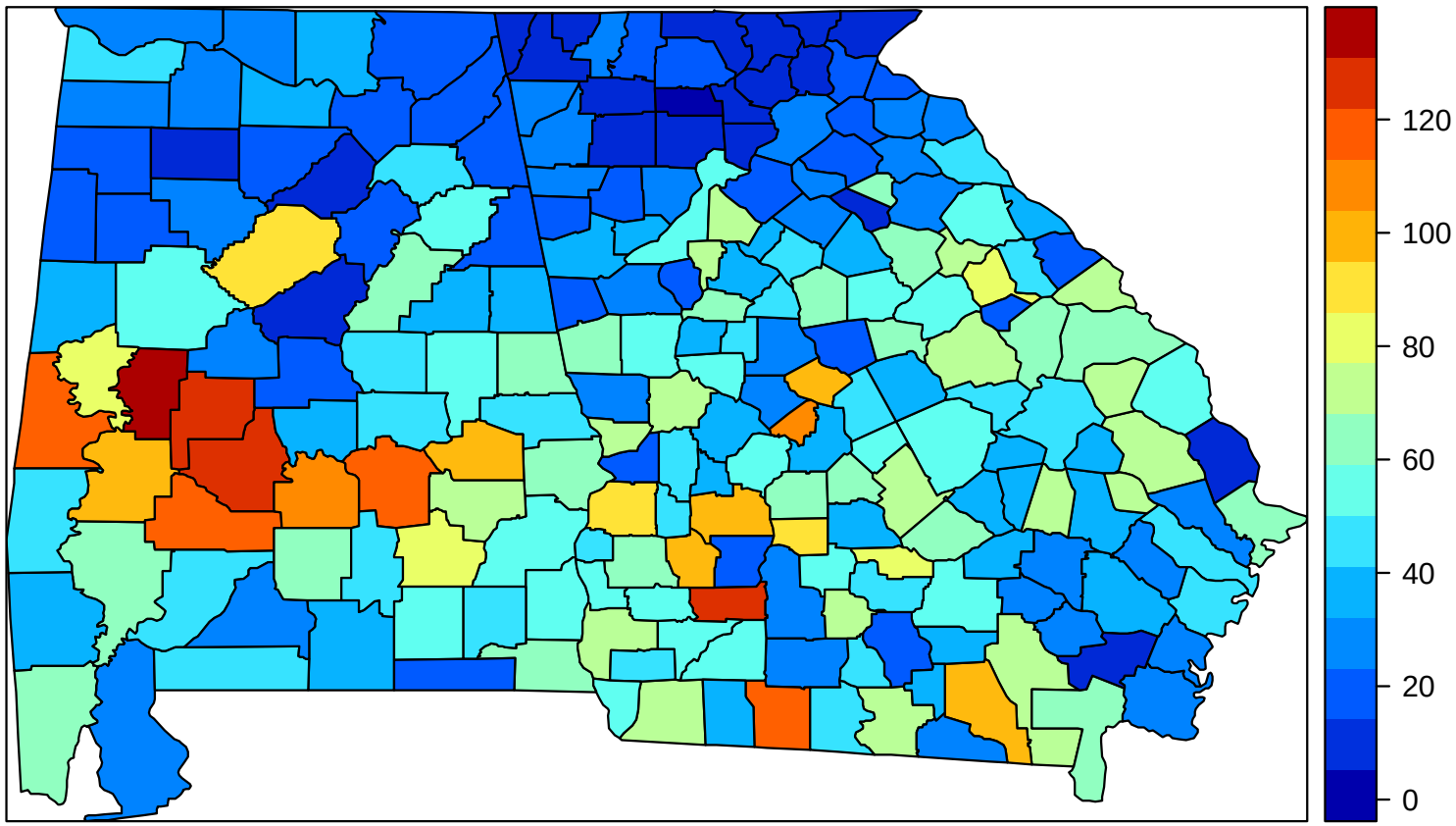
Lecture 3: Areal Data and Autocorrelation

SISMID 2025

Howard Chang
howard.chang@emory.edu

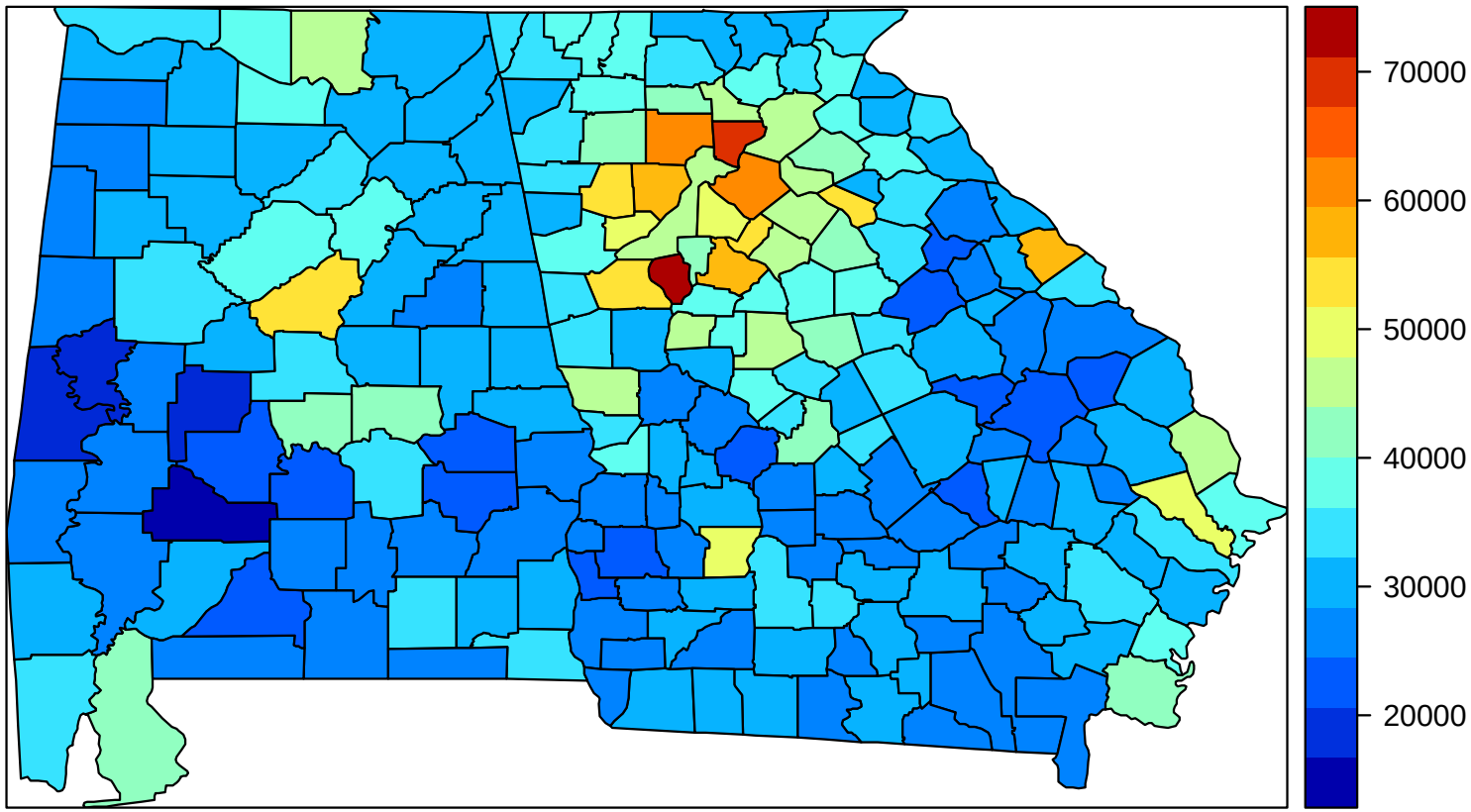
Motivating Data Example

County-level Incidence of Chlamydia Rate (per 10,000) in 2003



Motivating Data Example

County-level Median Household Income ($) in 2000



Spatial Analysis Questions

Areal data are typically represented as contiguous polygons with irregular or regular shapes (e.g grid).

Question 1: Is there spatial dependence? Do counties with high median income/chlamydia incidence tend to be closer to each other?

Question 2: What are some covariates that explain the observed spatial pattern? The covariates of interest often exhibit spatial pattern themselves.

Question 3: Can we identify areas that have values much higher or lower than expected? Perhaps these areas require further investigation or intervention.

Question 4: How do we smooth across polygons (leveraging spatial correlation) to account for random noise associated with each observation?

