Basics of Geographic Analysis in R

Spatial Autocorrelation and Spatial Weights

Yuri M. Zhukov

GOV 2525: Political Geography

February 25, 2013

Outline

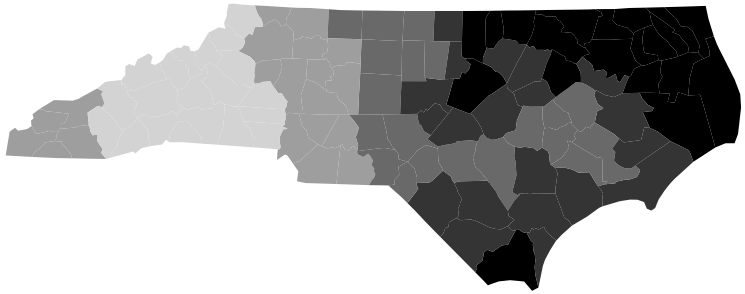
1. Introduction
2. Spatial Data and Basic Visualization in R
3. Spatial Autocorrelation
4. Spatial Weights
5. Spatial Regression

What is Spatial Autocorrelation?

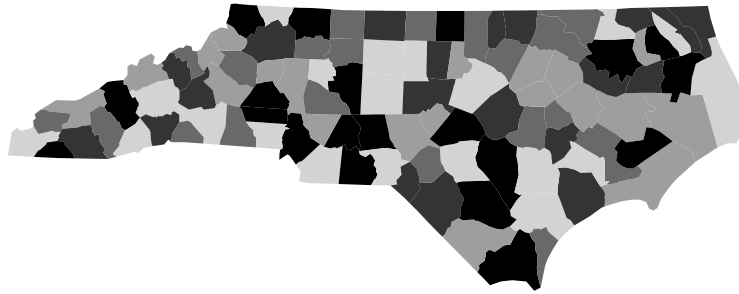
- Spatial autocorrelation measures the degree to which a phenomenon of interest is correlated to itself in space.
- Tests of spatial autocorrelation examine whether the observed value of a variable at one location is independent of values of that variable at neighboring locations.
- Positive spatial autocorrelation indicates that similar values appear close to each other, or cluster, in space
- Negative spatial autocorrelation indicates that neighboring values are dissimilar or, equivalenty, that similar values are dispersed.
- Null spatial autocorrelation indicates that the spatial pattern is random.

What is Spatial Autocorrelation?

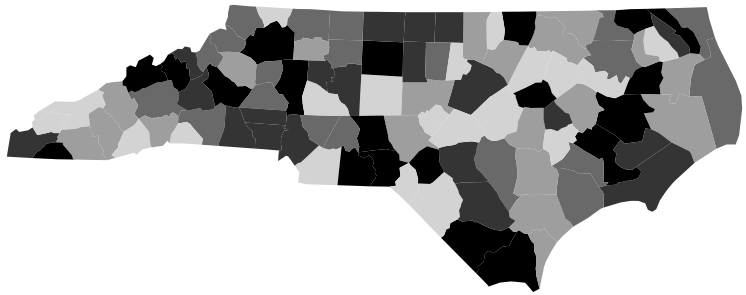
Positive autocorrelation



Negative autocorrelation



No autocorrelation



Global autocorrelation: Moran's $\mathcal{I}$

- The Moran's $\mathcal{I}$ coefficient calculates the ratio between the product of the variable of interest and its spatial lag, with the product of the variable of interest, adjusted for the spatial weights used.

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

- where y_i is the value of a variable for the ith observation, $\bar{y}$ is the sample mean and w_{ij} is the spatial weight of the connection between i and j.
- Values range from –1 (perfect dispersion) to +1 (perfect correlation). A zero value indicates a random spatial pattern.
- Under the null hypothesis of no autocorrelation, $\mathbb{E}[\mathcal{I}] = \frac{-1}{n-1}$

Global autocorrelation: Moran's $\mathcal{I}$

- Calculating the variance of Moran's $\mathcal{I}$ is a little more involved:

$$
\begin{aligned}
Var(\mathcal{I}) =& \frac{n\mathfrak{s}_1 - \mathfrak{s}_2\mathfrak{s}_3}{(n-1)(n-2)(n-3)(\sum_i \sum_j w_{ij})^2} \\
\mathfrak{s}_1 =& (n^2 - 3n + 3)\left(\frac{1}{2}\sum_i\sum_j (w_{ij} + w_{ji})^2\right) \\
& - n\left(\sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2\right) + 3(\sum_i\sum_j w_{ij})^2 \\
\mathfrak{s}_2 =& \frac{n^{-1}\sum_i (y_i - \bar{x})^4}{(n^{-1}\sum_i (y_i - \bar{x})^2)^2} \\
\mathfrak{s}_3 =& \frac{1}{2}\sum_i\sum_j (w_{ij} + w_{ji})^2 - 2n\left(\frac{1}{2}\sum_i\sum_j (w_{ij} + w_{ji})^2\right) \\
& + 6\left(\sum_i\sum_j w_{ij}\right)^2
\end{aligned}
$$

Global autocorrelation: Geary's $\mathcal{C}$

- The Geary's $\mathcal{C}$ uses the sum of squared differences between pairs of data values as its measure of covariation.

$$\mathcal{C} = \frac{(n-1)\sum_i \sum_j w_{ij}(y_i - y_j)^2}{2(\sum_i \sum_j w_{ij}) \sum_i (y_i - \bar{y})^2}$$

- where y_i is the value of a variable for the ith observation, $\bar{y}$ is the sample mean and w_{ij} is the spatial weight of the connection between i and j.
- Values range from 0 (perfect correlation) to 2 (perfect dispersion). A value of 1 indicates a random spatial pattern.

