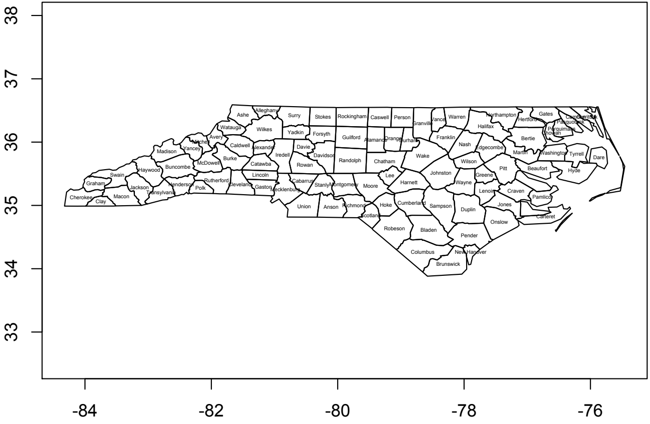


Figure: Monomier, N. Lying with Maps. Statistical Science 2005, 20(3) 215222.

Areal Data: Is there a spatial pattern?

- 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

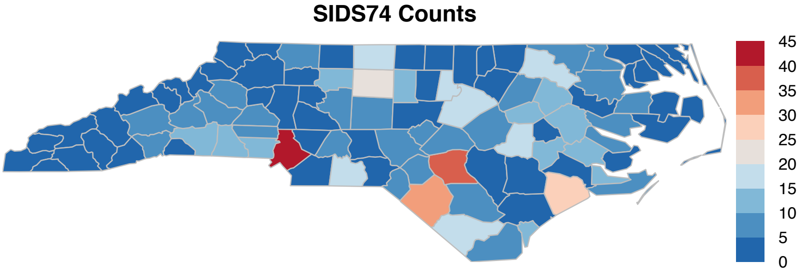
Sudden Infant Deaths in North Carolina



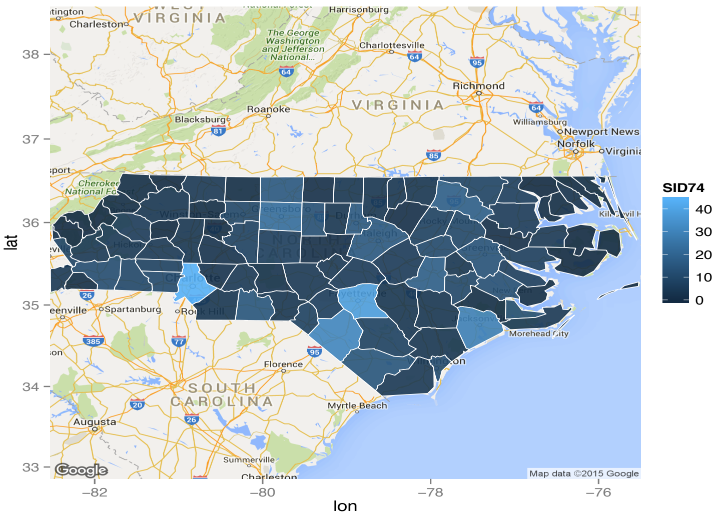
Areal Data Example: SIDS

- 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 to 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.