Proximity/Adjacency Matrix **W**

The spatial information of areal data cannot be represented by a single location. We need to describe how each observation is spatially connected to all other areal units.

Given observations $Y_1, \ldots, Y_n$, let **W** denote an $n \times n$ matrix, where element W_{ij} measures the spatial proximity between Y_i and Y_j.

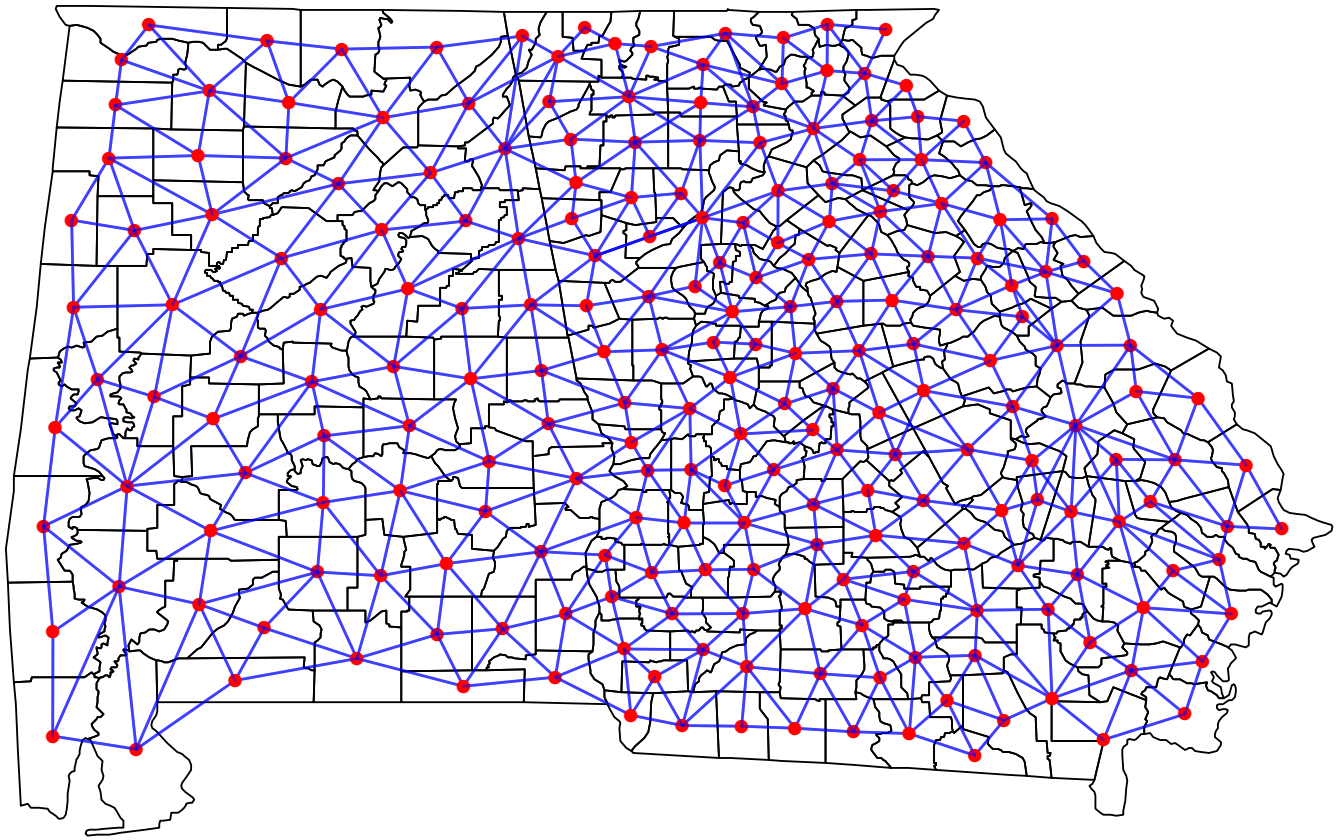
Some common choices include:

- $W_{ij} = 1$ if Y_i and Y_j share a border and $W_{ij} = 0$ otherwise.
- $W_{ij} = d_{ij}$ where d_{ij} is the distance between the centroids of area i and area j.
- $W_{ij} = 1$ if d_{ij} is below a threshold and $W_{ij} = 0$ otherwise.

Note that the above three choices give us a symmetric **W**. This not is a requirement. One example of an asymmetric **W** is when $W_{ij} = 1$ if Y_j is one of the K closest areas to Y_i.

Also, by definition, we have $diag(\mathbf{W}) = \mathbf{0}$.