Global autocorrelation: Join Counts

- When the variable of interest is *categorical*, a join count analysis can be used to assess the degree of clustering or dispersion.
- A binary variable is mapped in two colors (Black & White), such that a join, or edge, is classified as either *WW* (0-0), *BB* (1-1), or *BW* (1-0).
- Join count statistics can show
  - positive spatial autocorrelation (clustering) if the number of *BW* joins is significantly *lower* than what we would expect by chance,
  - negative spatial autocorrelation (dispersion) if the number of *BW* joins is significantly *higher* than what we would expect by chance,
  - null spatial autocorrelation (random pattern) if the number of *BW* joins is approximately *the same* as what we would expect by chance.

Global autocorrelation: Join Counts

- By the naive definition of probability, if we have n_B Black units and $n_W = n - n_B$ White units, the respective probabilities of observing the two types of units are:

$$P_B = \frac{n_B}{n} \qquad P_W = \frac{n - n_B}{n} = 1 - P_B$$

- The probabilities of BB and WW in two adjacent cells are

$$P_{BB} = P_B P_B = P_B^2 \qquad P_{WW} = (1 - P_B)(1 - P_B) = (1 - P_B)^2$$

- The probability of BW in two adjacent cells is

$$P_{BW} = P_B(1 - P_B) + (1 - P_B)P_B = 2P_B(1 - P_B)$$

Global autocorrelation: Join Counts

- The expected counts of each type of join are:

$$\mathbb{E}[BB] = \frac{1}{2}\sum_i \sum_j w_{ij} P_B^2 \qquad \mathbb{E}[WW] = \frac{1}{2}\sum_i \sum_j w_{ij}(1 - P_B)^2$$

$$\mathbb{E}[BW] = \frac{1}{2}\sum_i \sum_j w_{ij} 2P_B(1 - P_B)$$

- Where $\frac{1}{2}\sum_i \sum_j w_{ij}$ is the total number of joins (of any type) on a map, assuming a binary connectivity matrix.
- The observed counts are:

$$BB = \frac{1}{2}\sum_i \sum_j w_{ij} y_i y_j \qquad WW = \frac{1}{2}\sum_i \sum_j w_{ij}(1 - y_i)(1 - y_j)$$

$$BW = \frac{1}{2}\sum_i \sum_j w_{ij}(y_i - y_j)^2$$

- where $y_i = 1$ if unit i is Black and $y_i = 0$ if White.

Global autocorrelation: Join Counts

- The variance of BW is calculated as

$$\begin{aligned}
\sigma^2_{BW} =& \mathbb{E}[BW^2] - \mathbb{E}[BW]^2 \\
=& \frac{1}{4}\left(\frac{2\mathfrak{s}_2 n_B(n-n_B)}{n(n-1)} + \frac{(\mathfrak{s}_3 - \mathfrak{s}_1) n_B(n-n_B)}{n(n-1)} \right. \\
& \left. + \frac{4(\mathfrak{s}_1^2 + \mathfrak{s}_2 - \mathfrak{s}_3) n_B(n_B-1)(n-n_B)(n-n_B-1)}{n(n-1)(n-2)(n-3)} \right) - \mathbb{E}[BW]^2
\end{aligned}$$

$$\mathfrak{s}_1 = \sum_i \sum_j w_{ij}$$

$$\mathfrak{s}_2 = \frac{1}{2} \sum_i \sum_j (w_{ij} - w_{ji})^2$$

$$\mathfrak{s}_3 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2$$

Global autocorrelation: Join Counts

- A test statistic for the BW join count is

$$\mathcal{Z}(BW) = \frac{BW - \mathbb{E}[BW]}{\sqrt{\sigma^2_{BW}}}$$

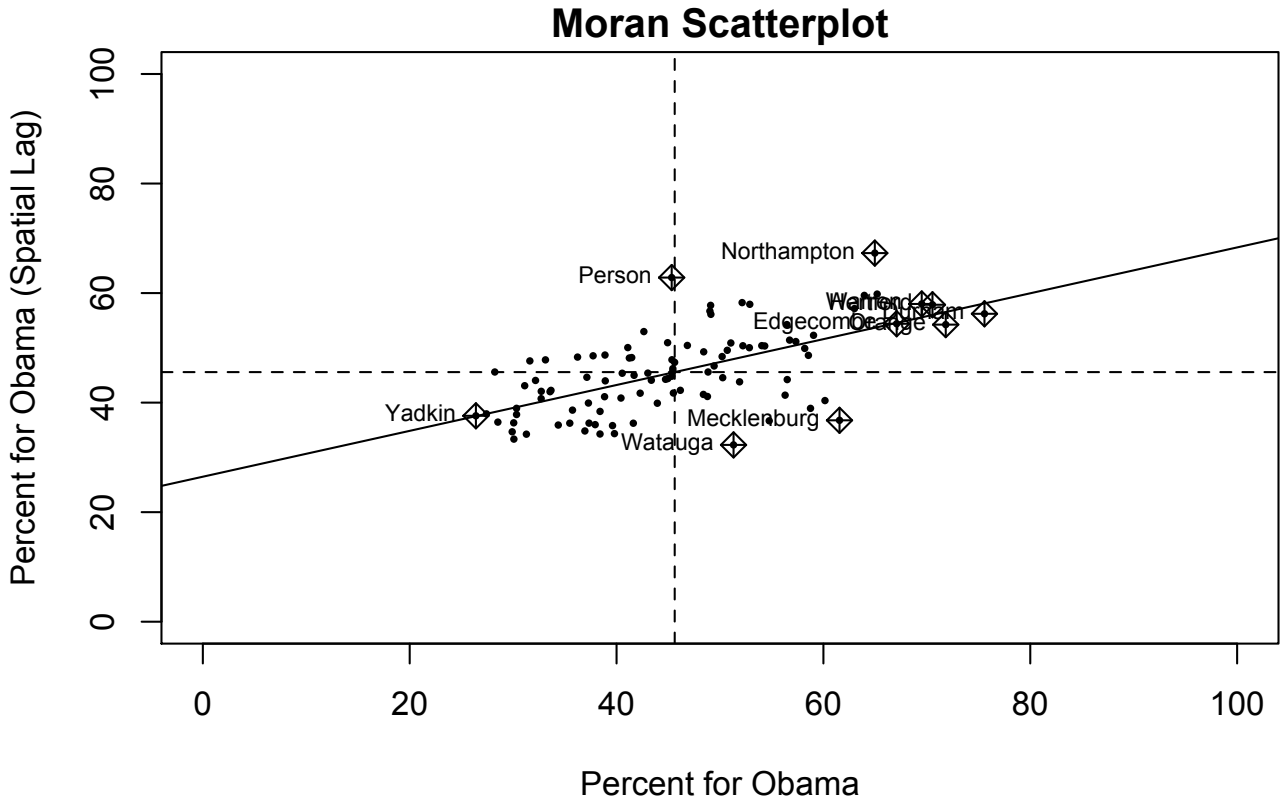
- The join count statistic is assumed to be asymptotically normally distributed under the null hypothesis of no spatial autocorrelation.
- The test of significance is then provided by evaluating the BW statistic as a standard deviate (Cliff and Ord, 1981).

