Spatial Data Analysis

Week 6: Areal Data I

Meredith Franklin

Department of Statistical Sciences and School of the Environment

October 10th, 2025

Areal Data I

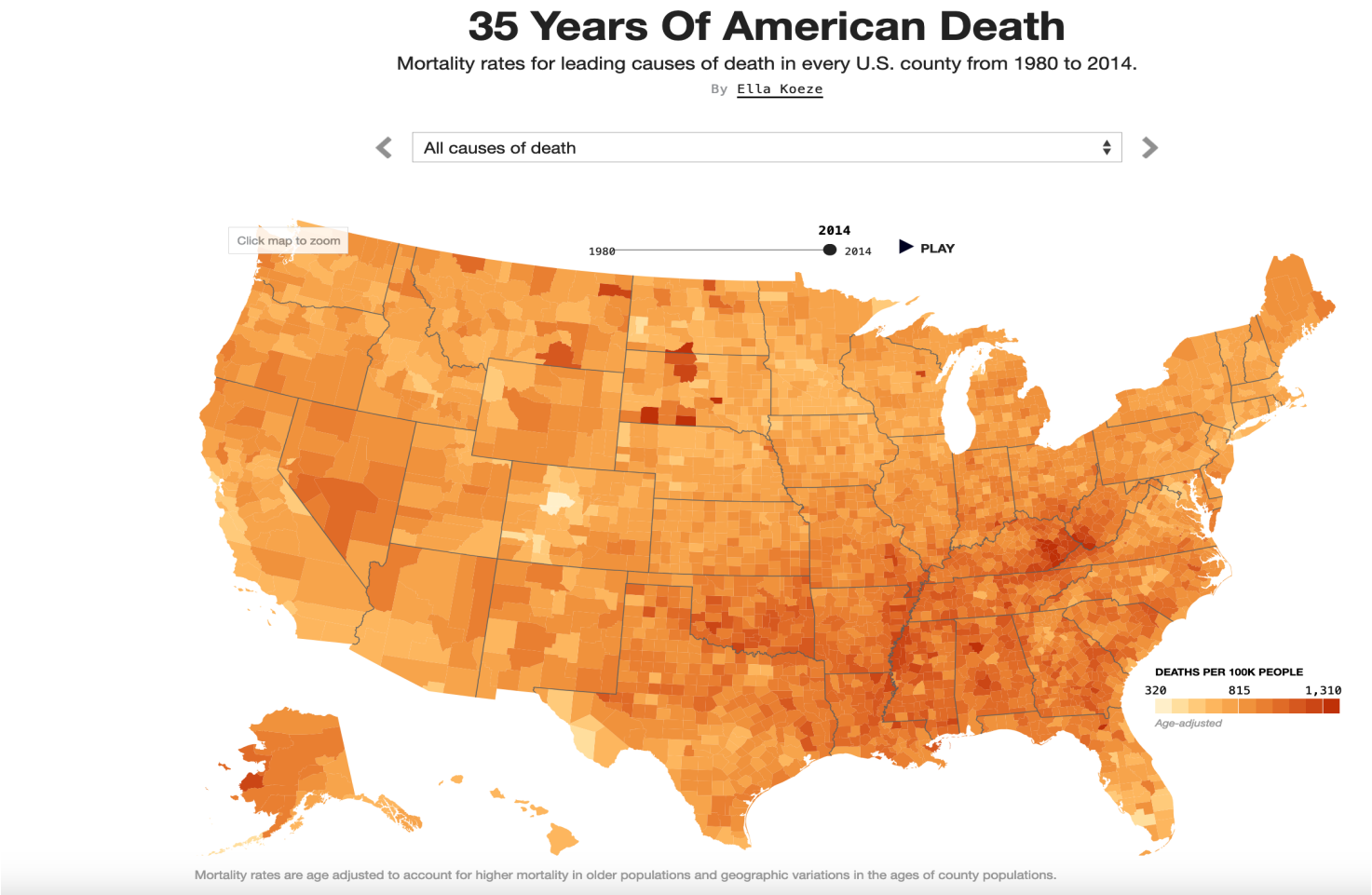
- Review of Areal Data Properties
- Definitions of Proximity
- Weights and Adjacency Matrices
- Indices of Spatial Autocorrelation

Areal Data Properties

- Data referenced at an aggregate level.
- Areal "units" are generally irregular geographic areas, and in spatial analysis we have a collection of areal units. Values are summaries for each areal unit (e.g., means, totals, rates); within-unit variation is not observed.
- Common areal data are census data (blocks, tracts, counties, states), postal codes, zip codes.
- We often have shapefiles (also simple features, GeoPackage (GPKG)) to deal with, as they are used for collections of polygons.
- Cautions: MAUP (modifiable areal unit problem) and ecological fallacy—critical when talking about aggregated data.

Areal Data: Example

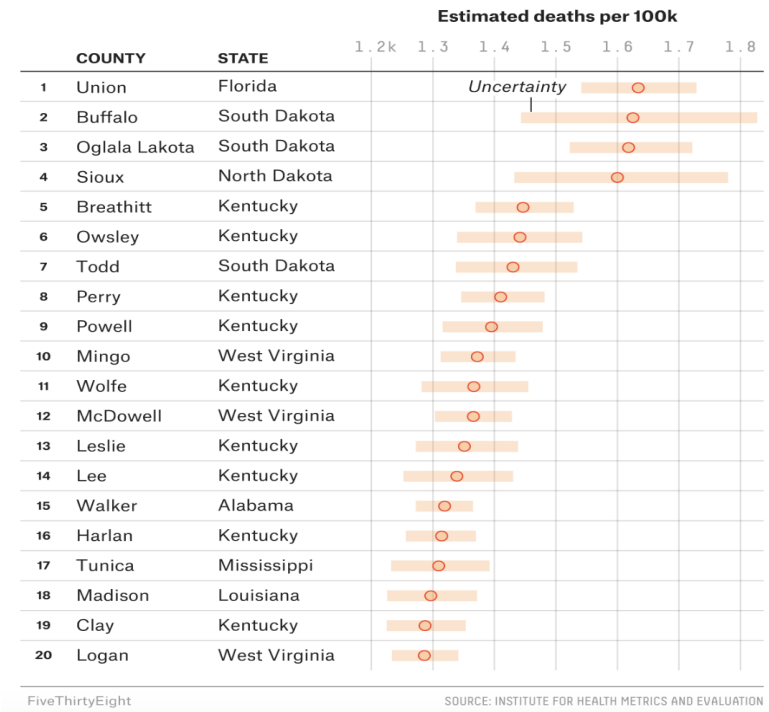
Cause-specific Mortality Rates by County



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

Areal Data: Example

- The Institute for Health Metrics and Evaluation (IHME), University of Washington, conducted a spatial analysis of these data (on a national level) `https://jamanetwork.com/journals/jama/fullarticle/2592499`.
- They used small-area estimation via generalized linear mixed-effects models with spatial and temporal random effects, fitted using the Template Model Builder (TMB) package in R.
- 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-function smoothing (e.g., splines).
- For areal data, we use neighbor 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 neighborhood information are more commonly used.
- We will use the R package `spdep` and `spatialreg` for areal data analysis.

Inferential issues with areal data

Is there a spatial pattern?