Counties Sharing a Border

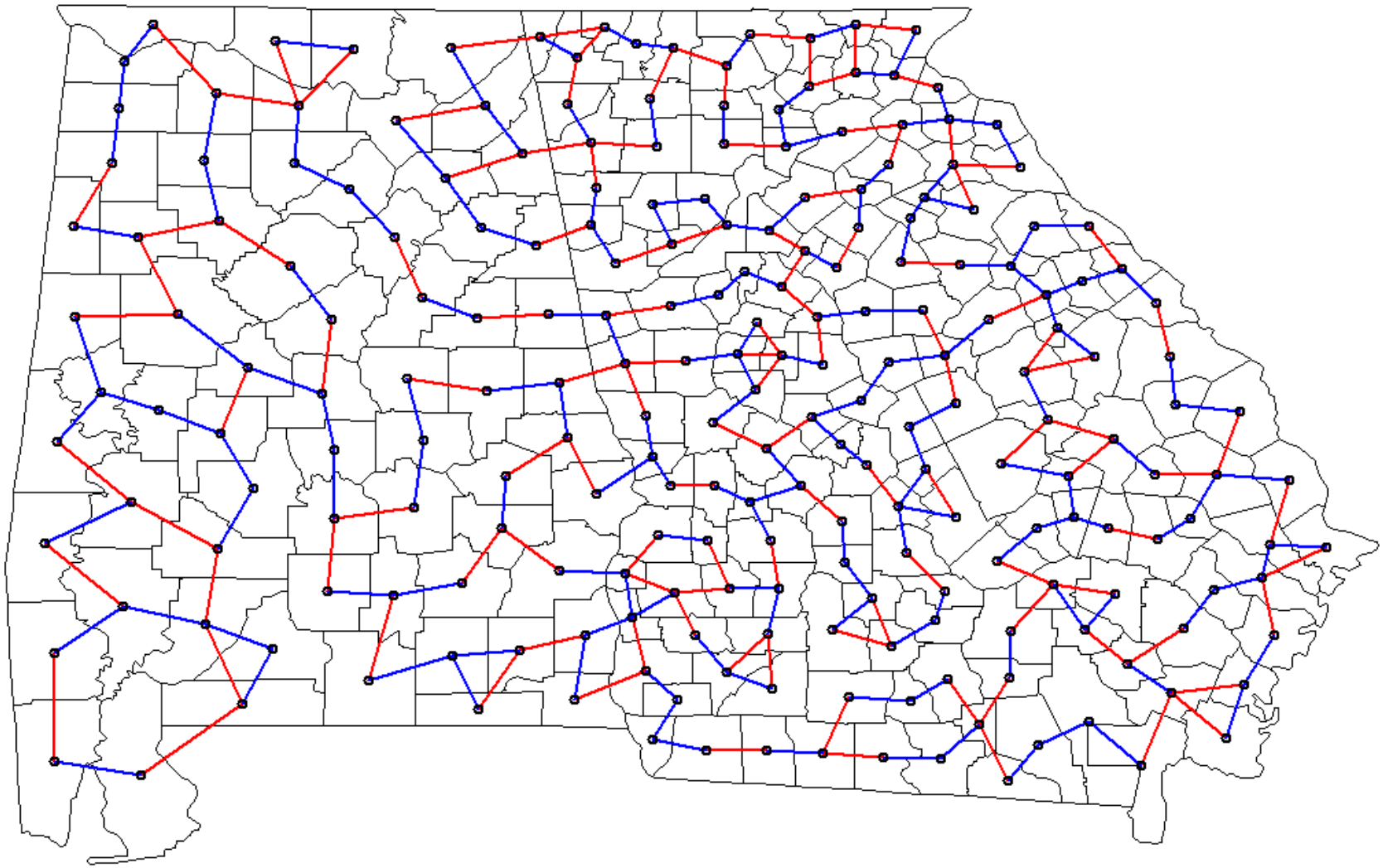


Number of regions = 226
Number of links = 1,258
Average number of links = 5.6

First- or Second-closest County

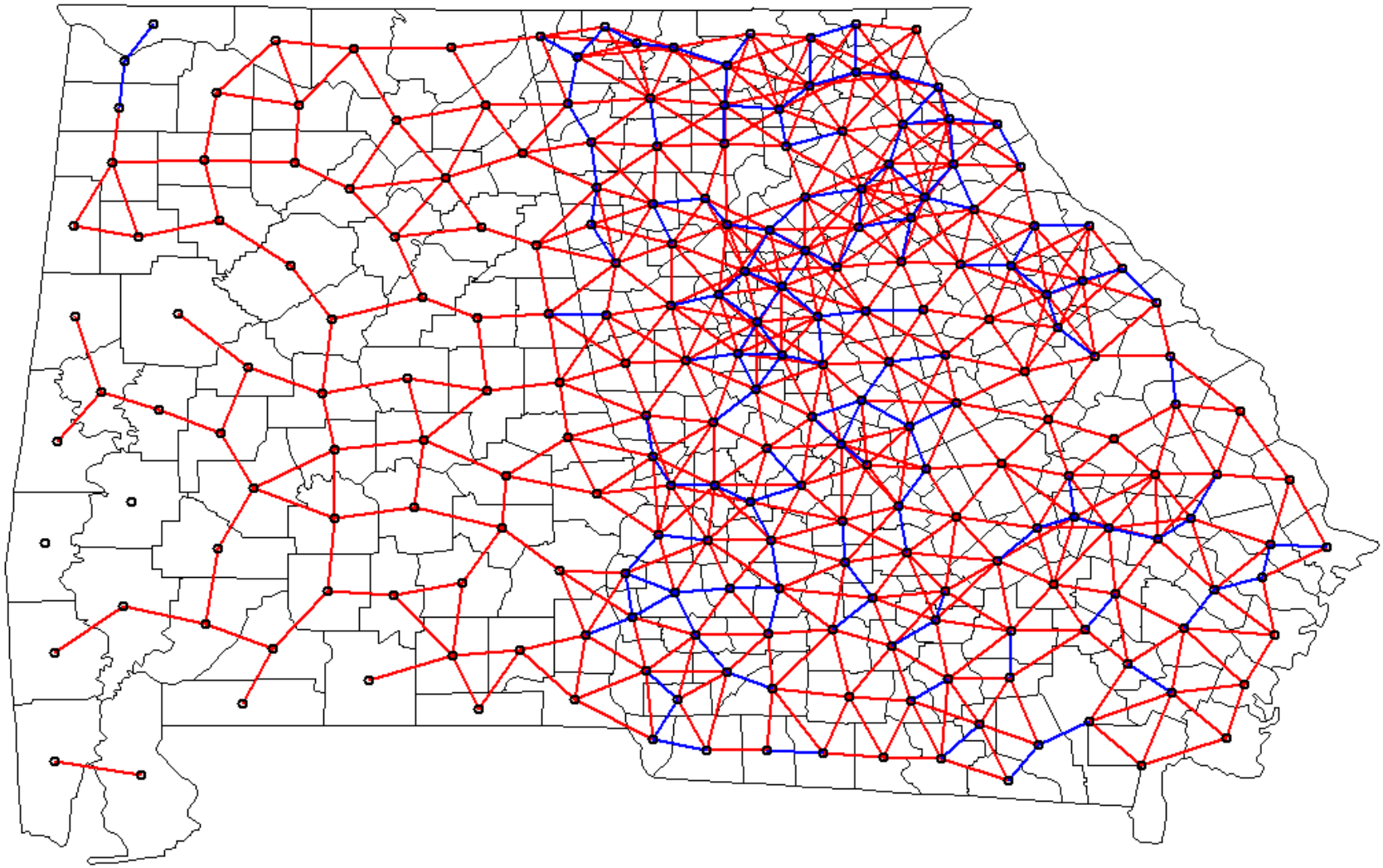
Blue = first closest county

Red = second closest county



Distance Buffer

Blue = centroids within 25km Red = centroids within 40km



Measure of Spatial Dependence

Spatial dependence implies that each Y_i is correlated with its neighbors.

Consider the following regression model

$$\mathbf{Y} = \rho\tilde{\mathbf{W}}\mathbf{Y} + \boldsymbol{\epsilon}$$

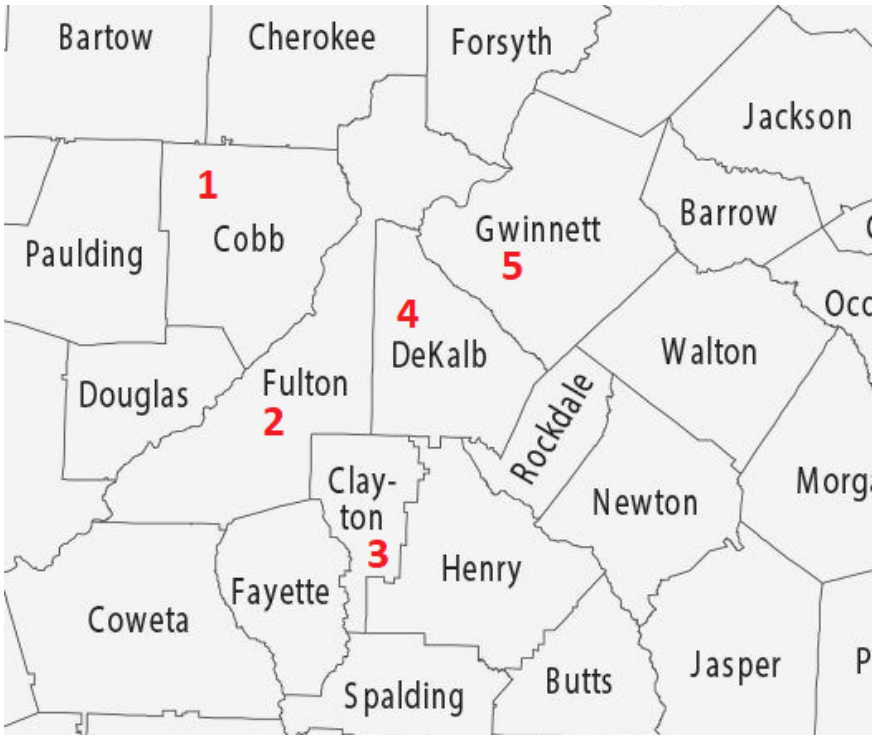
where $\widetilde{\mathbf{W}}$ is a row-standardized (sum to 1) proximity matrix

$$\widetilde{W}_{ij} = W_{ij} / \sum_{i=1}^{n} W_{ij} \ .$$

Then each element of $\widetilde{\mathbf{W}}\mathbf{Y}$ can be viewed as the weighted average of all Y_i's determined by the proximity matrix $\mathbf{W}$. Row-standardization accounts for the different number of neighbors for each polygon.

A positive/negative ρ value indicates positive/negative association between $\mathbf{Y}$ as its spatial averages.