Local autocorrelation

- Global tests for spatial autocorrelation are calculated from local relationships between observed values at spatial units and their neighbors.
- It is possible to break these measures down into their components, thus constructing local tests for spatial autocorrelation.
- These tests can be used to detect
  - Clusters, or units with similar neighbors
  - Enclaves, or units with dissimilar neighbors

Local autocorrelation

Below is a scatterplot of county vote for Obama and its spatial lag (average vote received in neighboring counties). The Moran's $\mathcal{I}$ coefficient is drawn as the slope of the linear relationship between the two. The plot is partitioned into four quadrants: low-low, low-high, high-low and high-high.



Local autocorrelation: Local Moran's $\mathcal{I}$

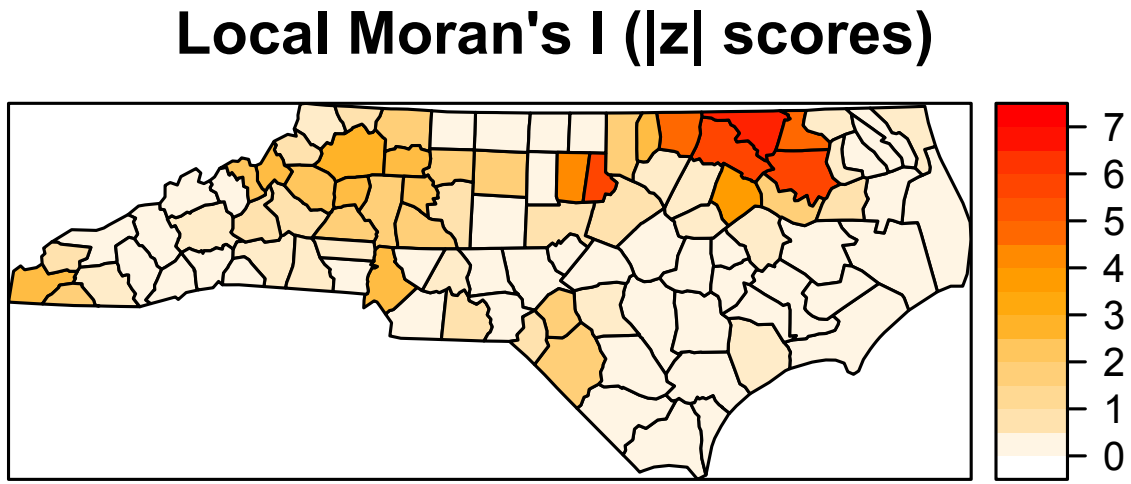
- A local Moran's $\mathcal{I}$ coefficient for unit i can be constructed as one of the n components which comprise the global test:

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

- As with global statistics, we assume that the global mean $\bar{y}$ is an adequate representation of the variable of interest.
- As before, local statistics can be tested for divergence from expected values, under assumptions of normality.

Local autocorrelation: Local Moran's $\mathcal{I}$

Below is a plot of Local Moran $|z|$-scores for the 2008 Presidential Elections. Higher absolute values of z scores (red) indicate the presence of "enclaves", where the percentage of the vote received by Obama was significantly different from that in neighboring counties.