Areal Data Example: SIDS

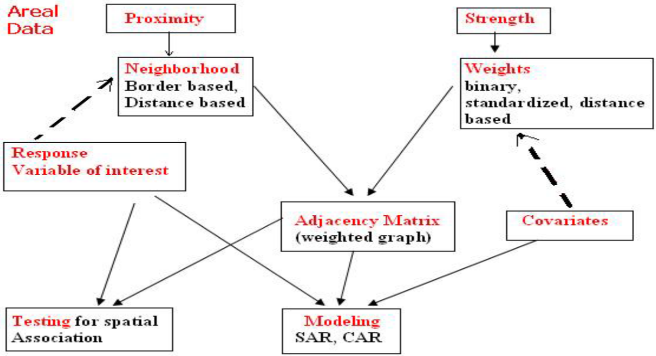


Areal Data Example: SIDS



Areal Data: Flowchart

How we analyze areal data



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

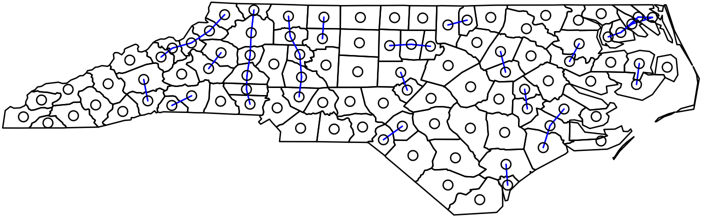
1. Border based (edge connections): areal units are neighbours if they share a border
   - $w_{ij} = 1$ if i and j share common boundary
2. Distance based: areal units are neighbours if they are within a distance of ϵ of each other
   - $w_{ij} = 1$ if the centroid of i is distance ϵ (ex. 25km) of the 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

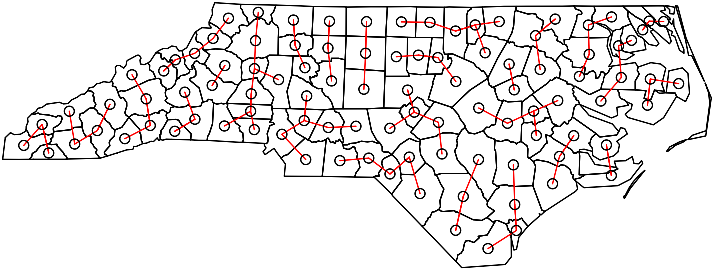
- Distance can be defined several ways:
  - Euclidean distance (or driving distance or driving time, etc) between centroids (straight line path)
  - Mean driving distance, mean driving time, walking distance, etc. (transit path, not necessarily a straight line)
- The connections between blocks under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

Areal Data: Distance Neighbours

Counties connected if 30km or less apart

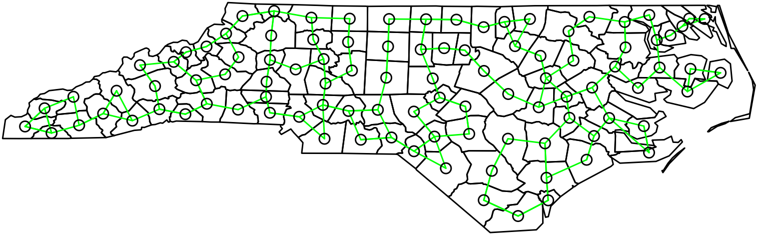


Counties each connected to its nearest neighbour



Areal Data: Distance Neighbours (2)

Counties each connected to its 2 nearest neighbours



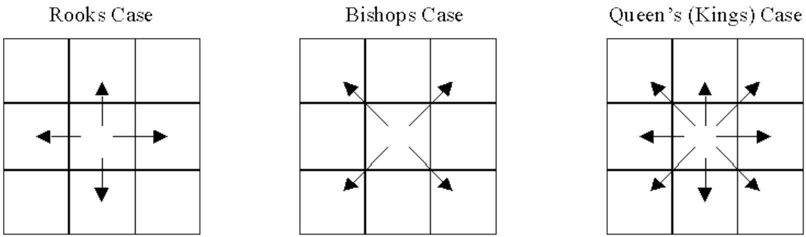
Areal Data: Proximity

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

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

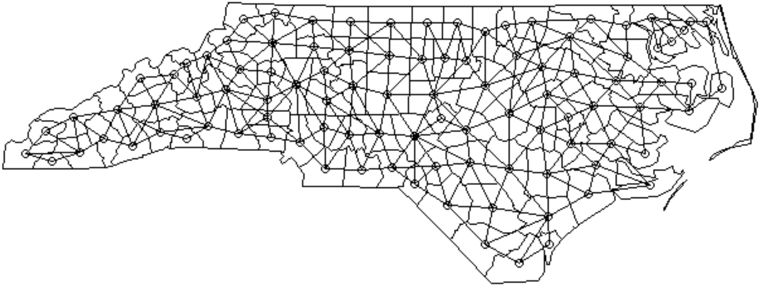
Where $w_{ij} = w_{ji}$ (symmetric)

Border/Edge Connectivity

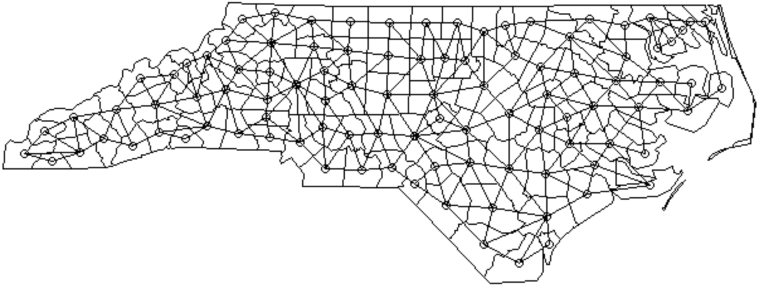


Queen a single shared boundary point means they are neighbours. Rook requires more than a single shared point to constitute neighbours.

