Visualizing Labor Migration Using Quantitative Data

Exploratory Spatial Data Analysis

Outline

- Overview of spatial data analysis
- Opportunities and challenges
- Overview of spatial effects
- Walk through the county homicide example

The Spatial Statistics Family

- Geostatistics
  - Focus on processes in a continuous space
- Point pattern analysis
  - Focus on the location of events
- The analysis of lattice data
  - Focus on processes in a discrete space
    - Regular vs. irregular lattices

Our focus is on the use of spatial econometrics to examine areal data.

The opportunities and challenges of spatial data analysis lie in the potential association between value similarity and spatial similarity.

What’s Special about Spatial Data?

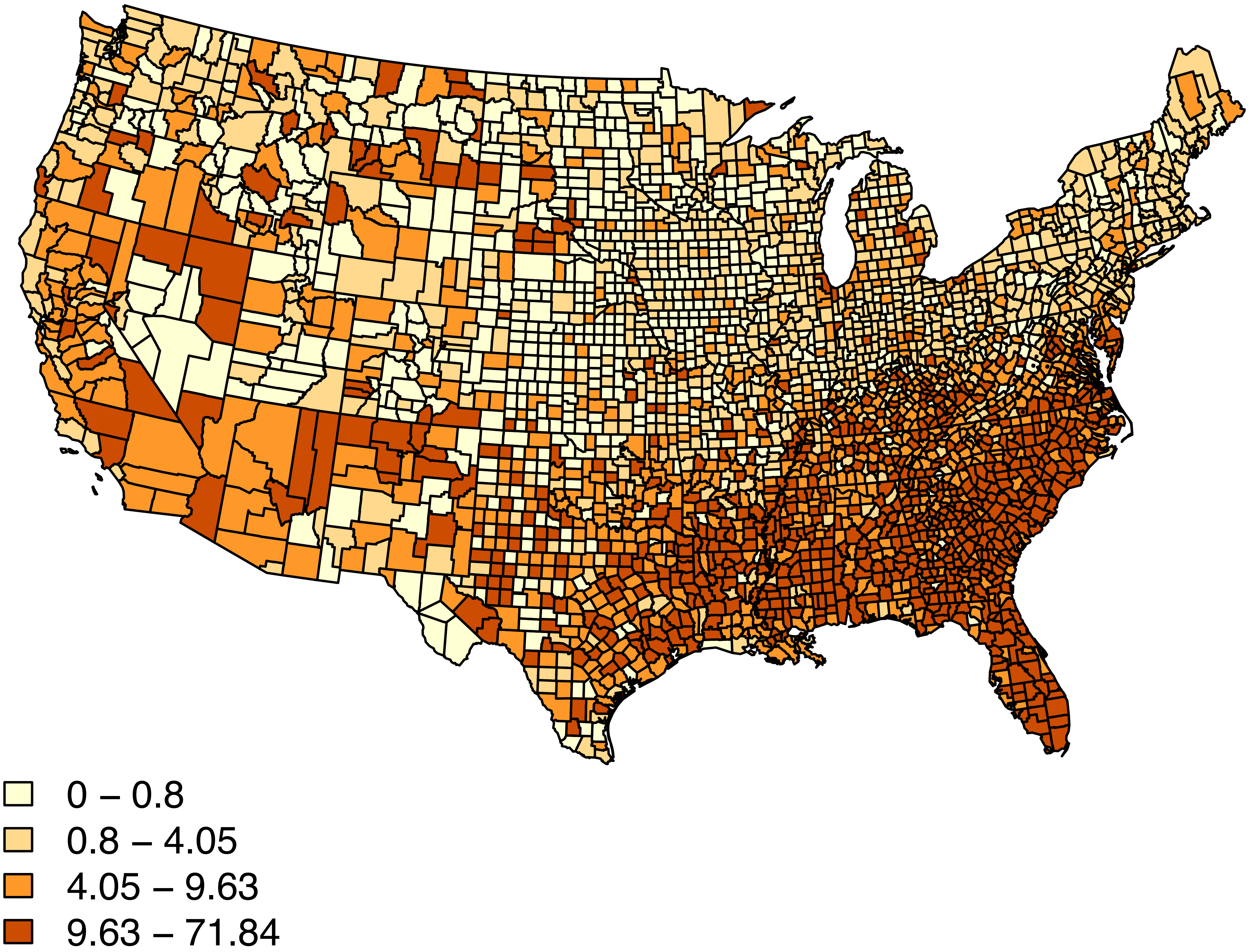
- The modifiable areal unit problem
- The problem of apparent populations
- Galton’s problem

Example: County Homicide Rates

- Baller et al. (2001) examine county-level homicide rates in the United States between 1960 and 1990
- Find evidence of regional differences in the underlying social process
  - Differences in both structural predictors and spatial processes
    - Lag in the South
    - Error in the Non-South (except in 1960)

Spatial data analysis is a little bit like solving a murder mystery in which the culprit is the underlying spatial process.

County Homicide Rates, 1970



Let’s find out how to do this all in R!

How do we explain this particular arrangement of values?

We focus on three possible suspects, either working alone or in combination with one another.