Words of Caution

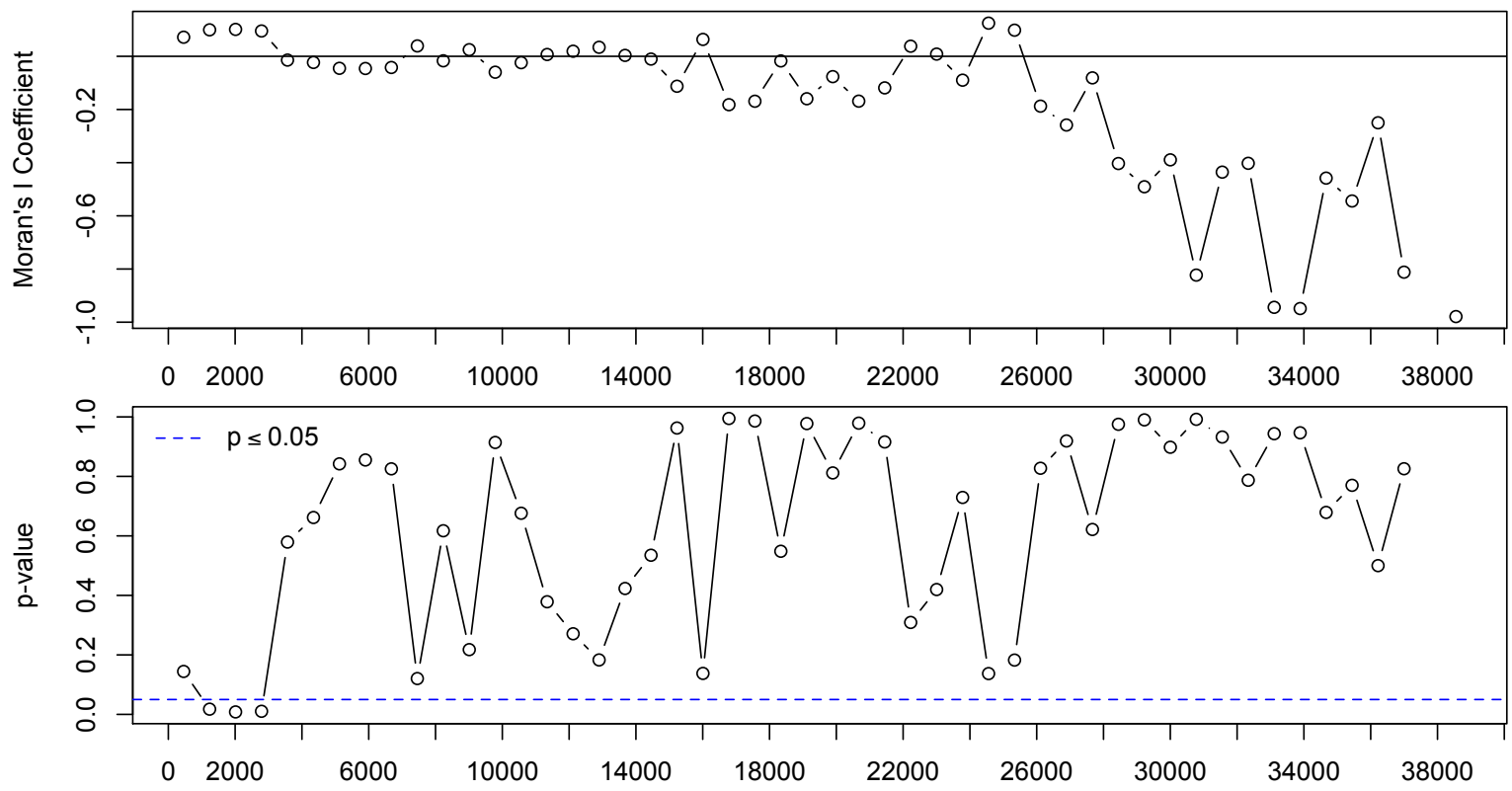
1. By themselves, spatial autocorrelation tests do not always produce useful insights into the DGP.

Words of Caution

1. By themselves, spatial autocorrelation tests do not always produce useful insights into the DGP.
2. These tests are also highly sensitive to one's choice of spatial weights. Where the weights do not reflect the "true" structure of spatial interaction, estimated autocorrelation (or lack thereof) may actually stem from misspecification.

Words of Caution

Below is a correlogram of Moran's $\mathcal{I}$ coefficients for Polity IV country democracy scores in 2008. The x-axis represents distances between country capitals, in kilometers. Here, democracy is significantly ($p \leq .05$) spatially autocorrelated only at distances of 3,000 km and below. So, autocorrelation estimates will depend highly on choice of lag distance.



Words of Caution

1. By themselves, spatial autocorrelation tests do not always produce useful insights into the DGP.
2. These tests are also highly sensitive to one's choice of spatial weights. Where the weights do not reflect the "true" structure of spatial interaction, estimated autocorrelation (or lack thereof) may actually stem from misspecification.
3. As originally designed, spatial autocorrelation tests assumed there are no neighborless units in the study area.

Outline

1. Introduction
2. Spatial Data and Basic Visualization in R
3. Spatial Autocorrelation
4. Spatial Weights
5. Spatial Regression

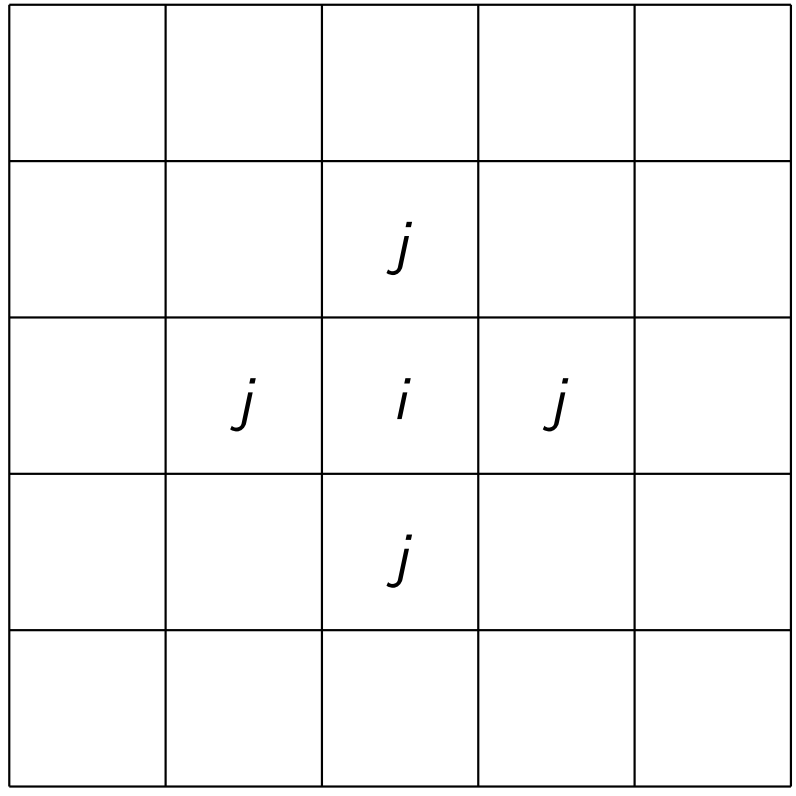
Choosing your neighbors?