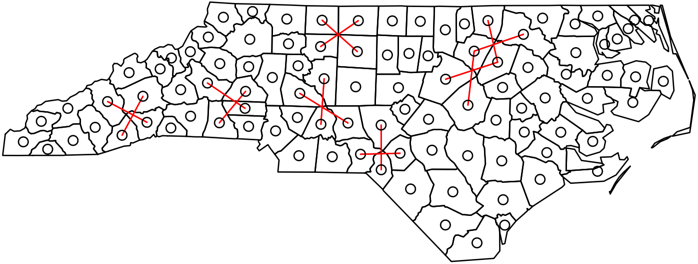
Border/Edge Connectivity: Queen



Border/Edge Connectivity: Rook



Border/Edge Connectivity: Difference Queen-Rook



Areal Data: Proximity

Fractional borders

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

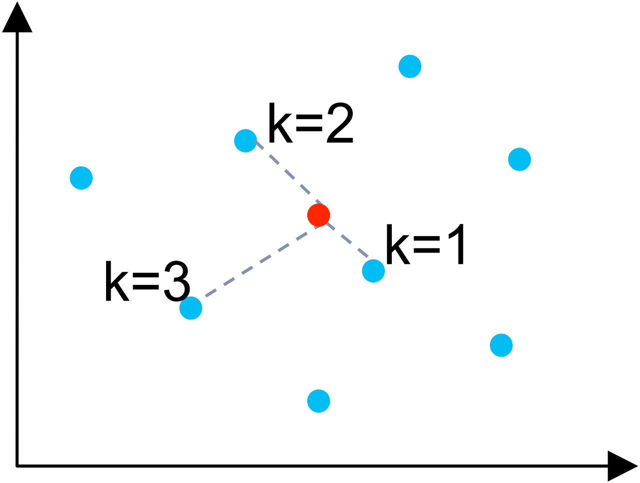
Where l_{ij} is the length of the common border between regions i and j, and l_i is the perimeter of region i.

Neighbour Based

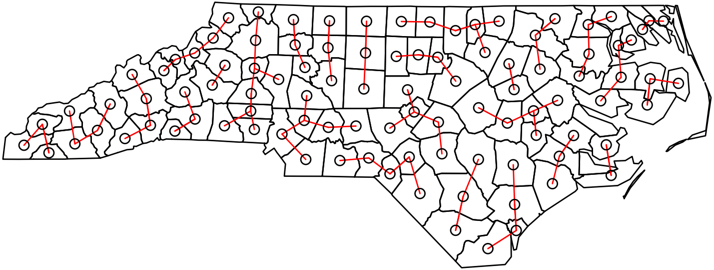
$$w_{ij} = \begin{cases} 1 & \text{if centroid of } j \text{ is a } k \text{ nearest neighbour of } i \\ 0 & \text{otherwise} \end{cases}$$

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

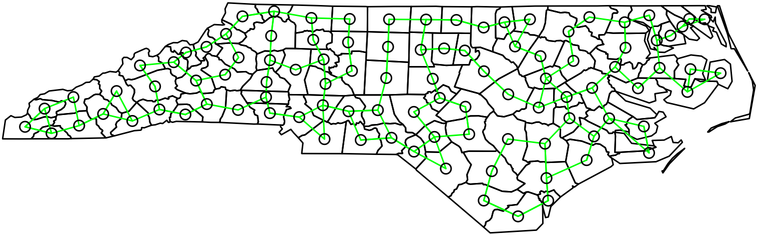
k Nearest Neighbours (kNN)