An Example: 5-county Atlanta



$$W = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

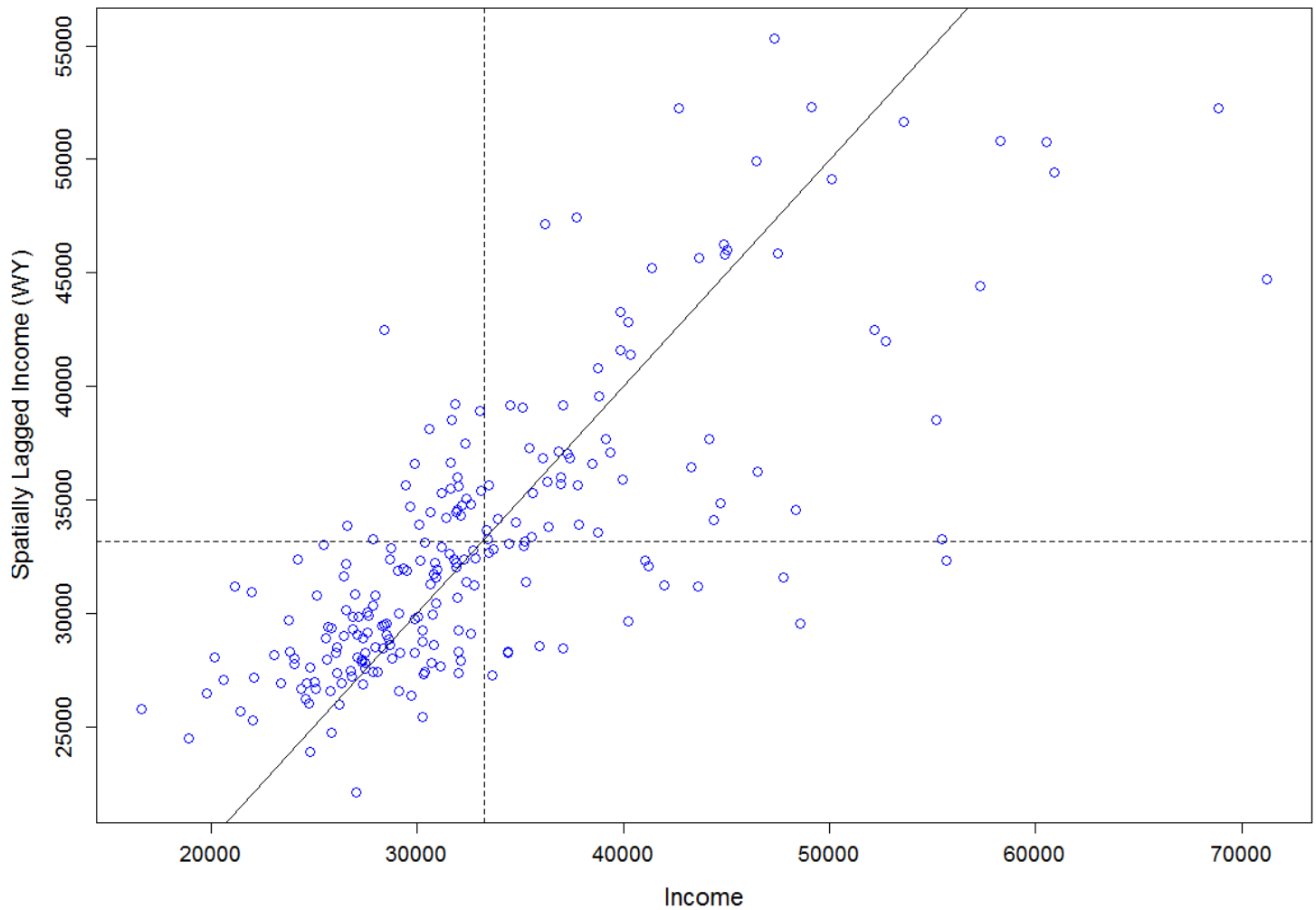
An Example: 5-county Atlanta

$$W = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \qquad \widetilde{W} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \rho\widetilde{W} = \rho \begin{bmatrix} y_2 \\ (y_1 + y_3 + y_4)/3 \\ (y_2 + y_4)/2 \\ (y_2 + y_3 + y_5)/3 \\ y_4 \end{bmatrix}$$

Measure of Spatial Dependence (Moran's Plot)

Spatially-weighted Income ($\tilde{\mathbf{W}}\mathbf{Y}$) versus Observed Income ($\mathbf{Y}$)



Moran's I

Given any proximity matrix $\mathbf{W}$, the Moran's I statistics is

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where $\bar{Y}$ is the average of all Y_i's.

Note that if we use the row-standardized proximity matrix $\tilde{\mathbf{W}}$, then

$$I = (\mathbf{Y}'\tilde{\mathbf{W}}\mathbf{Y})/(\mathbf{Y}'\mathbf{Y})$$

which is identical to the slope coefficient from a linear regression of $\tilde{\mathbf{W}}\mathbf{Y}$ on $\mathbf{Y}$.

- $I > 0$ indicates positive spatial dependence.
- $I < 0$ indicates negative spatial dependence.

Moran's I

- Moran's I is unit-less.
- Moran's I depends on the proximity/weight matrix $\mathbf{W}$.
- Moran's I has a range given by the minimum and maximum eigenvalue of $(\mathbf{W} + \mathbf{W}')/2$

For the income data, we have a Moran's I of 0.541.

The range of possible Moran's I values using a $\mathbf{W}$ defined as the all neighbours sharing a common boundary is -0.566 to 1.016.

Moran's I: Hypothesis Test

In exploratory analysis, we are interested in testing whether I is significantly larger or smaller than 0.

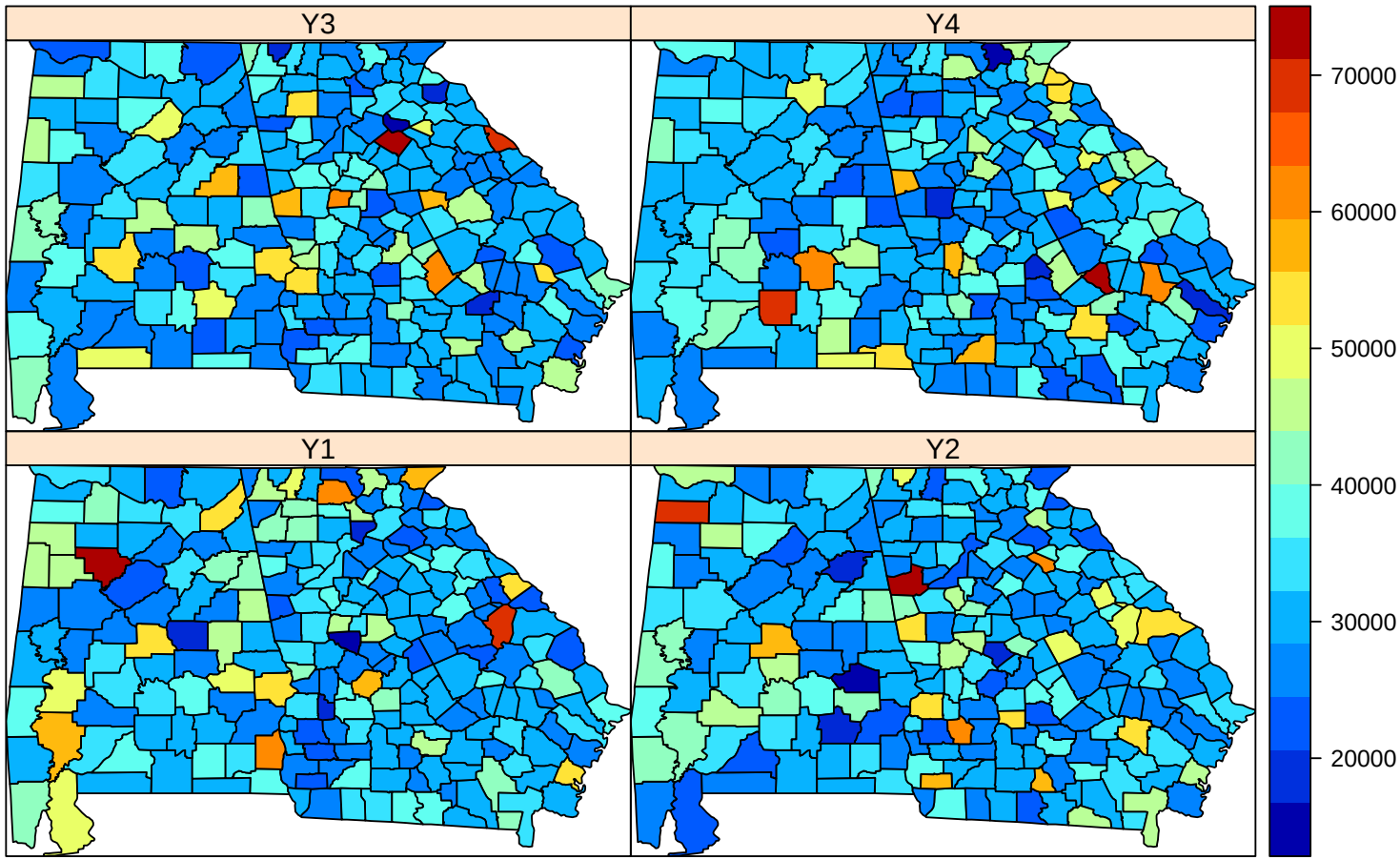
Approach 1: Asymptotic Normality:

- Under the null hypothesis $I = 0$, the expected value of I has mean $-1/(n-1)$ and is asymptotically normal. The standard error is very complicated and depends on whether we assume the outcome Y is Gaussian.
- For our previous income analysis, the Z-value is about 13, indicating strong evidence for the presence of spatial autocorrelation.

Approach 2: Permutation Test:

- The null hypothesis implies that the observed Y_i's is independent of their locations.
- The sampling variability in I under the null hypothesis can be approximated by randomly permuting Y_i across the areal units.