- Most spatial weights matrices **W** are based on some version of a connectivity matrix **C**.
- **C** is an $n \times n$ binary matrix, where $i = \{1, 2, \ldots, n\}$ and $j = \{1, 2, \ldots, n\}$ are the units in the system (for example, countries in the international system).
- Entry $c_{ij} = 1$ if two units $i \neq j$ are considered connected, and $c_{ij} = 0$ if they are not.
- The tricky part is how the word "connected" is defined.

Areal Contiguity I: Regular Grids

Rook's case

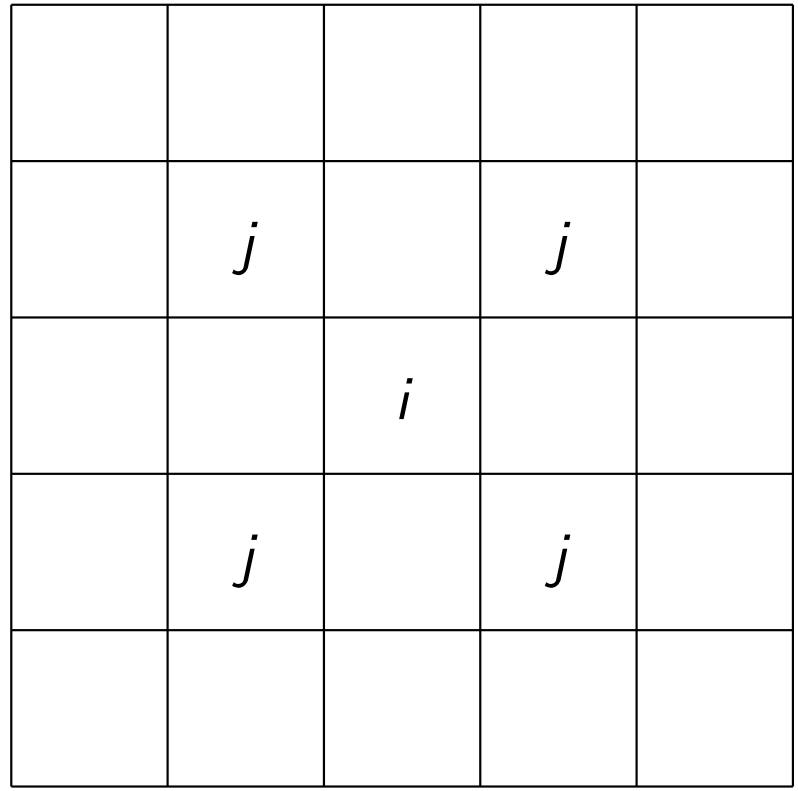
Cells sharing a common edge are considered contiguous



Areal Contiguity I: Regular Grids

Bishop's case

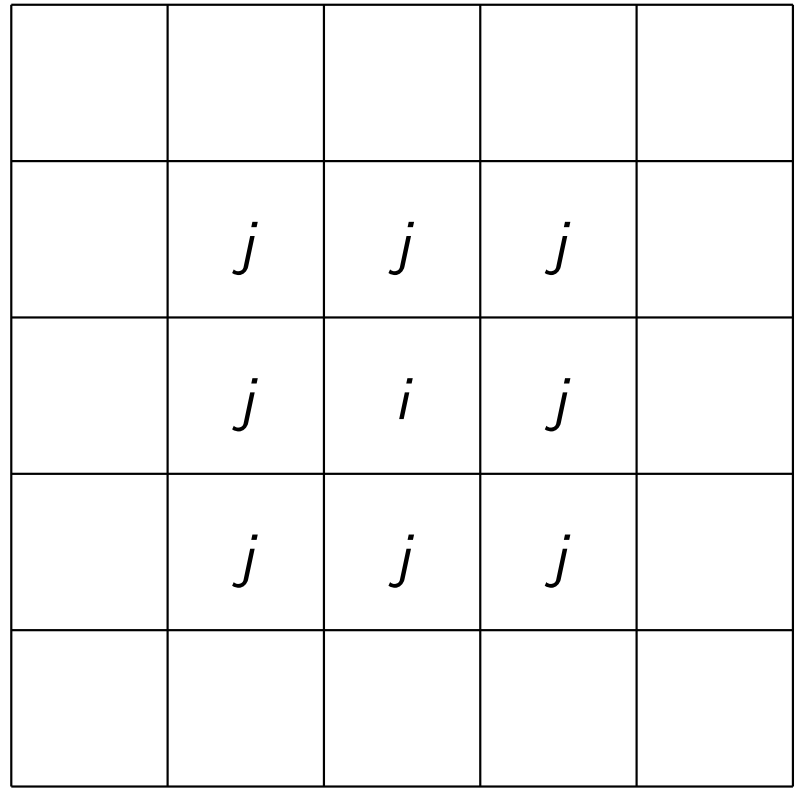
Cells sharing a common vertex are considered contiguous