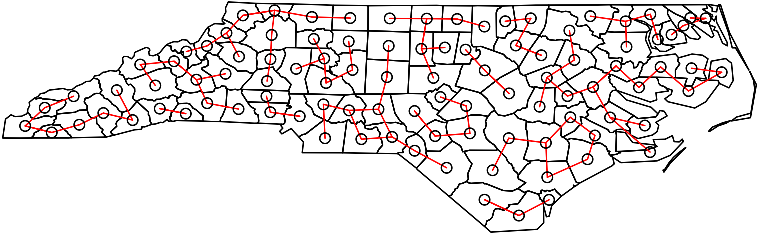
Neighbour Based: 1NN



Neighbour Based: 2NN



Neighbour Based: Difference Between 1NN and 2NN



Areal Data: Proximity

Distance Based

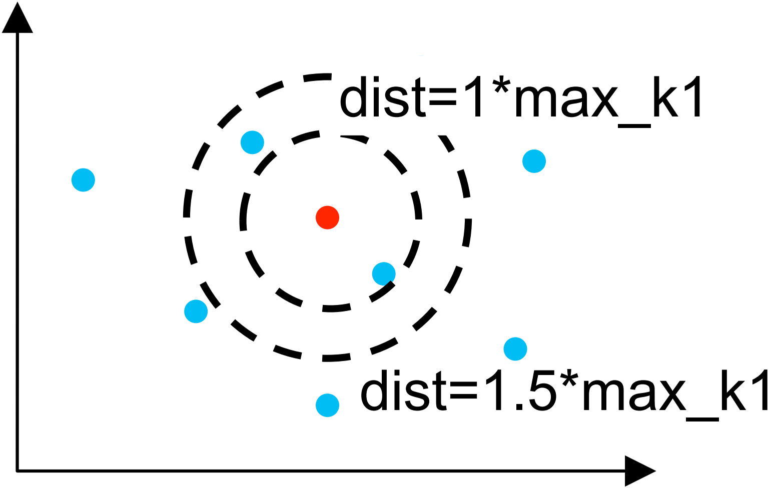
$$w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

For some specified distance threshold ϵ
Alternatively,

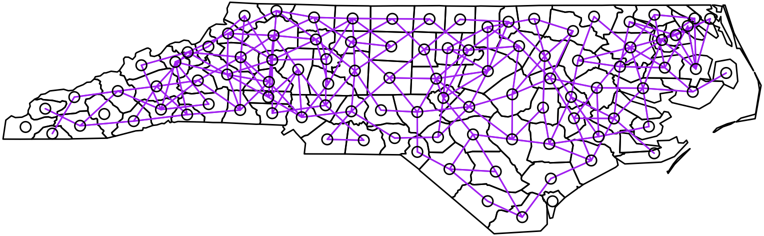
$$w_{ij} = \begin{cases} d_{ij}^{-\rho} & \text{if } \rho > 0 \\ 0 & \text{otherwise} \end{cases}$$

For some power, ρ (recall idw)

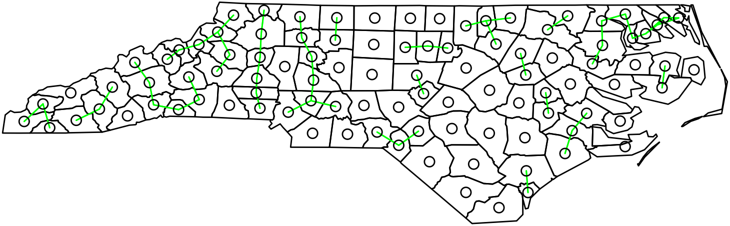
Distance based neighbours, ϵ



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

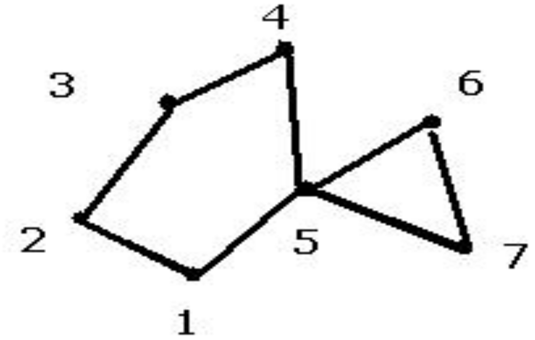


Distance based neighbours, ϵ between 10 and 30km



Areal Data: Adjacency

Creating the adjacency matrix from connectivity graphs



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

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

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

Example con't:

Y=

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

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=

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

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:

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