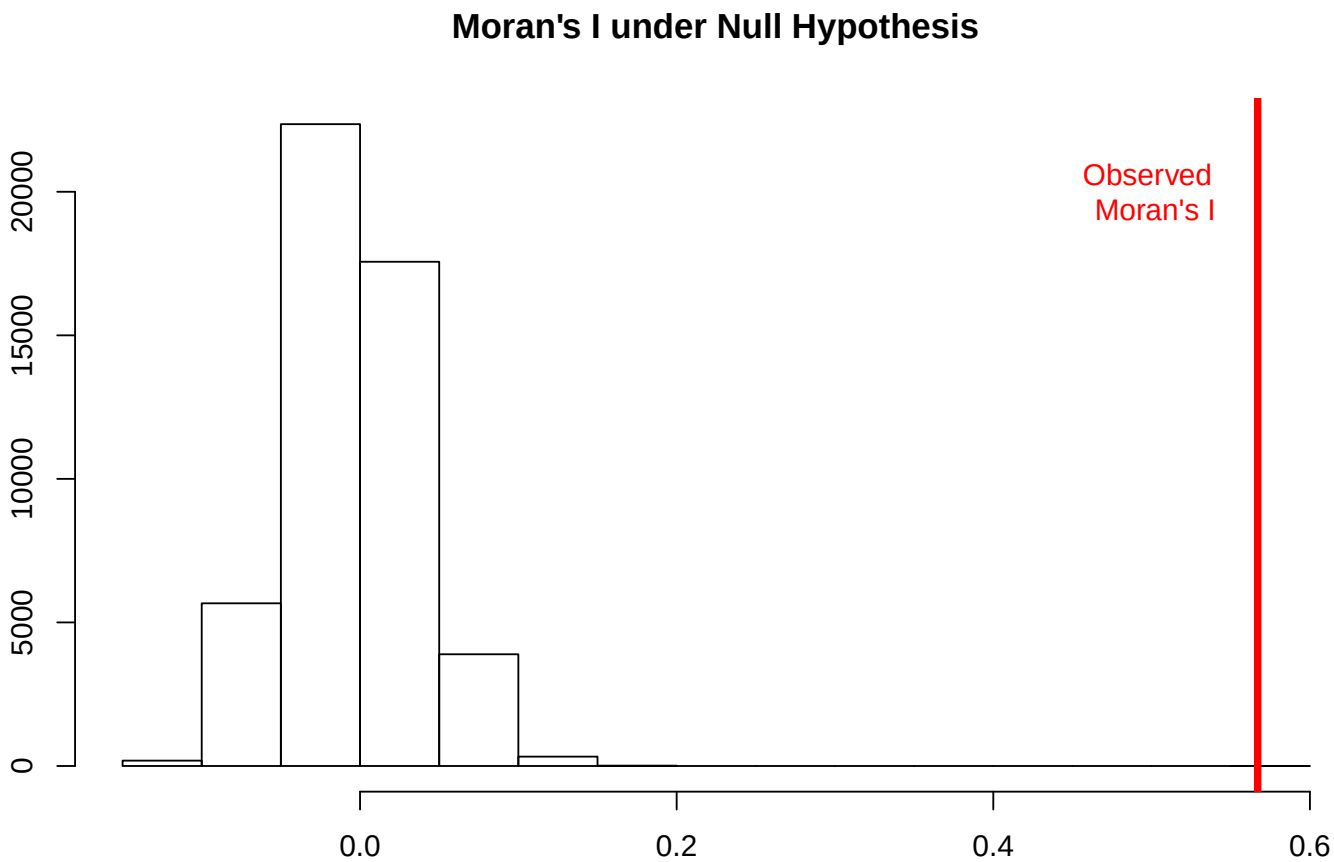
Moran's I: Permutation



The corresponding Moran's I are: -0.02, -0.05, -0.03, 0.02.

Moran's I: Permutation

Permutation p-value $< 1 \times 10^{-4}$ based on 50,000 permutations.



Global versus Local Dependence

The Moran's I statistic measures global spatial dependence because it uses one number of summarize the entire spatial region.

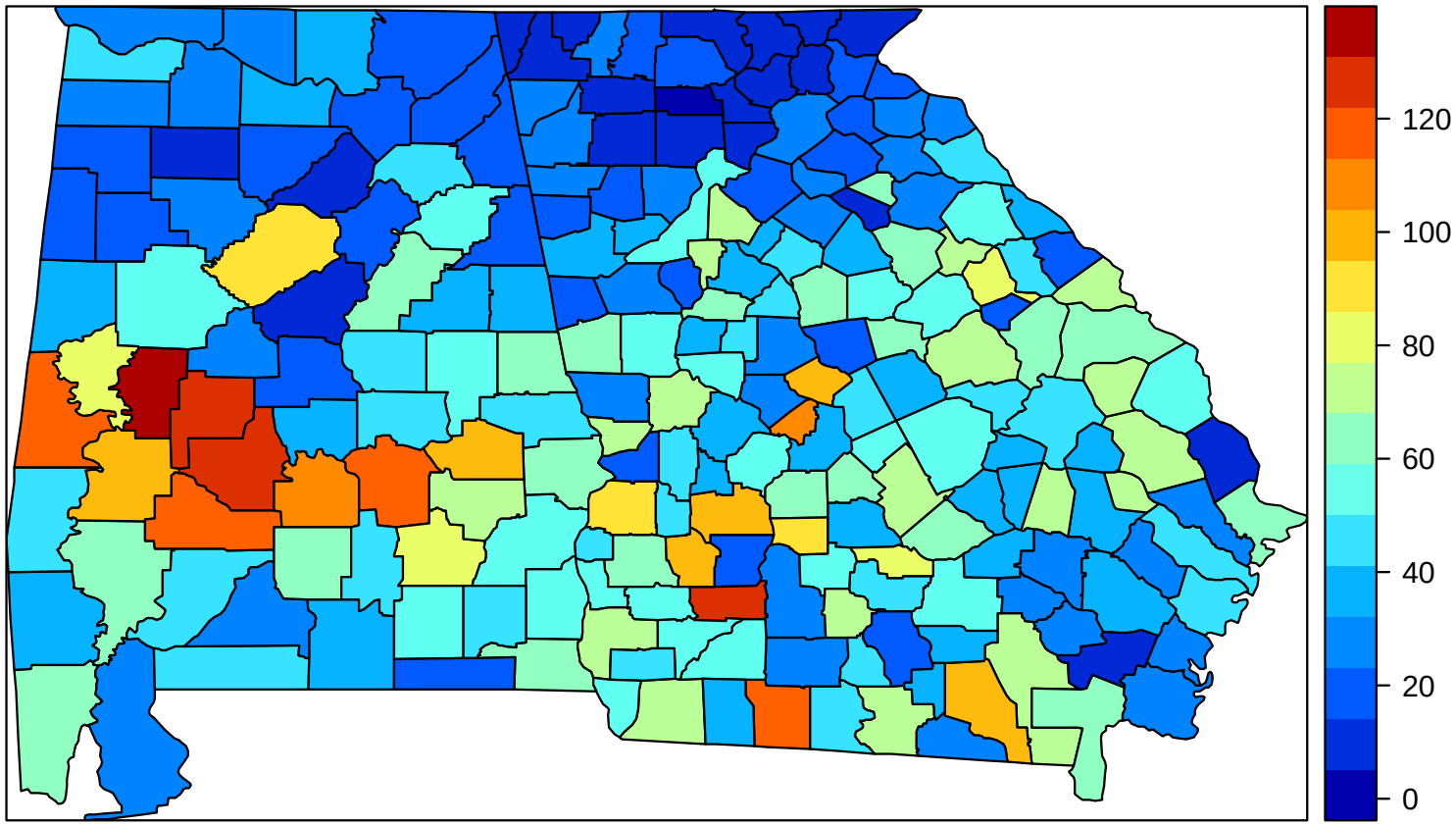
Sometimes *clustering* can occur at spatial sub-regions. A local measure that varies spatially can be useful.

Local Indicator of Spatial Association (LISA)

- The LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation.
- The sum of LISA's for all observation is proportional to a global indicator of spatial association.