Areal Contiguity I: Regular Grids

Queen's case

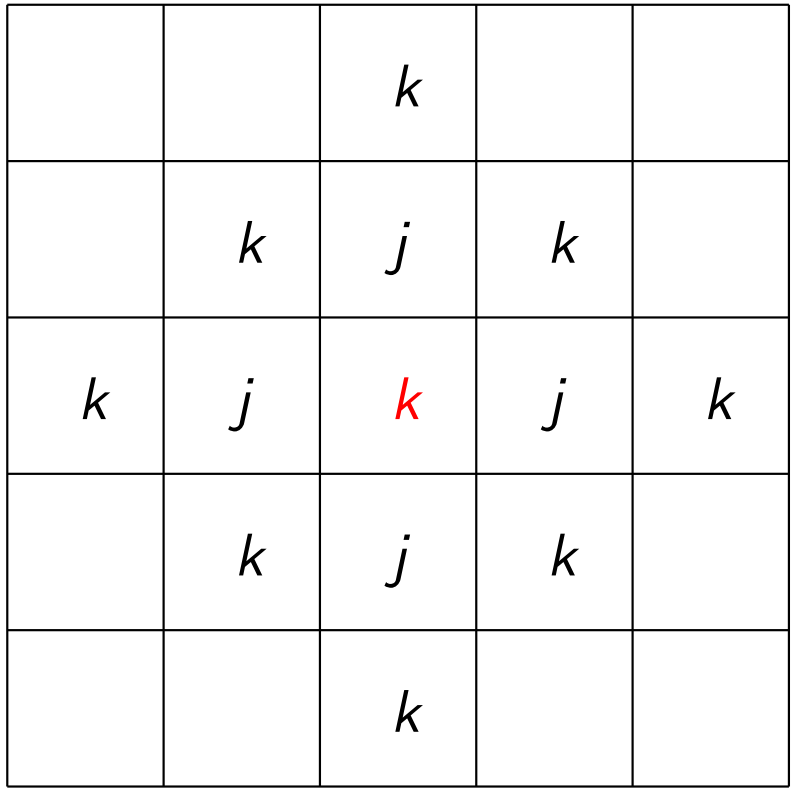
Cells sharing a common edge or common vertex are considered contiguous



Areal Contiguity I: Regular Grids

Second-order neighbors: (rook's case)

Cells sharing a common edge with first-order neighbors are considered contiguous

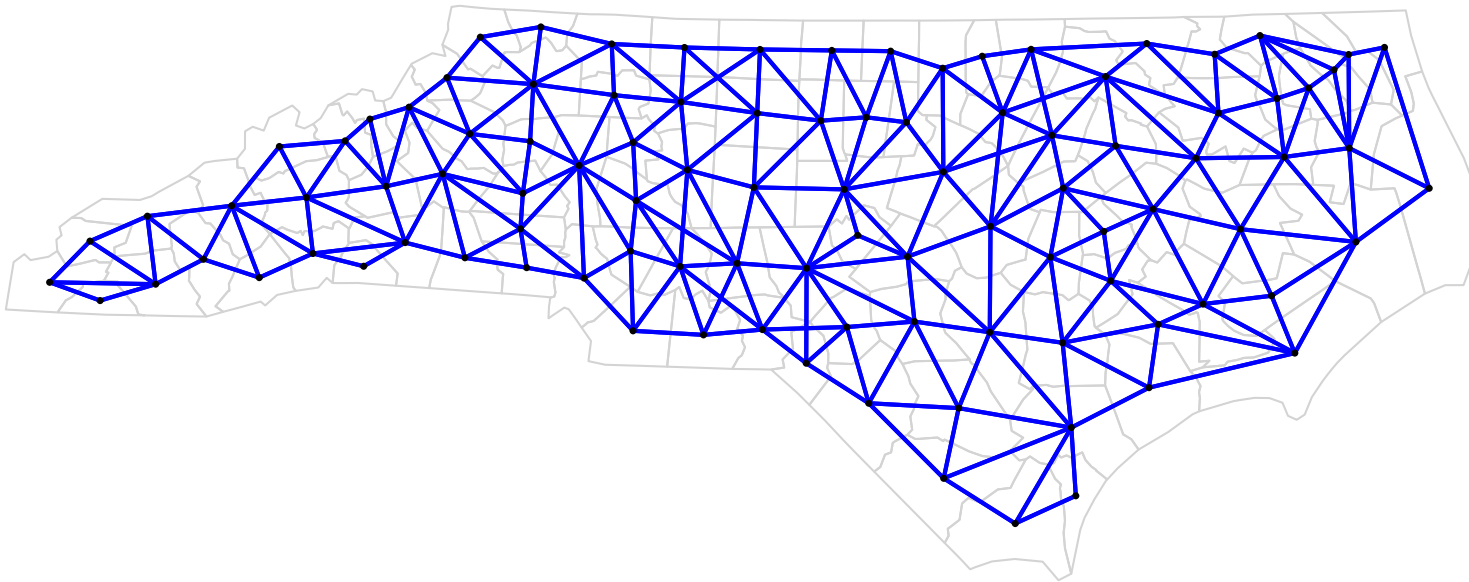


Areal Contiguity I: Regular Grids

- These conceptions of contiguity are useful when dealing with regular square grids or rectangular lattices, where the spatial structure can be easily summarized in elegant mathematical terms.
- But when spatial units consist of irregularly-shaped polygons, as is the case in most applied work (countries, census tracts, various administrative units), this simple characterization breaks down...