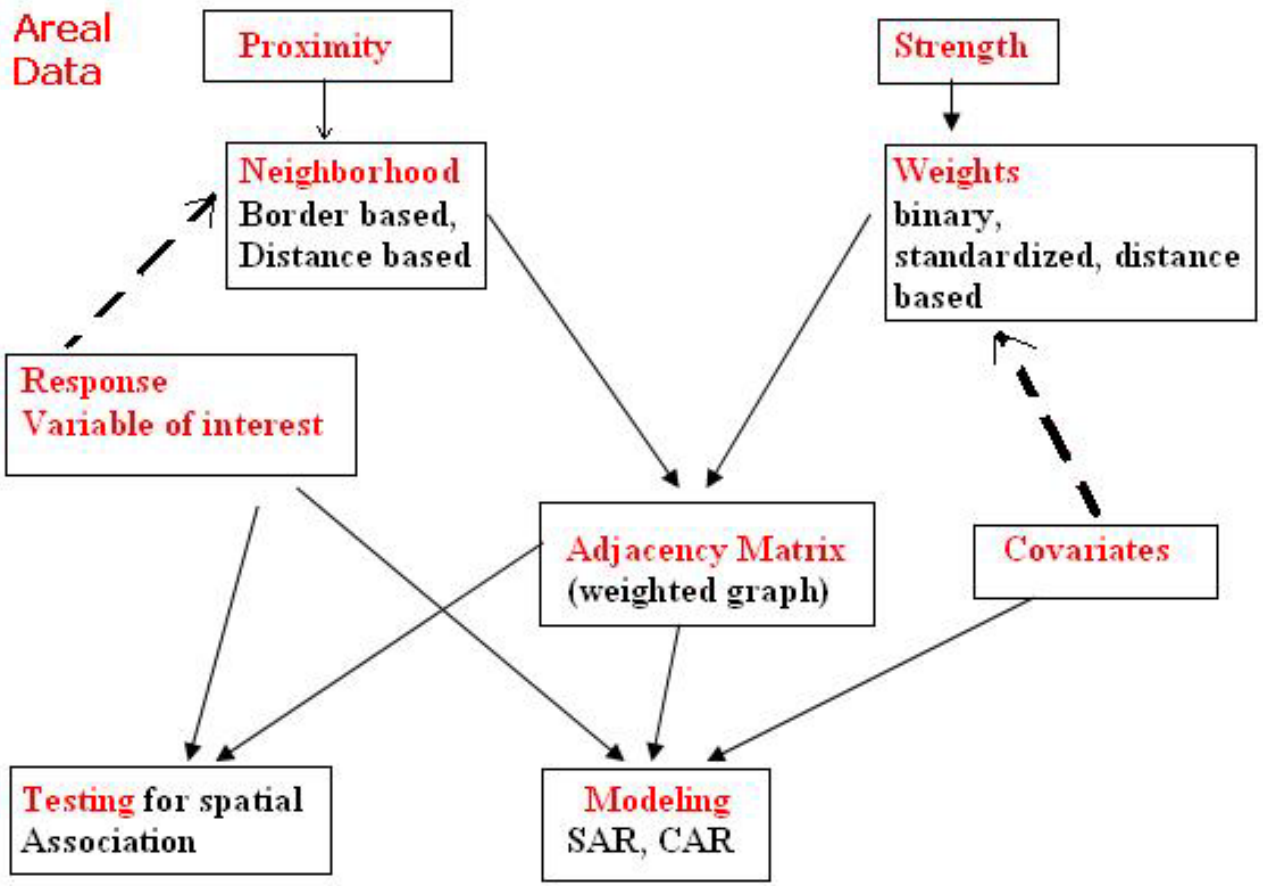
- A spatial pattern suggests that areal observations close to each other have more similar values than those far from each other.
- You might think that there is a pattern through visualization, but this is often subjective.
- Independent measurements will have no pattern, and would look completely random, but there may actually be an underlying pattern.
- If there is a spatial pattern, how strong is it? How do we quantify it? Is it a global or local spatial pattern?

Areal Data Analyses

- Response of interest Y_i measured in block or areal unit B_i
- The B_i are supplemented with neighborhood 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 Moran's I or Geary's C statistic
  - Testing for spatial pattern: Local testing using local Moran's I_i or Getis-Ord G_i^* 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

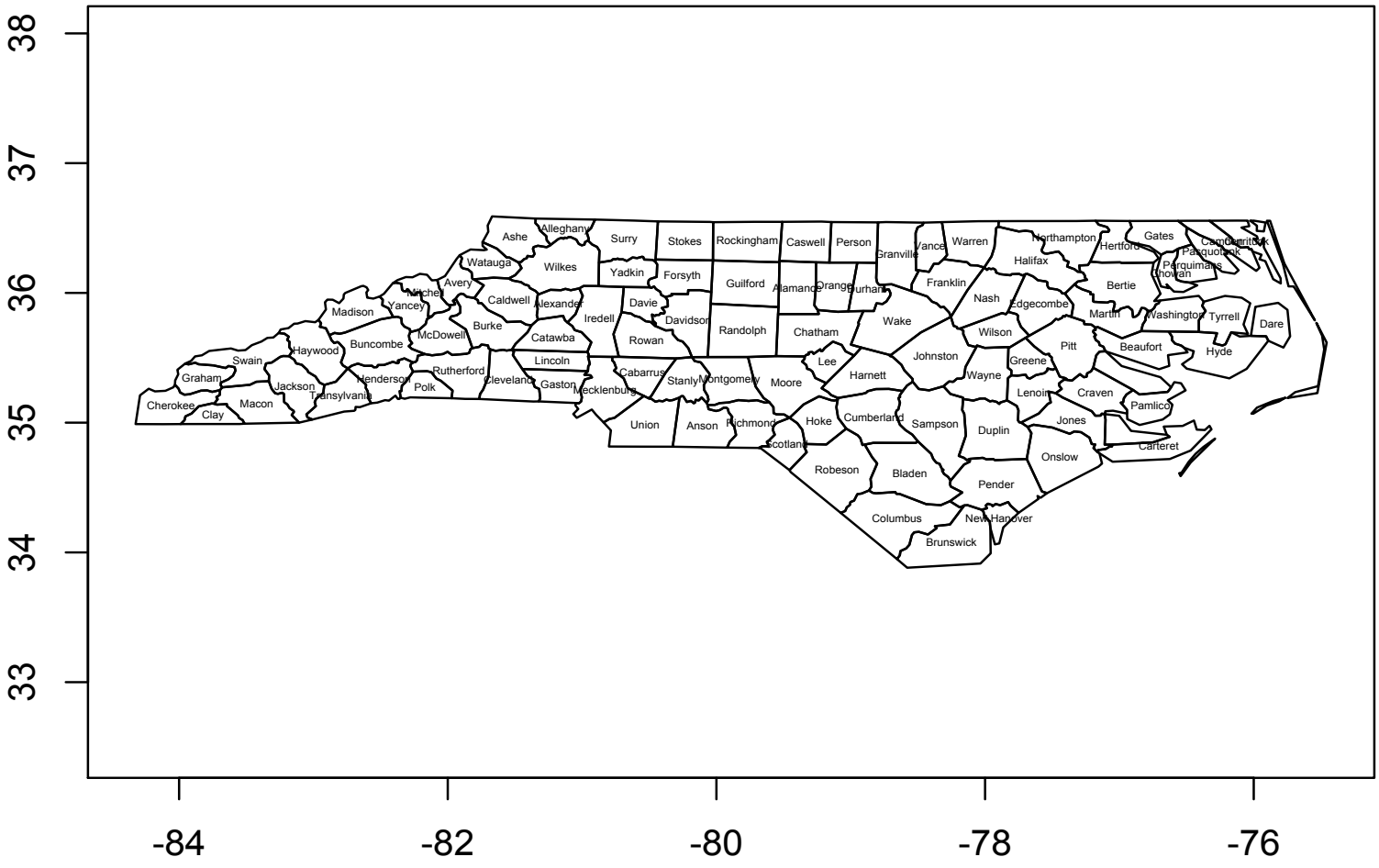


Areal Data Example: SIDS

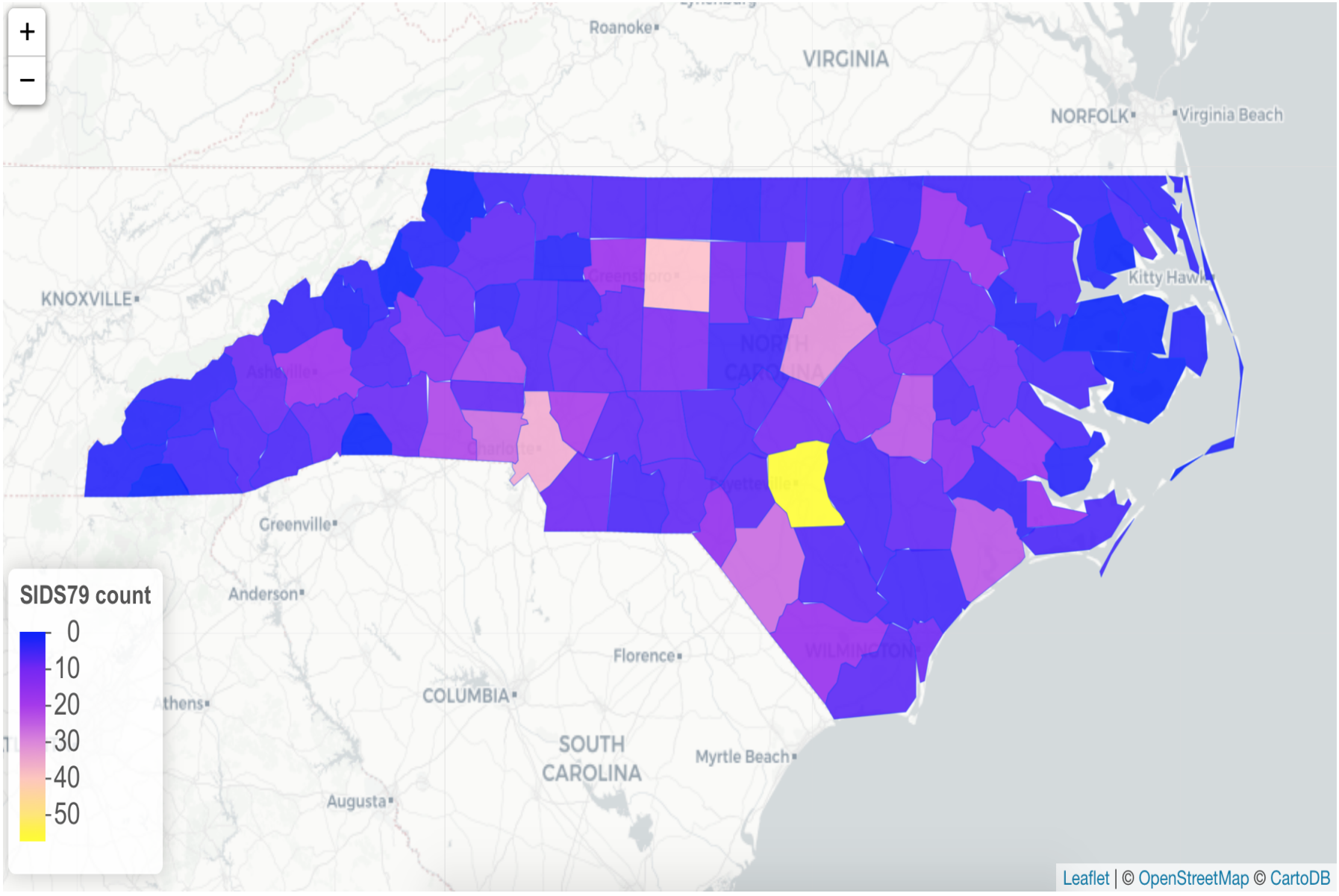
- Data for 100 counties in North Carolina
- Includes counts of live births and sudden infant deaths for two periods: July 1974-June 1978 and July 1979-June 1984.
- SIDS: sudden death of an infant ¡1 year that remains unexplained after thorough investigation (including autopsy, scene exam, and clinical history).
- 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.
- Rates should use counts with an offset (log births).