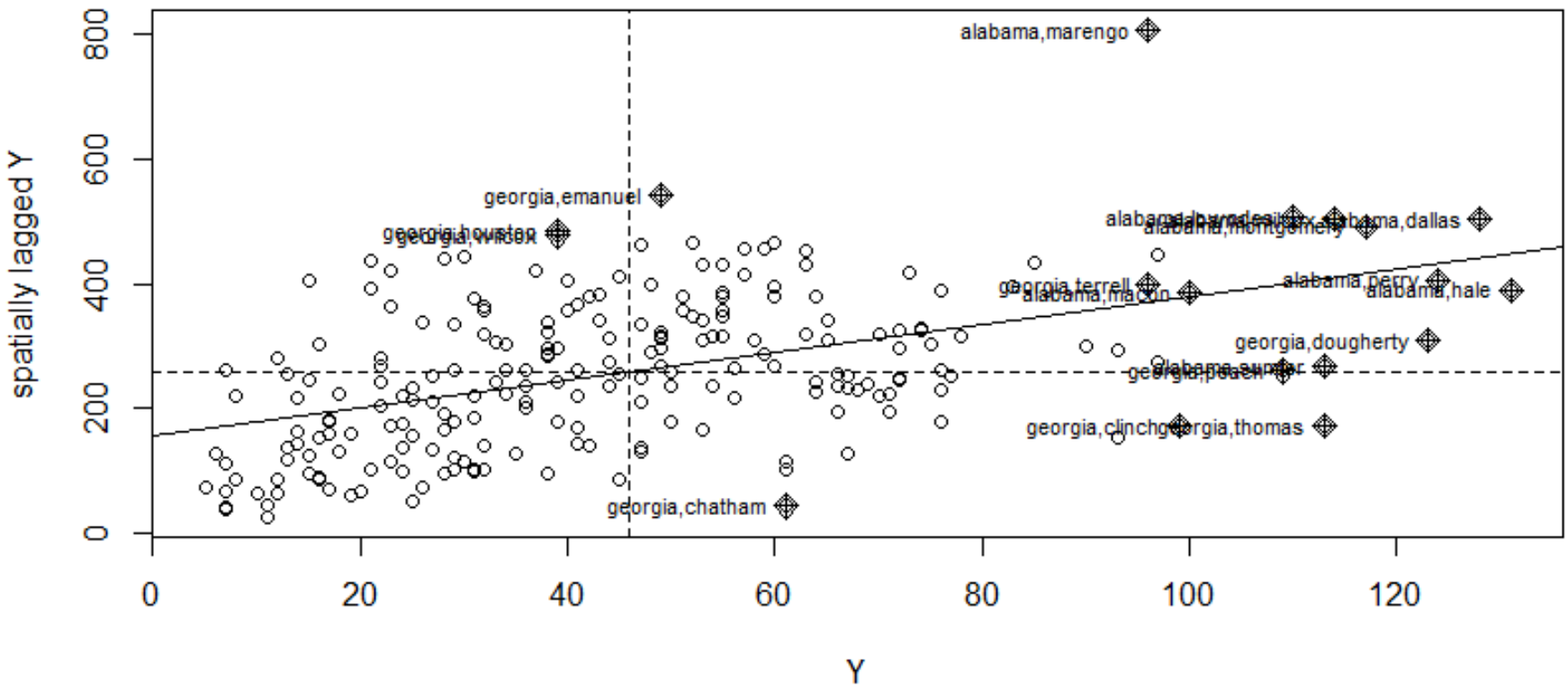
Data Example

County-level Incidence of Chlamydia per 10,000 in 2003



Moran's I Plot for Chlamydia Incidence

Global Moran's I = 0.357
The labeled points have high influence.



Local Moran's I (Anselin 1995)

The local Moran's I statistic, at location i, is defined as

$$I_i = \frac{Y_i - \bar{Y}}{\sum_{k=1}^{n}(Y_k - \bar{Y})^2/(n-1)} \sum_{j=1}^{n} W_{ij}(Y_j - \bar{Y}) \quad .$$

- I_i measures how different Y_i is from its spatial weighted average.
- I_i is scaled by the total variability $s^2 = \sum_k (Y_k - \bar{Y})^2/(n-1)$.

I_i is large when

- $Y_i - \bar{Y}$ is large $\rightarrow$ Y_i is different from the average value.
- $\sum_j W_{ij}(Y_j - \bar{Y})$ is large $\rightarrow$ Y_i's neighbor average is different from the average.

I_i is positive = clustering; I_i is negative = repulsion.

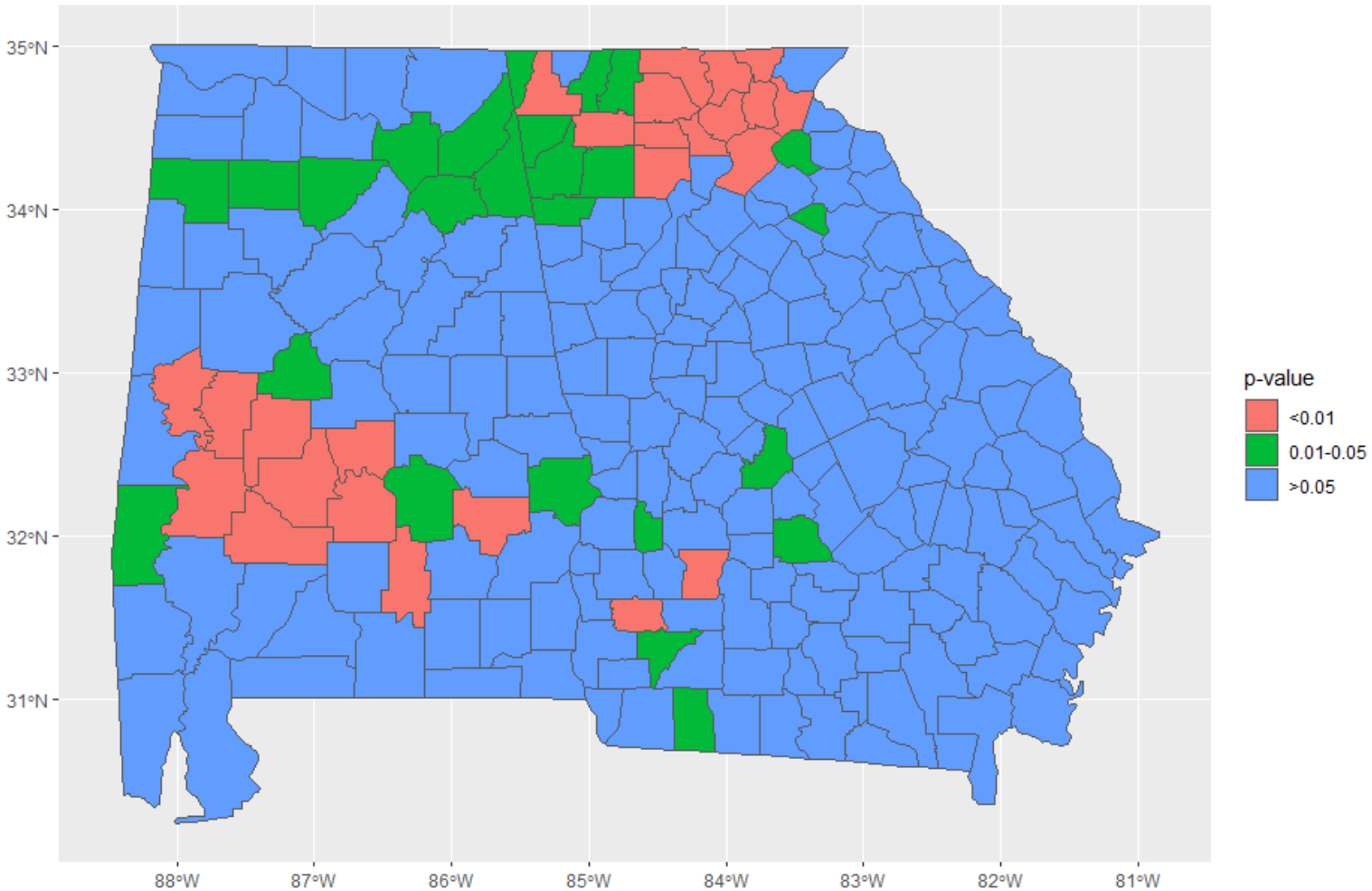
Local Moran's I

Note how the Global Moran's I can be decomposed.

$$
\begin{aligned}
I &= \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\
&= \left(\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \right) \sum_{i=1}^{n} \left[\frac{(Y_i - \bar{Y})}{\sum_{k=1}^{n}(Y_k - \bar{Y})^2} \sum_{j=1}^{n} W_{ij}(Y_j - \bar{Y}) \right] \\
&= \text{Constant} \times \sum_{i=1}^{n} I_i \ .
\end{aligned}
$$

- Sum of I_i is proportional to the global Moran's I.
- For each I_i, we can perform hypothesis test using its asymptotic variance.