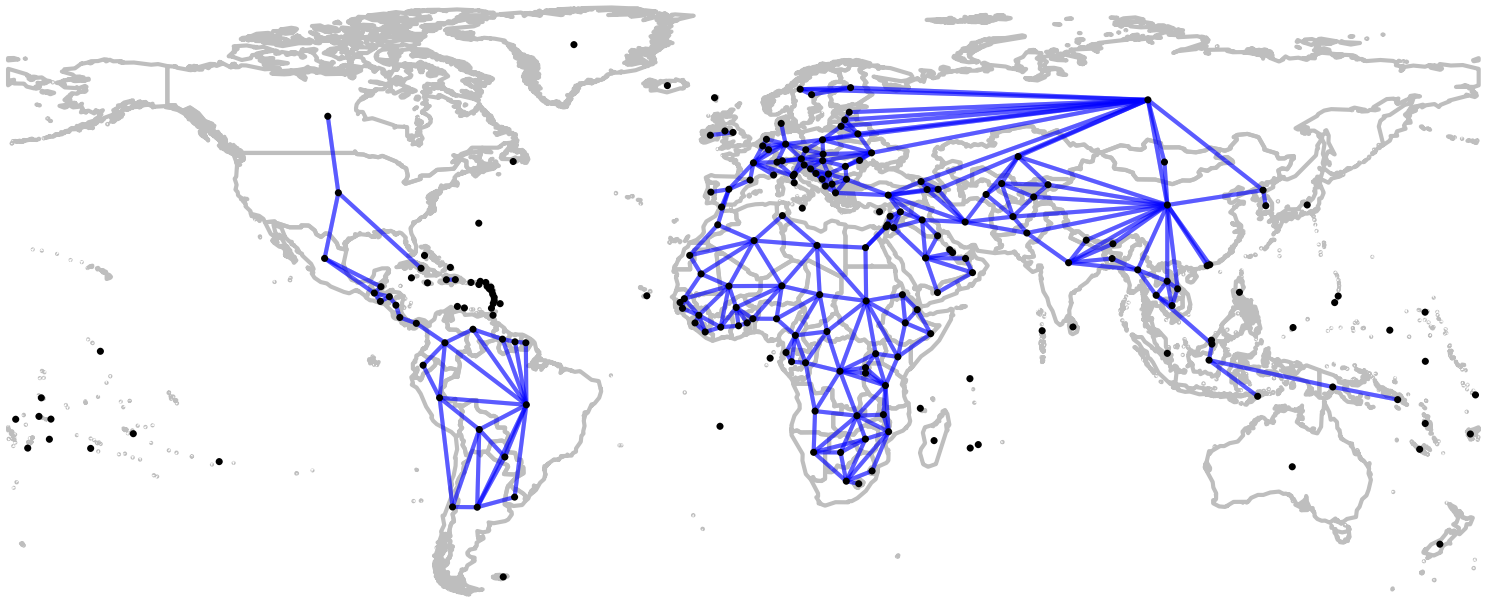
Areal Contiguity II: Polygons

Figure: Contiguity neighbors



Areal Contiguity II: Polygons

Figure: Contiguity neighbors



Interpoint Distance

Thresholding

$$\mathbf{c}_{THRES}(i,j) = 1\{i,j \in \mathbf{S} : d(i,j) \leq r\}$$

k nearest neighbor

$$\mathbf{c}_{KNN}(i,j) = 1\{i,j \in \mathbf{S} : d(i,j) \leq d_{(k)}(i,-)\}$$

Sphere of Influence

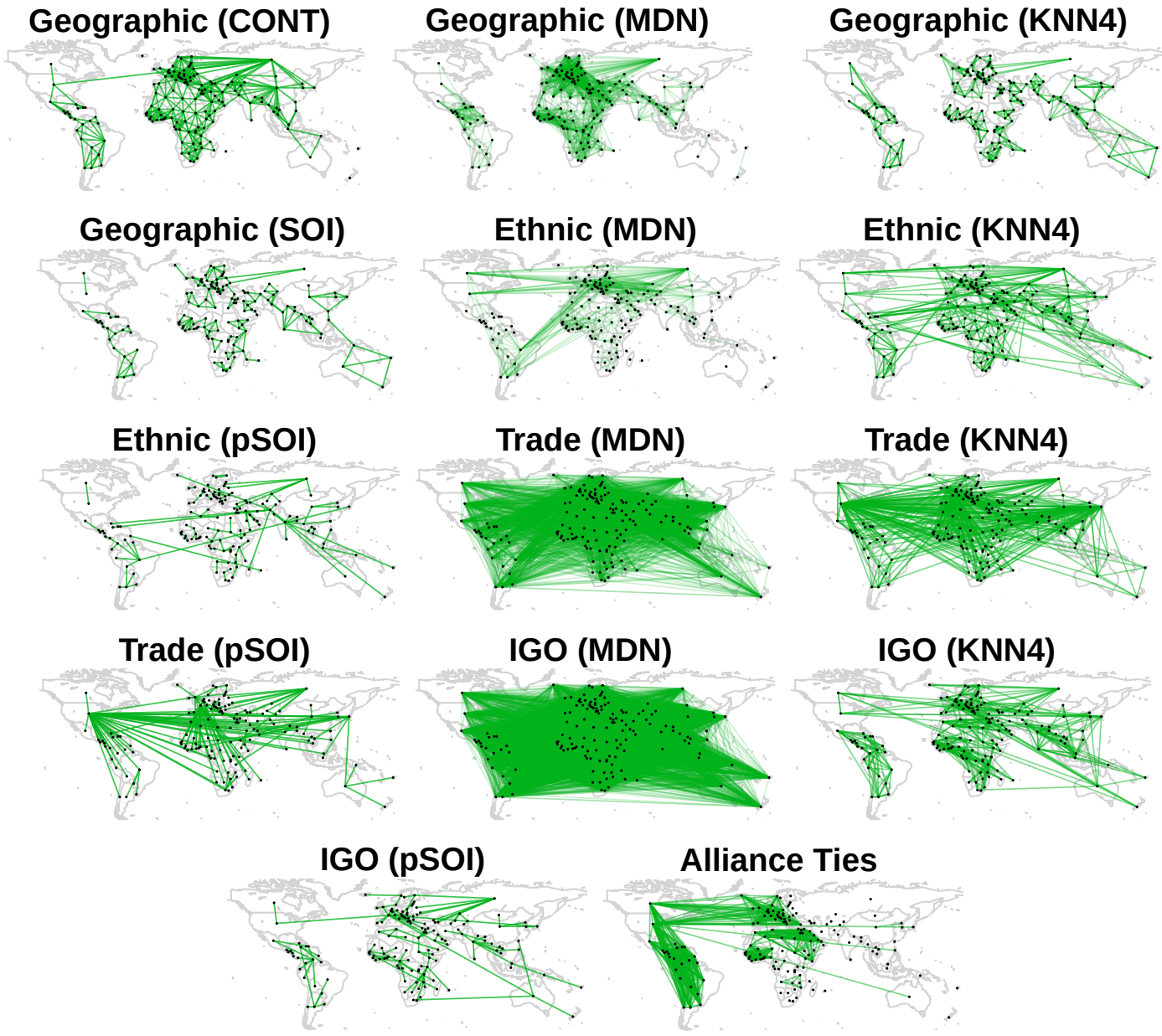
$$\mathbf{c}_{SOI}(i,j) = 1\{i,j \in \mathbf{S} : O_i \cap O_j \neq \emptyset\}$$

Network neighbors

The structure of spatial dependence can be non-geographic. Any theoretically-relevant dyadic relationship can form the basis of connectivity.

- **Individual level**: friendship, frequency of communication, citations, kinship.
- **Organizational level**: market competition, joint enterprises, personnel exchanges.
- **International level**: alliance relationship, trade flows, joint organizational membership, diplomatic contacts, cultural exchanges, migration flows.

Other options



From Connections to Weights

- Once a definition of connectivity is made, one must translate binary indicators into weights, which will form the elements w_{ij} of matrix **W**.
- A plethora of options exist: inverse distance (IDW), negative exponentials of distance, length of shared boundary, relative area, accessibility...
- The rows of **W** are often row-standardized, so that $\sum_{j=1}^{n} w_{ij} = 1$
- Row standardization facilitates interpretation of lagged variables as a weighted average of neighboring values.
- This also ensures that principal eigenvalue is 1 (useful for optimization in regression models).
- **Bottom line**: the weights should bear a direct relation to one's theoretical conceptualization of the structure of dependence.

Sparse vs. Dense Matrices

Sparsity carries a number of substantive and computational advantages:

- Dense matrices are noisy and contain a potentially large number of irrelevant connections.
- Dense matrices will bias downward indirect effects of a change in observation j (the individual weights of non-zero entries in row-standardized weight matrices will be smaller).