Areal Data Example: SIDS

Sudden Infant Deaths in North Carolina



Areal Data Example: SIDS



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. $W = [w_{ij}]$ denote the nxn matrix encoding neighbor relations amongh $Y_1, \ldots, Y_n$.
- That is, w_{ij} connect different values of the process $Y_1, \ldots, Y_n$, $i = 1, \ldots, n$ in some fashion.
- Generally w_{ii} is zero.

Proximity

Examples of weights

1. **Contiguity (border-based):** areal units are neighbors if they share a border (rook) or a border/vertex (queen)
   - $w_{ij} = 1$ if i and j share a common boundary (rook); optionally include shared vertex (queen)
2. **Distance-based:** neighbors defined by centroid distances.
   - $w_{ij} = 1$ if the centroid of j is *within* distance ε of the centroid of i
   - $w_{ij} = 1$ if j is the nearest neighbor of i
   - $w_{ij} = 1$ if j is among the k nearest neighbors of i (often asymmetric unless symmetrized)
   - As in geostatistics, distance can be defined in several ways: Euclidean or great-circle distance between centroids (straight-line); Network-based travel distance or time (driving, walking, transit).

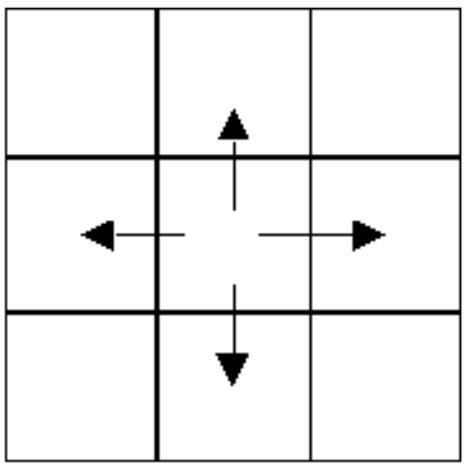
Binary contiguity: neighbors share a border

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

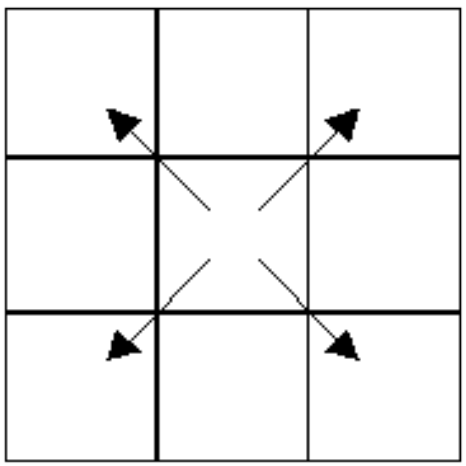
For contiguity weights, $w_{ij} = w_{ji}$ (symmetric).

Border/Edge Connectivity

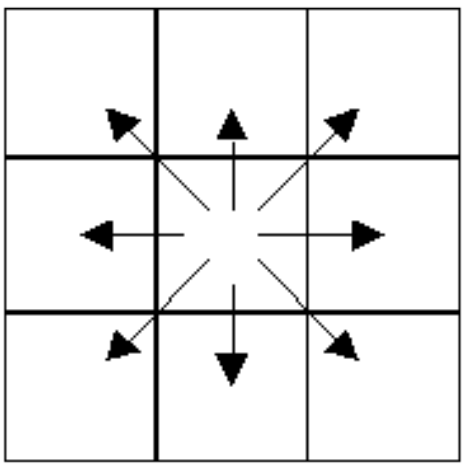
Rooks Case



Bishops Case

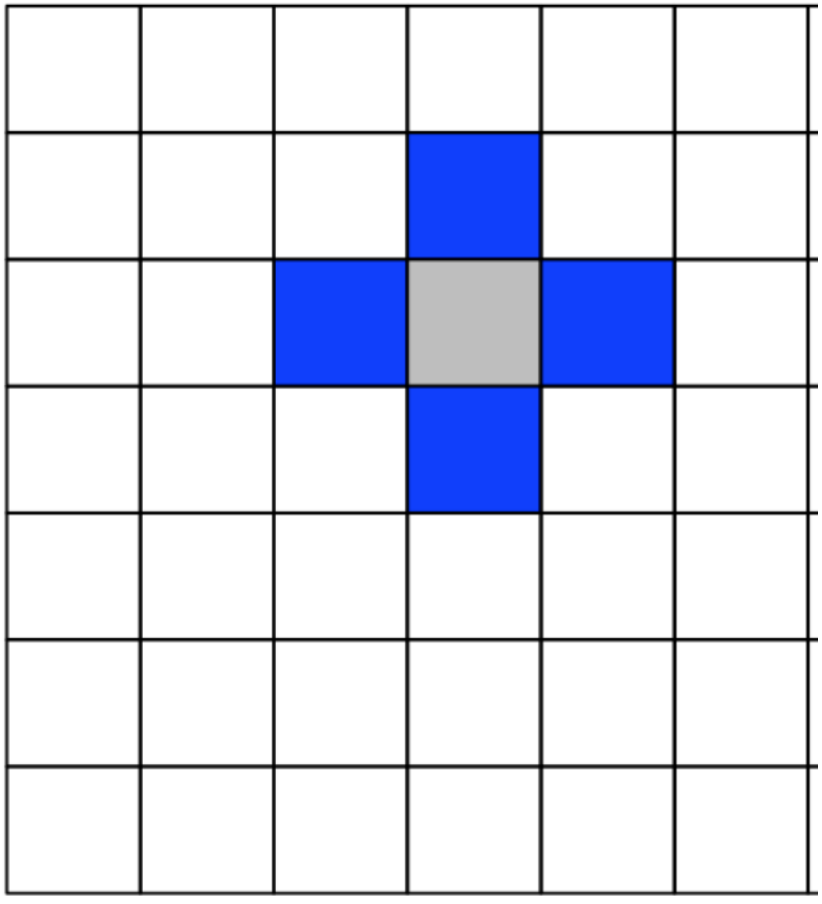


Queen's (Kings) Case

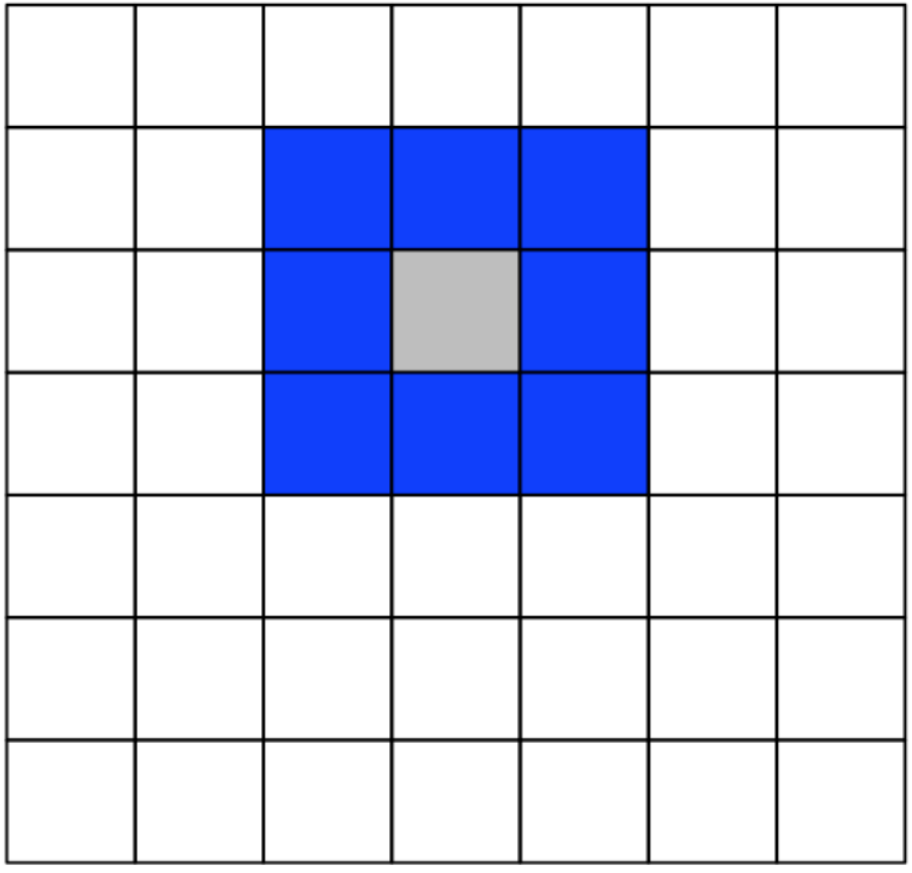


Queen a single shared boundary point means they are neighbors. Rook requires more than a single shared point to constitute neighbors.