Cluster Detecting with Local Moran's I



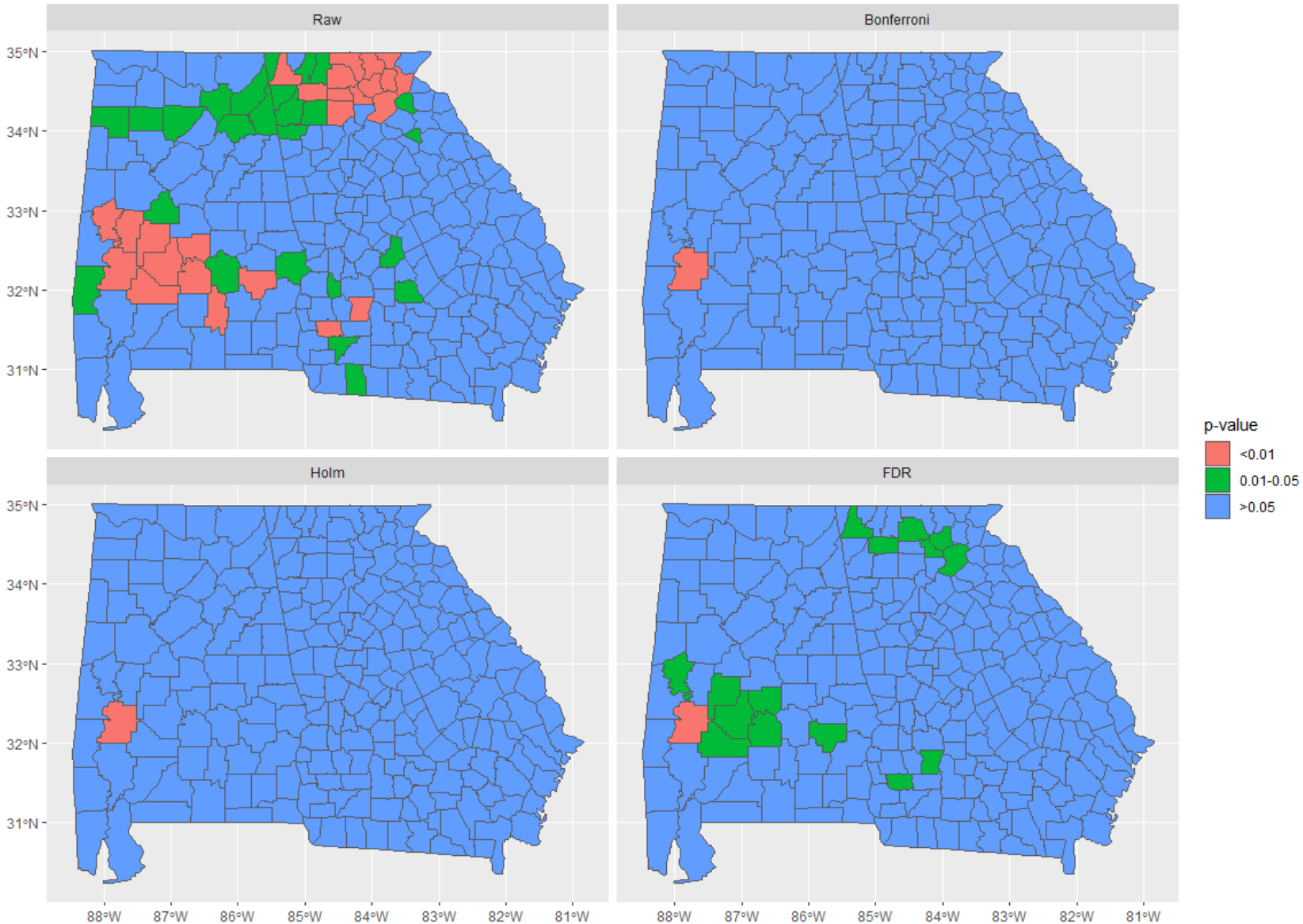
Control for Multiple Testing

For cluster detection, a hypothesis test is performed at each spatial unit. To protect against inflated type I error rate, a multiple testing correction is often implemented.

Let p_i and $\tilde{p}_i$ be the original and the corrected p-value, respectively.

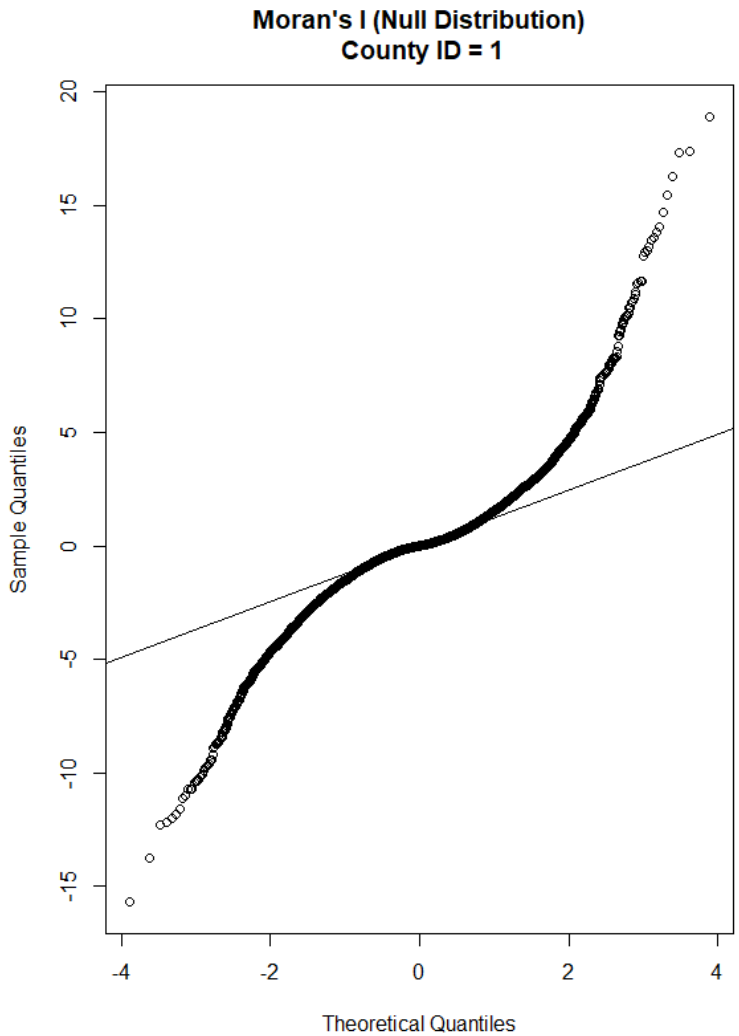
1. Bonferroni
   - $\tilde{p}_i = \min\{1, p_i \times n\}$
   - Can be overly conservative (i.e. true family-wise type I error $<< \alpha$).
2. Holm
   - $\tilde{p}_{[i]} = \min\{1, p_{[i]} * \times (n - i + 1)\}$, where $p_{[i]}$ is the i^{th} smallest p-value.
   - Controls for family-wise type I error rate α but less conservative than Bonferroni (i.e. more power).
3. False-discovery rate (Benjamin-Hochberg)
   - Controls for the expected proportion of falsely rejected H_0 out of total rejections.
   - Less stringent than family-wise type I error (mainly for screening).