- Dense matrices can be computationally intensive to the point that even simple matrix operations are infeasible.

Sparse vs. Dense Matrices

Consider the following example with 2000 U.S. Census data:

Tracts

$n = 65,443$

31.90 GB of storage required for dense matrix, .01 GB for sparse matrix.

Block Groups

$n = 208,790$

324.80 GB of storage required for dense matrix, .03 GB for sparse matrix.

Blocks

$n = 8,205,582$

501,659.33 GB of storage required for dense matrix, 1.10 GB for sparse.

Here, dense and sparse matrices have n^2 and $6/n$ nonzero elements, respectively. For spatially random data on a plane, each unit will have an average of 6 contiguity neighbors.

Ordering of Weights Matrix

Ordering of rows and columns matters greatly for computation times.

- Consider an $n \times n$ permutation matrix $\mathbf{P}$, which has exactly one entry 1 in each row and each column and 0's elsewhere. Each permutation matrix can produce a reordered weights matrix $\mathbf{W}_P$, by the operation $\mathbf{W}_P = \mathbf{PWP}'$.
- Note that $\mathbf{P}^{-1} = \mathbf{P}'$, $|\mathbf{P}| = 1$ and $|\mathbf{P}(\mathbf{I}_n - \rho\mathbf{W})\mathbf{P}'| = |\mathbf{P}||\mathbf{I}_n - \rho\mathbf{W}||\mathbf{P}'| = |\mathbf{I}_n - \rho\mathbf{W}| = |\mathbf{I}_n - \rho\mathbf{PWP}'|$
- Thanks to these properties, log-determinant calculation and other matrix operations will not be affected by the reordering of $\mathbf{W}$.
- But computation times for these operations are affected.

Ordering of Weights Matrix

Efficiency is increased if ordering is **geographic** (north-south or east-west)

- This ordering concentrates nonzero elements around the diagonal, which reduces the bandwidth of a matrix ($max|i-j|$ for nonzero elements).
- For a sample of 62,226 U.S. Census Tracts, calculation of a single log-determinant requires over 12 GB of memory for a randomly ordered weights matrix, making calculation infeasible on most machines.
- The same operation takes less than a minute for a geographically-ordered matrix.

Examples in R

Switch to R tutorial script.