Border/Edge Connectivity

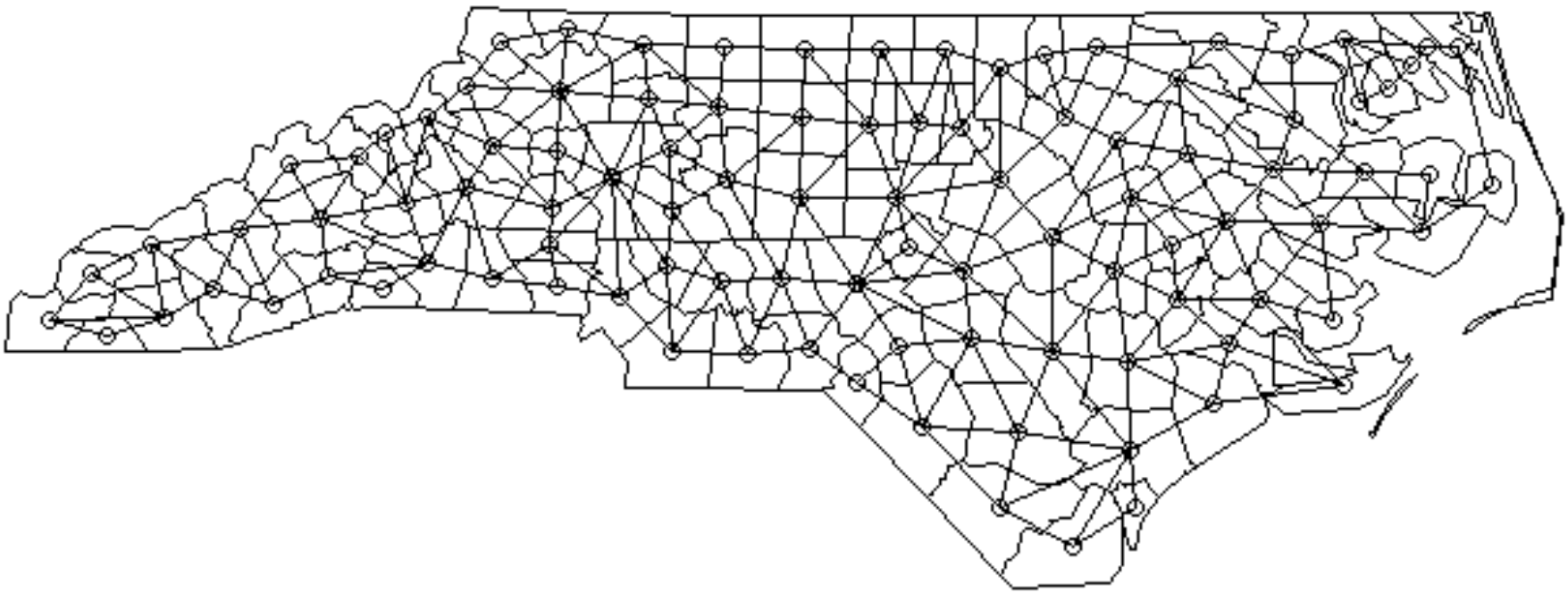


Rook

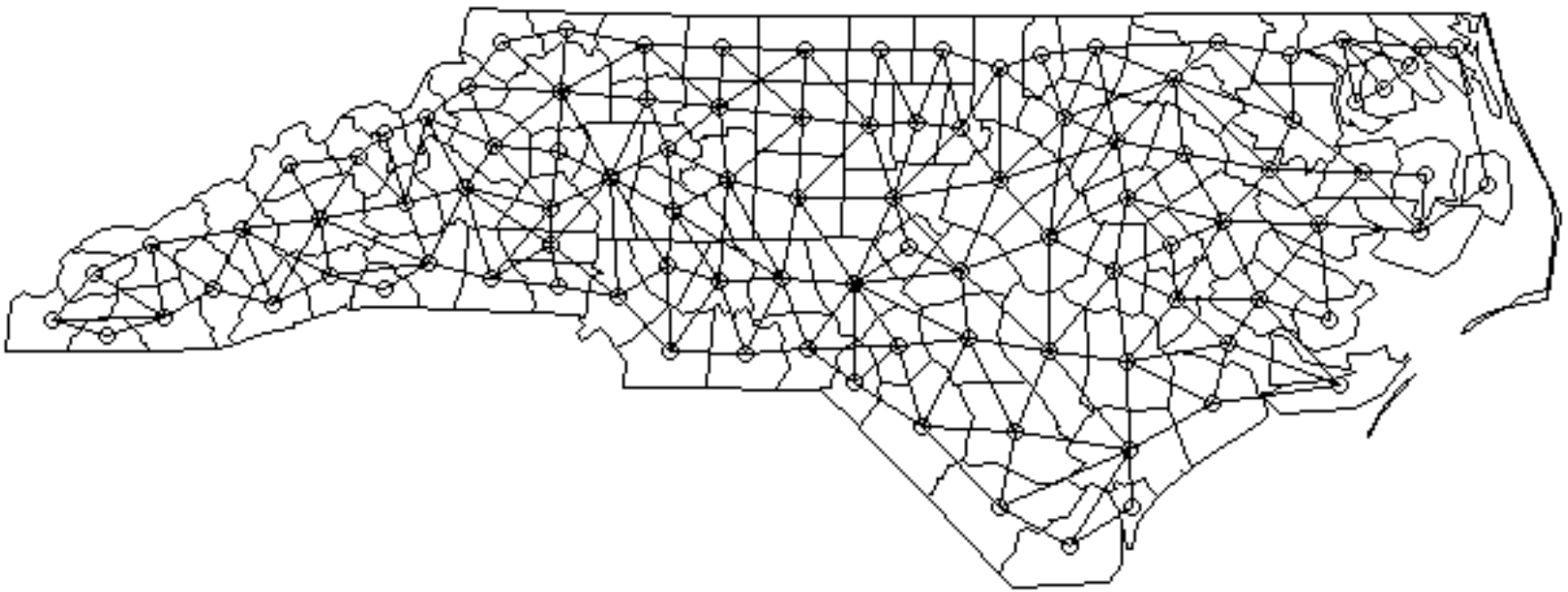


Queen

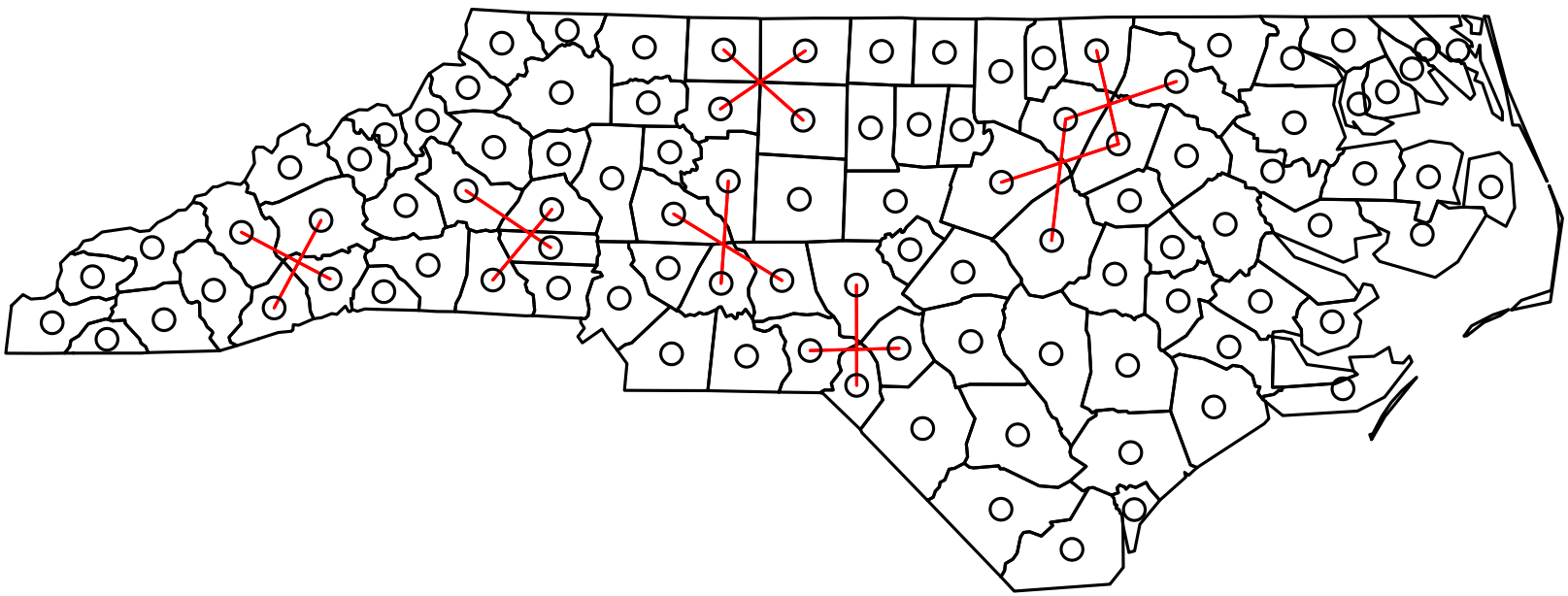
Border/Edge Connectivity: Rook



Border/Edge Connectivity: Queen



Border/Edge Connectivity: Difference Rook-Queen



Border-Based Proximity

Defining Fractional Borders