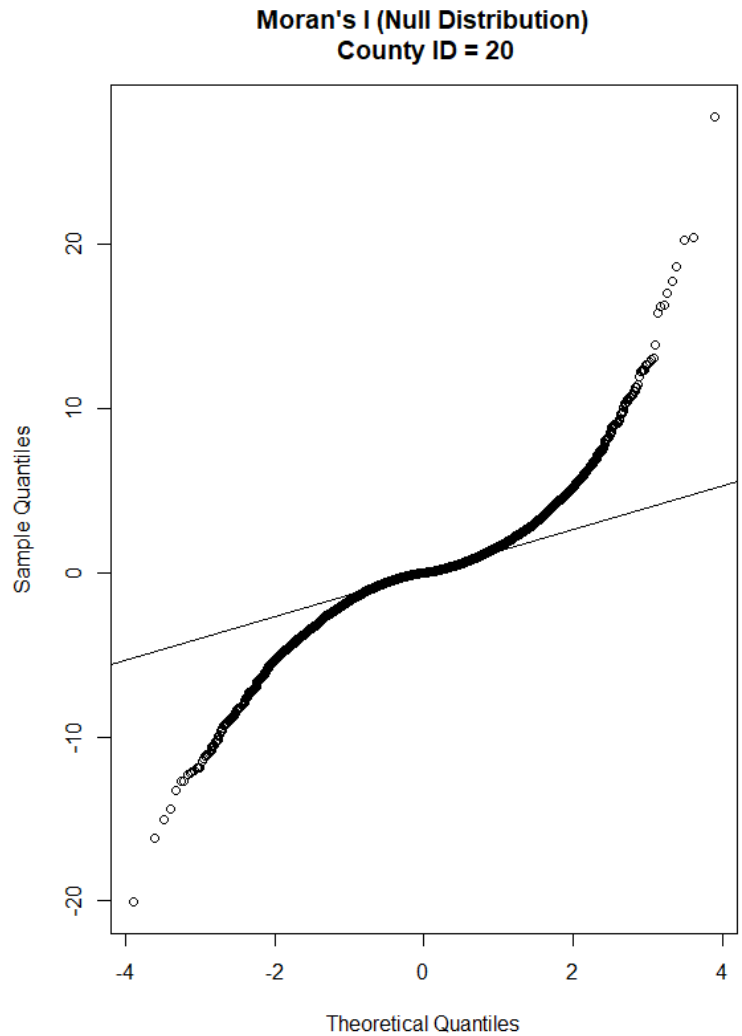
Local Dependence with Local Moran's I



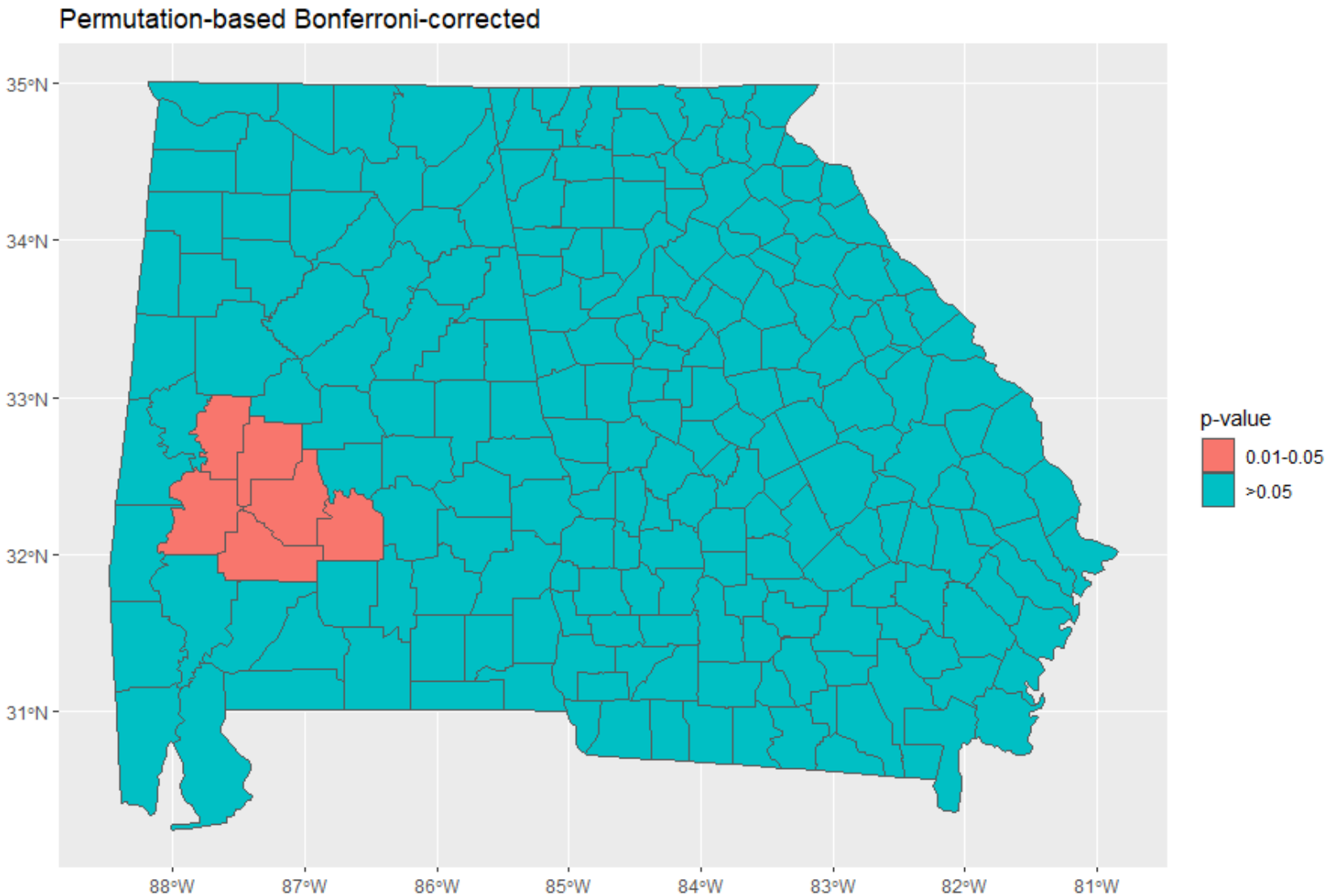
Monte Carlo-Based Inference

Local inference is more prone to normality violation because the neighbors around each spatial unit is often very small.





Permutation Test for Local Moran's I



Cluster Detection and Inference

In a permutation, we re-arrange observed data across the entire study region. Is this a valid approach for estimating local dependence?

An alternative method is to take a model-based approach to simulate data under a null hypothesis.

What is the distribution of the outcome?

Disease Cluster Detection

For disease mapping applications, there has been extensive work on cluster detection. This main difference is the addition of a parametric assumption on the count data.

The general framework is a Poisson model

$$O_i \sim \text{Poisson } (\theta_i E_i) \ .$$

- O_i is the observed case count.
- E_i is the expected number of cases. Usually $E_i = P_i \theta_0$ where P_i is the at risk population size and θ_0 is the overall incidence ratio $= O_+/P_+$.
- θ_i is the location-specific relative rate.

Without clustering, conditioned on the total number of cases, the vector of observed counts $O_1, O_2, \ldots, O_n$ follows a multinomial distribution with probabilities $E_1/E_+, E_2/E_+, \ldots, E_n/E_+$.

Test of Homogeneity: χ^2 Goodness-of-fit

The first exploratory test is whether the relative risks are constant across regions. Since each region has an expected value, we can perform a chi-square goodness-of-fit test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - \theta_i E_i)^2}{\theta_i E_i}$$

which follows a χ^2 distribution with n degrees of freedom. If E_i's are calculated from the data, then the degrees of freedom $= n - 1$.

However, significance is best determined by simulation.

- Permutation
- Parametric Simulation: Assume O_i follows Poisson, multinomial, or negative binomial.

Test of Homogeneity: Test for Overdispersion

The Potthoff-Whittinghill test is a specific test for relative risk homogeneity but with a specific alternative hypothesis:

$$H_0 = \theta_1 = \theta_2 = \ldots = \theta_n = \theta_0$$
$$H_1 = \theta_i \sim \text{Gamma } (\theta_0^2/\sigma^2, \theta_0/\sigma^2)$$

H_1 assigns a Gamma distribution to the location-specific relative risks. The consequence is that this model corresponds to a negative-binomial distribution for O_i.

The Potthoff-Whittinghill test statistic is given by

$$PW = E_+ \sum_{i=1}^{n} \frac{O_i(O_i - 1)}{E_i}$$

which follows an asymptotic normal distribution with mean $O_+(O_+ - 1)$ with variance $2nO_+(O_+ - 1)$ under H_0.

Is There Risk Heterogeneity?

χ^2 Goodness-of-Fit Test

- Test statistic = 21,755
- P-values:
  - < 0.001 (Asymptotic)
  - < 0.001 (Poisson simulation)
  - < 0.001 (Multinomial simulation)
  - < 0.394 (Negative binomial simulation)

Potthoff-Whittinghill Overdispersion Test

- Test statistic = 6,032,784,494
- P-values:
  - < 0.001 (Asymptotic)
  - < 0.001 (Poisson simulation)
  - < 0.001 (Multinomial simulation)

These two tests do not explicitly consider spatial structure (coming in Lecture 4)!