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

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

Neighbor-Based Proximity

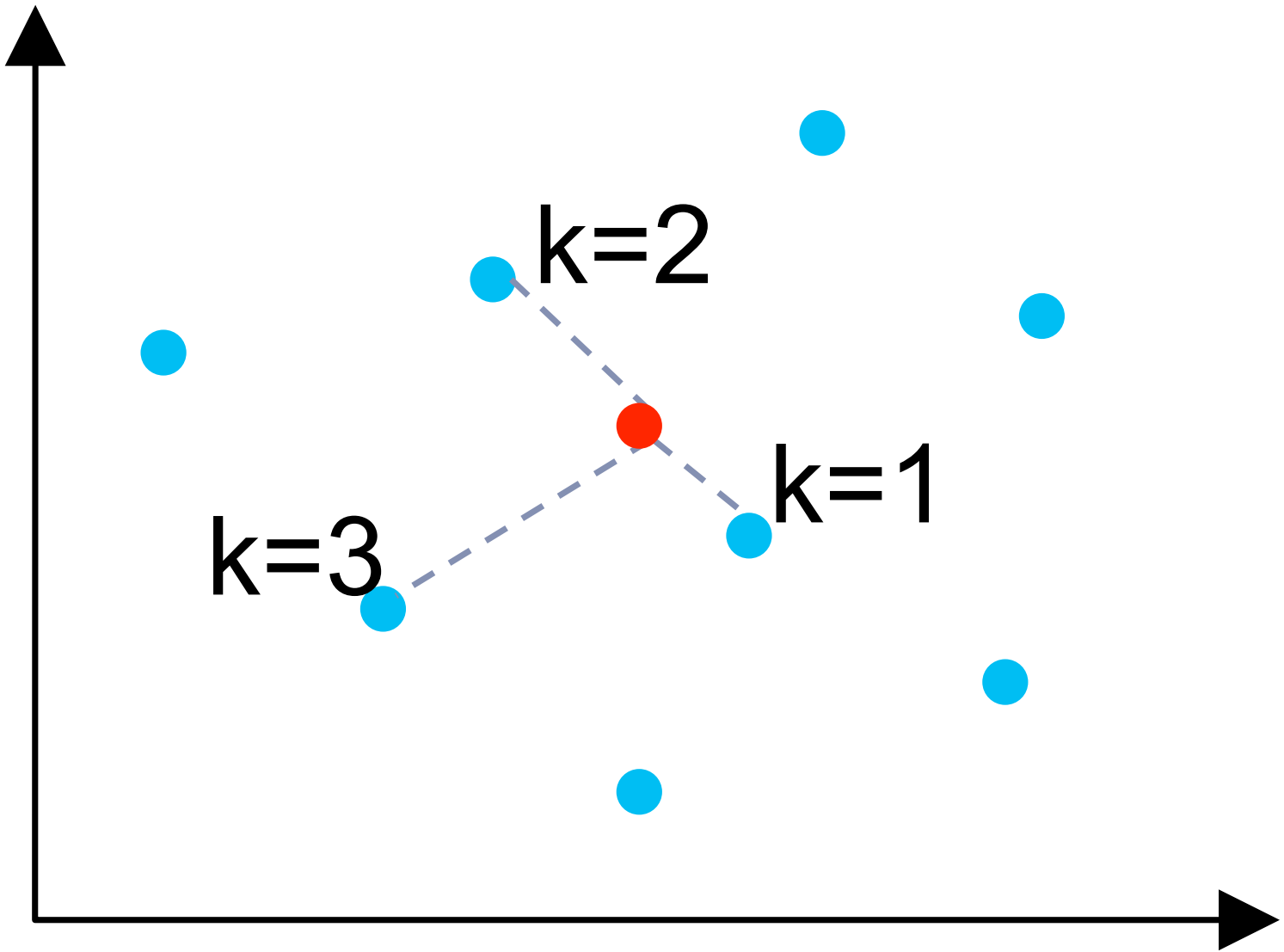
Centroid-Based Definition

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

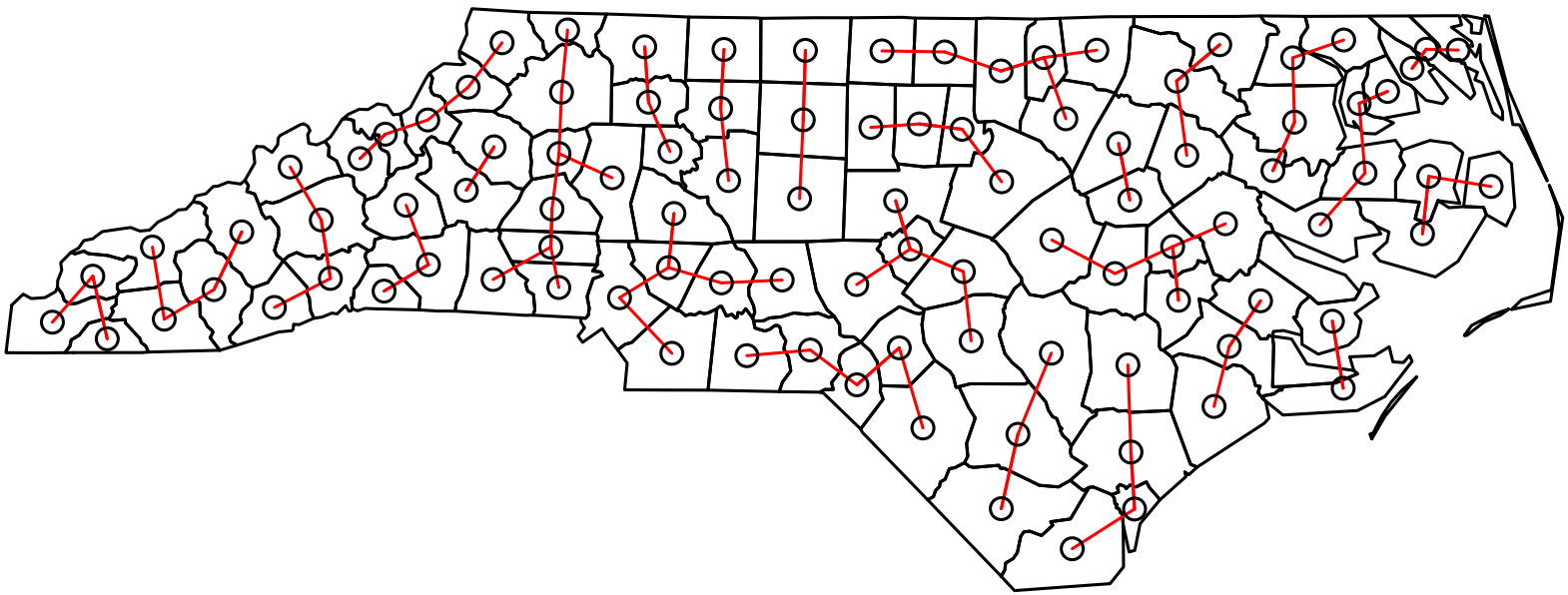
Weights w_{ij} and w_{ji} are not necessarily symmetric.

k-Nearest Neighbors (kNN) Proximity

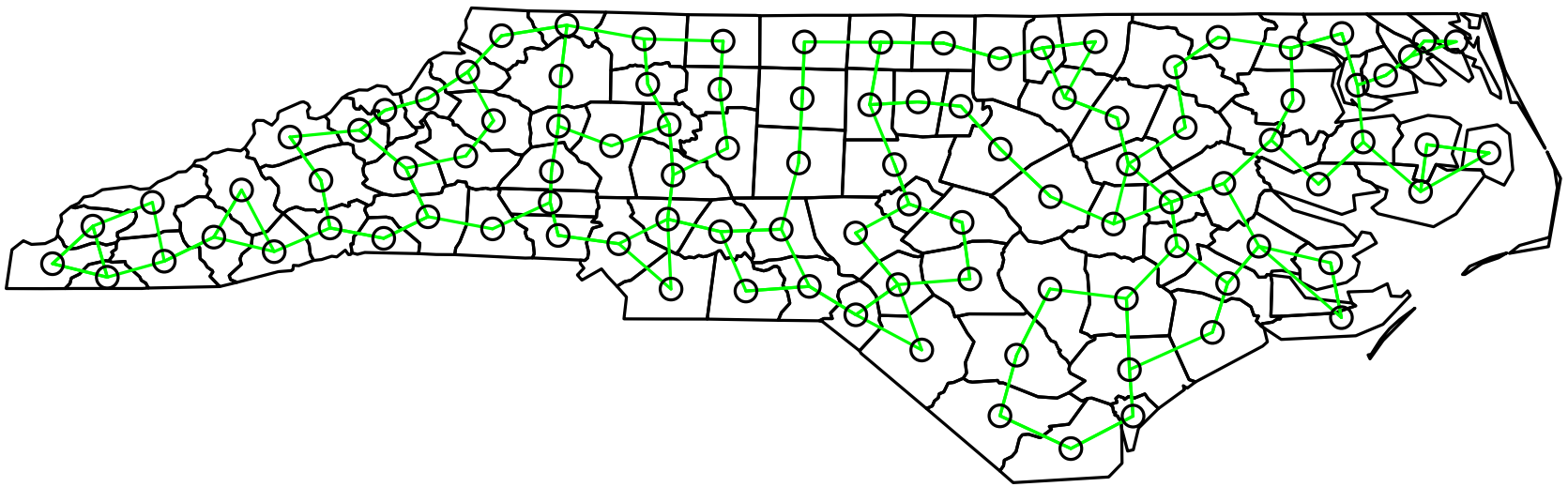
Conceptual Illustration



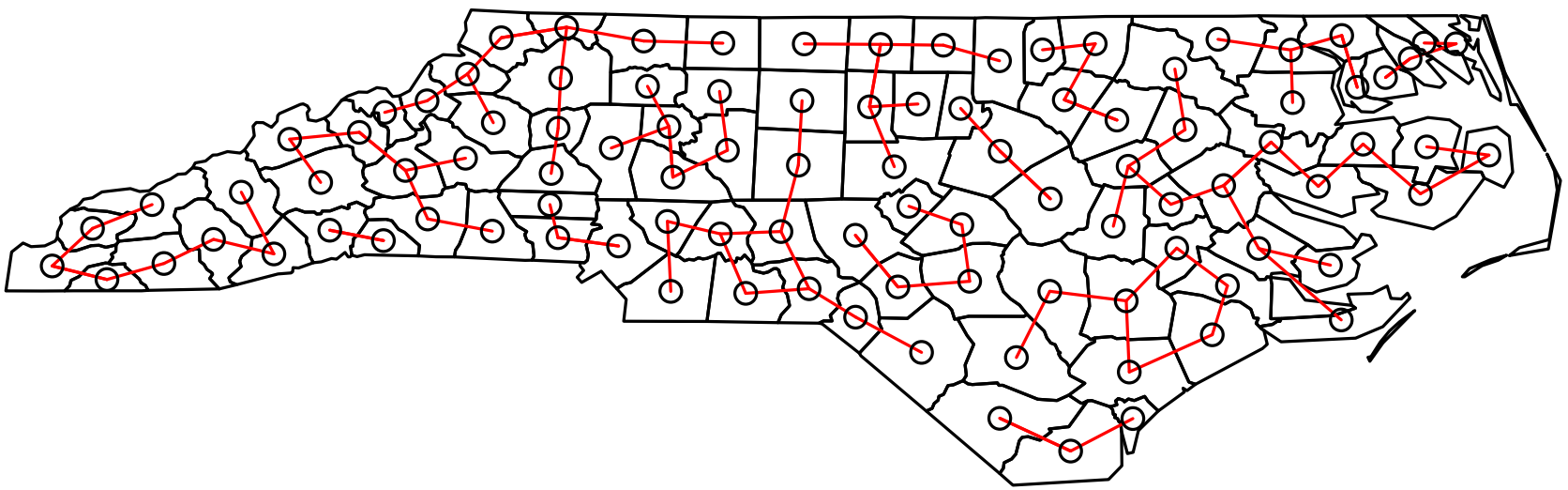
kNN Proximity: 1 Nearest Neighbor (1NN)



kNN Proximity: 2 Nearest Neighbors (2NN)



kNN Proximity: Difference Between 1NN and 2NN



Distance-Based Proximity

Distance Threshold Definition

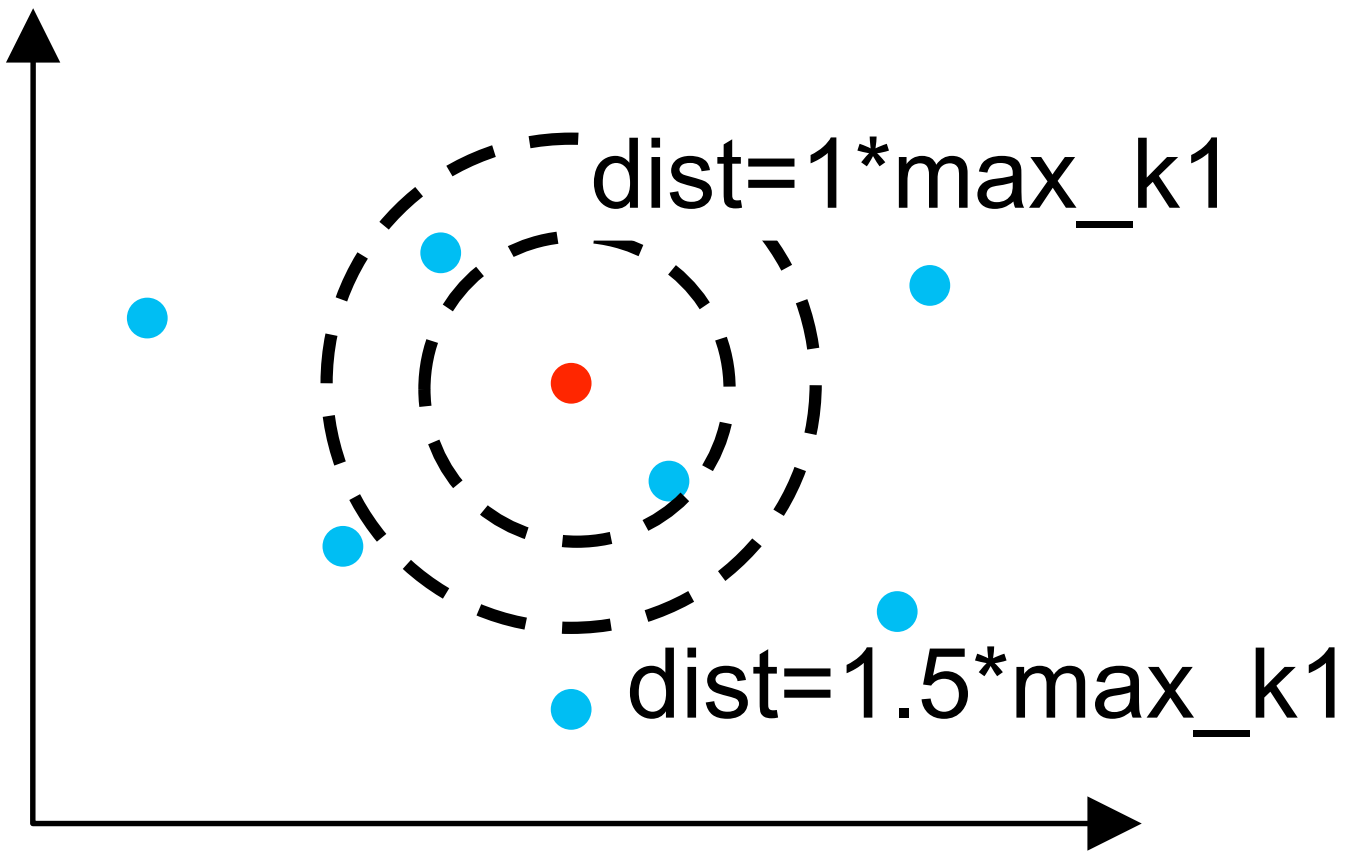
$$w_{ij} = \begin{cases} 1, & \text{if } d_{ij} < \varepsilon, \\ 0, & \text{otherwise.} \end{cases}$$

Alternatively, use distance decay:

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

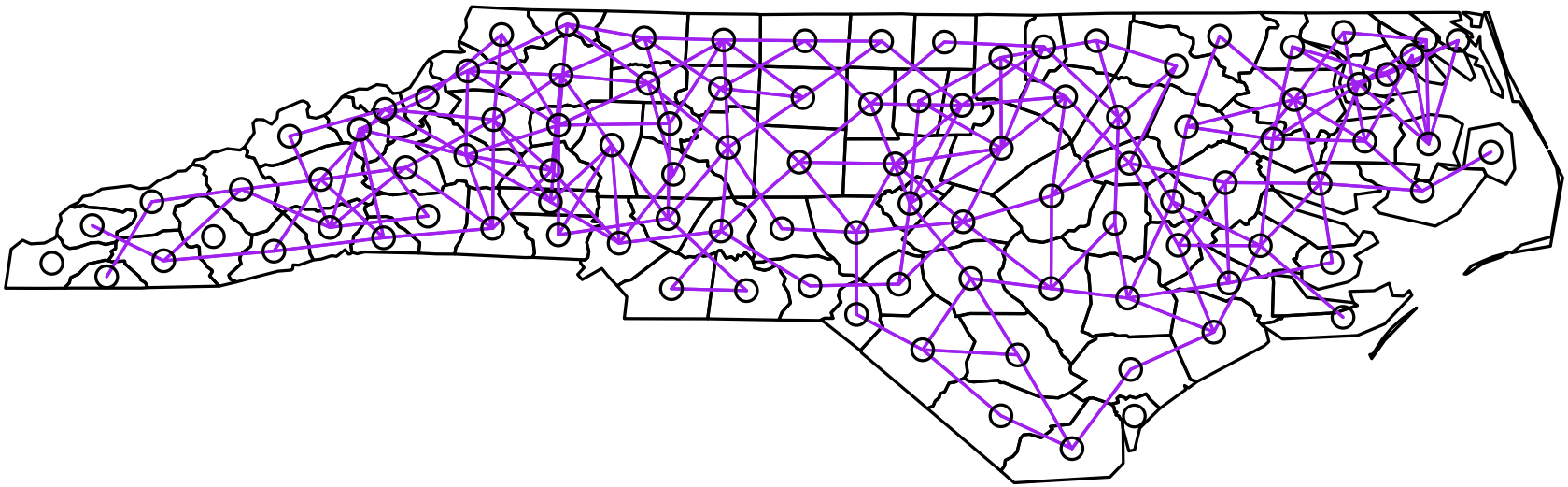
For some specified distance threshold ε or power ρ (recall inverse distance weighting).

Distance-Based Neighbors (ε)



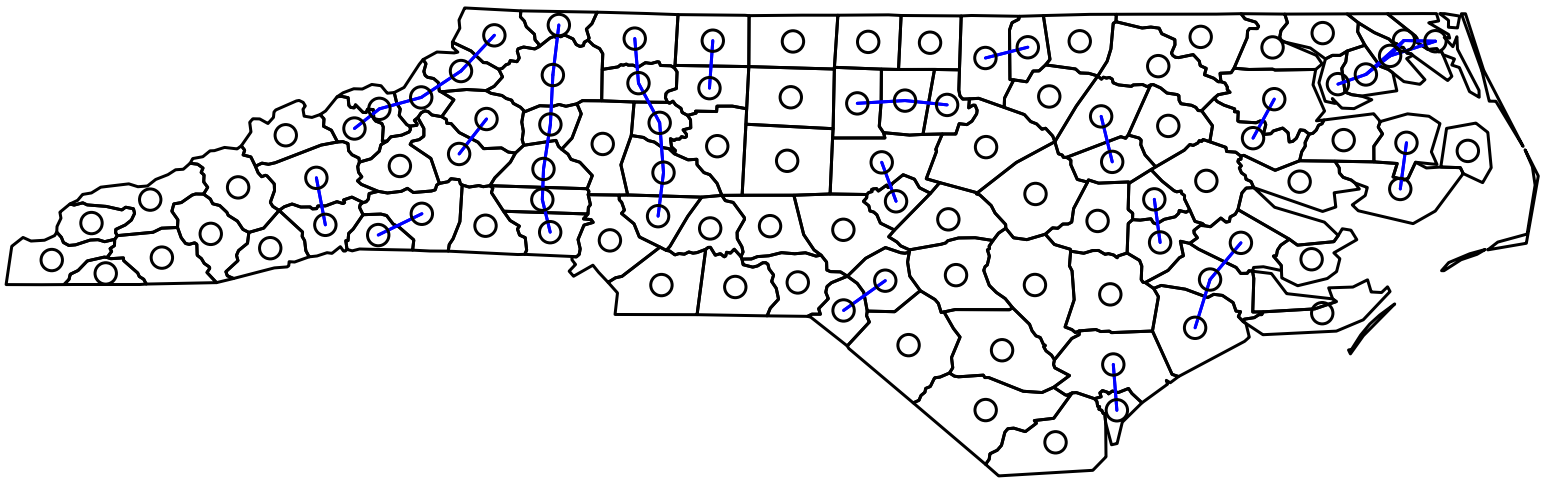
Distance-Based Neighbors

ε between 1 and 1.5 $\times$ maximum kNN distance



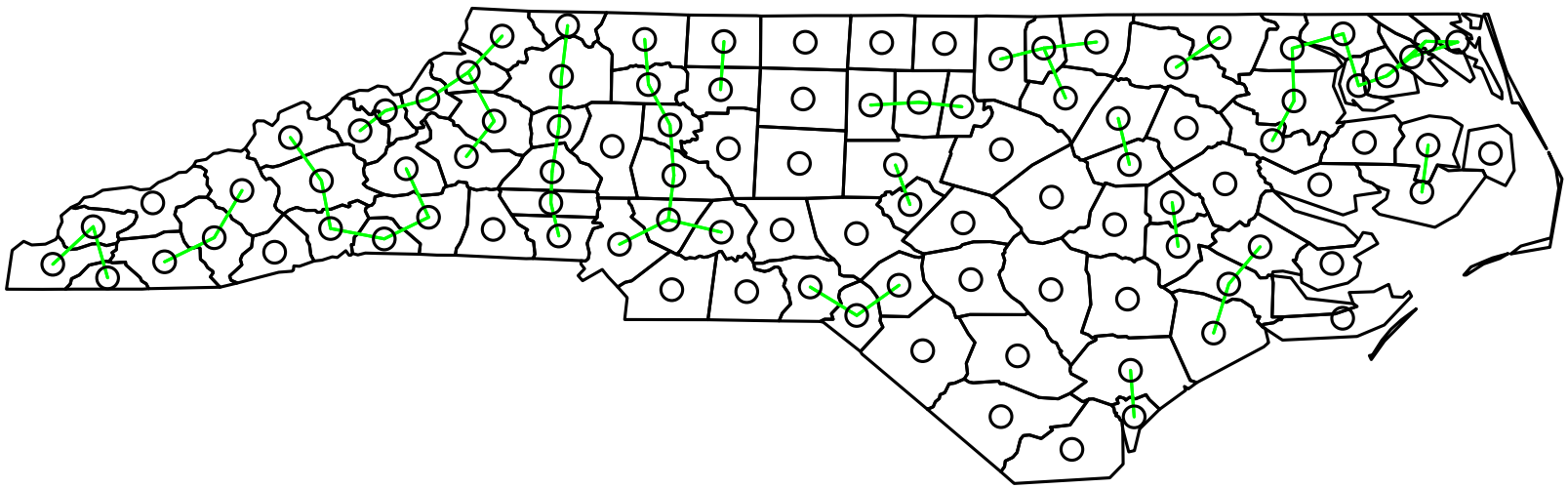
Distance-Based Neighbors

$\varepsilon = 30$ km (connected if $\leq$ 30 km apart)



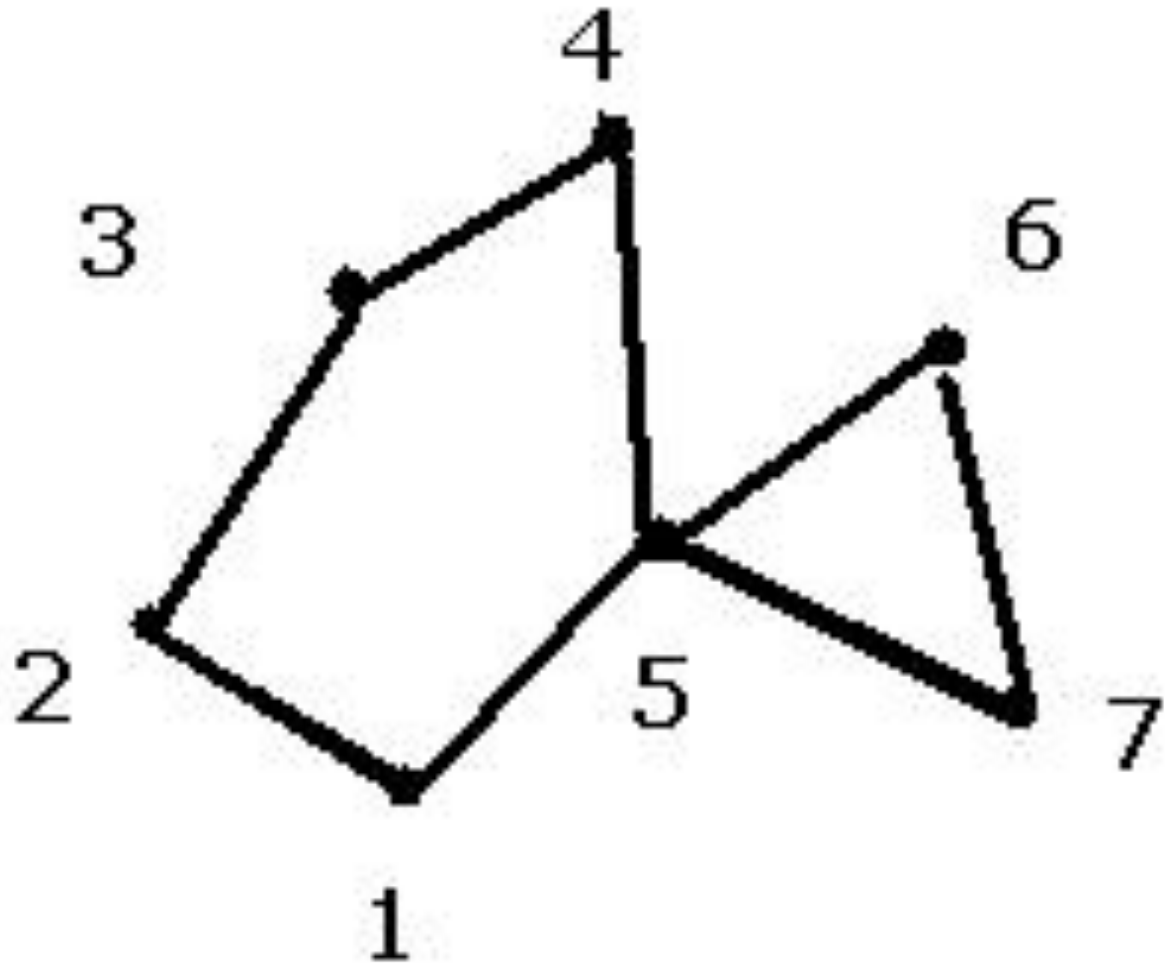
Distance-Based Neighbors

ε between 10 and 30 km



Adjacency Matrices

From Connectivity Graphs to Matrices



Example Neighbor List

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

Weights and the Adjacency Matrix

- The adjacency matrix W contains elements w_{ij} representing neighbor relationships.
- Once the neighbor list (fixed distance or kNN) is created, we assign spatial weights.
- Weights can be binary or variable.
- Even with binary weights (0/1), handling "no-neighbor" observations can be an issue.
- Binary weighting assigns 1 to neighboring features and 0 to all others.

Weights and the Adjacency Matrix: Binary Weights

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

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

Weights and the Adjacency Matrix: Row Standardization

- Row-standardized weights create proportional weights for unequal neighbor counts.
- They increase the influence of links from observations with few neighbors.
- Each neighbor weight is divided by the sum of all neighbor weights for that observation.
- Example: obs i has 3 neighbors $\rightarrow$ each gets weight $1/3$; obs j has 2 $\rightarrow$ each $1/2$.
- Use when comparing spatial parameters across datasets 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

Weight Matrix: Binary Weights

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

Weight Matrix: Row Standardized

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

Areal Spatial Smoothers

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

- Suppose we observe Y_i (e.g., the SIDS rate) for county i. We can form a smoothed estimate from its neighbors j:

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

- The new value $\hat{Y}_i$ is a function of its spatial neighbors j.
- This "borrows strength" from nearby areal units, making values look more like their neighbors.

Measures of Spatial Similarity

- Goal: summarize similarity between nearby areal units.
- **Spatial autocorrelation**: correlation of the same measurement across different areal units.
- The similarity of values at locations B_i and B_j is weighted by their proximity.
- The weights w_{ij} define proximity.

Measures of Spatial Association

Measuring strength of association

- We want to quantify how strongly nearby units are more (or less) alike than distant units.
- We also want to assess whether the observed similarity (or dissimilarity) is unlikely due to chance.
- Let Y_i be the response at the ith areal unit B_i and Y_j at B_j.
- Let sim_{ij} measure similarity (or dissimilarity) between B_i and B_j.
- Let w_{ij} measure spatial proximity between B_i and B_j.
- A general global statistic takes the form of a weighted cross–product:

$$C = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \, \text{sim}_{ij}.$$

Measures of Spatial Association

Example (setup):

$Y =$

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

$$\text{sim}_{ij} = (Y_i - Y_j)^2, \qquad w_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ share a boundary,} \\ 0, & \text{otherwise.} \end{cases}$$

Measures of Spatial Association

Example (cont'd): Weight matrix 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

Compute pairwise sim_{ij} from Y, then the global measure:

$$C = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \, \text{sim}_{ij}.$$

Measures of Spatial Association

Example (cont'd)

- If C is small: neighbors are very similar $\Rightarrow$ positive spatial autocorrelation.
- If C is large: neighbors are dissimilar.

Measuring strength of association

- A normalized global measure uses a weighted average:

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\,\mathrm{sim}_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}.$$

Global Indexes of Spatial Autocorrelation

- Summarize the degree to which similar observations tend to occur near each other.
- Global indexes aggregate over the whole study area (clustering signal), not individual clusters.
- Common structure: compute similarity at i,j and weight by proximity w_{ij}.
- High similarity with high weight $\Rightarrow$ close and alike; low similarity with high weight $\Rightarrow$ close and different.

Global Indexes of Spatial Autocorrelation

General Weighted-Average Form

- Weights w_{ij} define proximity.
- A generic index takes the form:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,\mathrm{sim}_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}.$$

Global Indexes of Spatial Autocorrelation

Moran's I

- Moran (1950) defines similarity via deviations from the mean:

$$\text{sim}_{ij} = (y_i - \bar{y})(y_j - \bar{y}), \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- With sample variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, a common form is:

$$I = \frac{1}{s^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}.$$

Moran's I: Properties

- I is a random variable determined by the joint distribution of the y_i and the spatial weights.
- Clustered (similar neighbors) $\Rightarrow I > 0$.
- Regular/alternating (dissimilar neighbors) $\Rightarrow I < 0$.
- Under spatial independence (randomization),

$$\mathbb{E}[I] = -\frac{1}{n-1} \quad \text{and} \quad \mathbb{E}[I] \to 0 \text{ as } n \to \infty.$$

- Asymptotically,

$$\frac{I + \frac{1}{n-1}}{\sqrt{\text{Var}(I)}} \overset{approx}{\sim} \mathcal{N}(0,1).$$

Moran's I: Testing and Software

- **Null hypothesis**: no spatial association (e.g., y_i i.i.d.).
- Test statistic:

$$z = \frac{I - \mathbb{E}[I]}{\sqrt{\mathrm{Var}(I)}}, \quad \mathbb{E}[I] = -\frac{1}{n-1}.$$

- Compare z to the standard normal distribution.
- In `R`: `spdep::moran.test`, which reports the standardized deviate (approximately $\mathcal{N}(0,1)$) and a p-value.

Global Indexes of Spatial Autocorrelation

Interpretation of Moran's I

- $I > 0$: Positive spatial autocorrelation (similar values cluster).
- $I < 0$: Negative spatial autocorrelation (neighbors are dissimilar).
- $I = 0$: No spatial autocorrelation (random spatial arrangement).

Moran's I is often between -1 and 1, but exact bounds depend on the data and the spatial weight structure.

Global Indexes of Spatial Autocorrelation

Geary's c

- Geary (1954) devised the *contiguity ratio* (Geary's c), using dissimilarity:

$$\text{sim}_{ij} = (y_i - y_j)^2.$$

- A common form is:

$$c = \frac{n-1}{2\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}.$$

- Like Moran's I, it is a weighted average, scaled by overall variation about the mean $\bar{y}$.

Geary's c: Properties and Software

- Range: $c \in [0, 2]$; $c \approx 0 \Rightarrow$ strong positive autocorrelation; $c \approx 2 \Rightarrow$ strong negative autocorrelation.
- c is not a Pearson correlation (it is related to the Durbin–Watson statistic).
- Low c indicates positive autocorrelation; high c indicates negative autocorrelation.
- Expected value under spatial independence: $\mathbb{E}[c] = 1$ (so $c = 1$ indicates no spatial autocorrelation).
- In `R`: `spdep::geary.test`.

Moran's I vs Geary's c

Moran's I

- Based on cross-product of deviations:

$$(y_i - \bar{y})(y_j - \bar{y})$$

- Measures *global* similarity (covariance-like).
- Sensitive to overall spatial clustering.
- Values roughly in $[-1, 1]$ (not strictly bounded).
- Easier to interpret as a correlation analogue.
- More influenced by global trends or gradients.

Geary's c

- Based on squared differences:

$$(y_i - y_j)^2$$

- Measures *local* dissimilarity (difference-like).
- More sensitive to local variation.
- Ranges from 0 (strong positive autocorrelation) to 2 (strong negative autocorrelation).
- $c = 1$ indicates no spatial autocorrelation.

Use cases: Moran's I for detecting overall clustering; Geary's c for identifying sharp local contrasts or boundaries.