Spatial Processes

- Structural similarity
- Heterogeneity
  - Discrete vs. continuous
- Dependence
  - Error vs. lag

When it comes to figuring out which of these are in play, we have a number of clues at our disposal…

Spatial Autocorrelation

- Concept
  - Interdependence between neighboring observations
    - Positive vs. negative
- Measure
  - Moran's *I*
    - Global vs. local

Moran's *I*

- Global
  - Measures the overall level of autocorrelation
  - Equal to the slope of the line associated with the regression of $\mathbf{Wy}$ on $\mathbf{y}$
  - Depicted using a Moran scatterplot
- Local
  - Identifies clustering or instability
  - Equal to the set of case-level contributions to the global measure
  - Depicted using a LISA map
- Parametric and non-parametric significance tests for both

Spatial Weights Matrices

- $\mathbf{W}$ is an $n \times n$ adjacency matrix depicting the relationship between pairs of observations
  - Contiguity vs. distance
    - Binary vs. valued ties
  - Symmetric vs. asymmetric matrices
    - Valid but problematic
- Typically row-standardized
  - Helps interpretation and estimation
- Diagonal entries (i.e., self-ties) are usually set to 0 by convention
- Can be generalized to accommodate non-spatial ties

Example: Network Multiplexity and the Paris Commune

- Gould (1991) examines patterns of mass insurgency during the Paris Commune of 1871
- Uses three different weights matrices
  - Spatial adjacency
  - Enlistment patterns
  - Transpose of enlistment patters
- Enlistment patterns are the most important

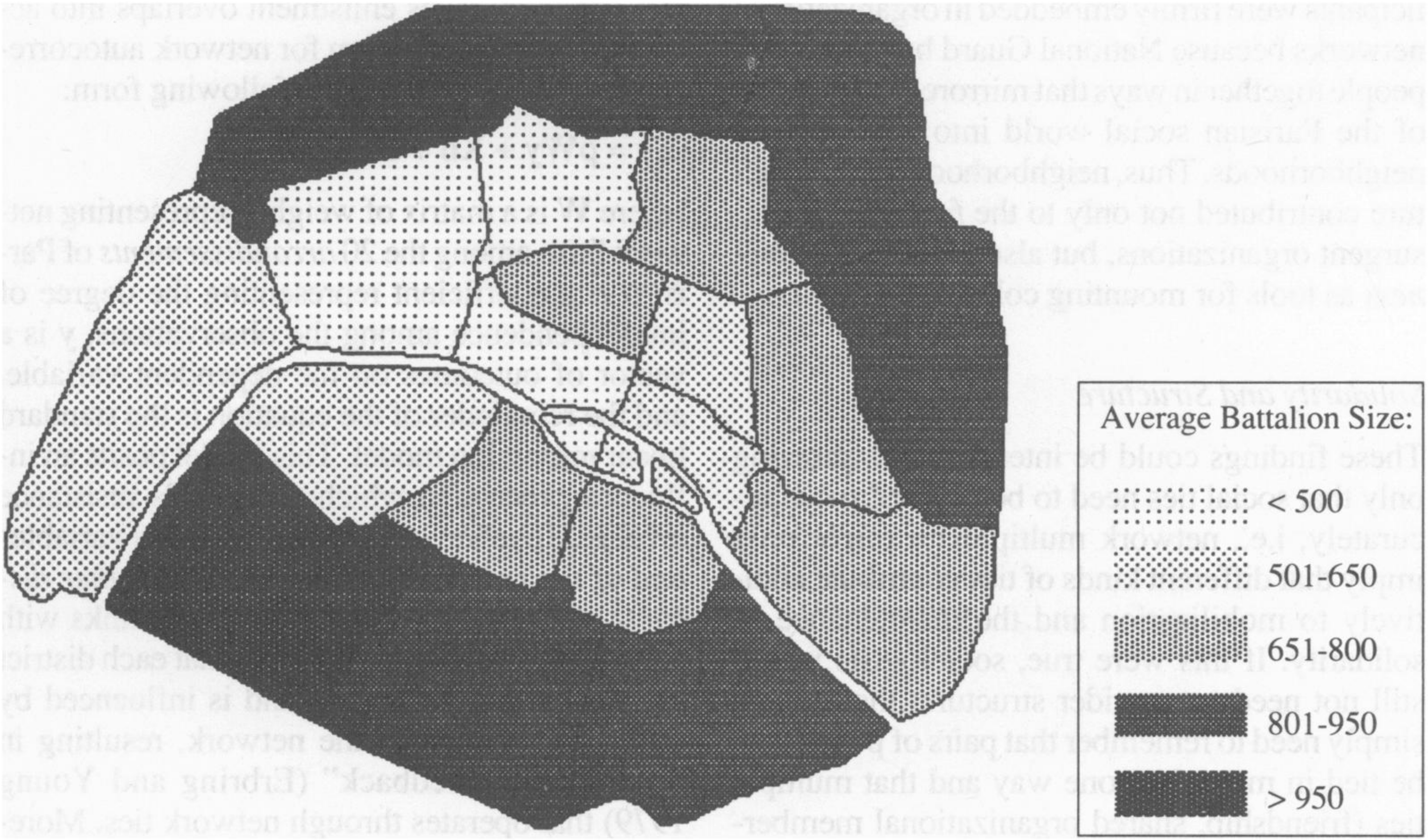


Figure 2. Average Battalion Size in Late May, by *Arrondissement*: Paris Commune, 1871

Note: Although battalion sizes vary from early to late May, the differences are not noticeable in this form of presentation.

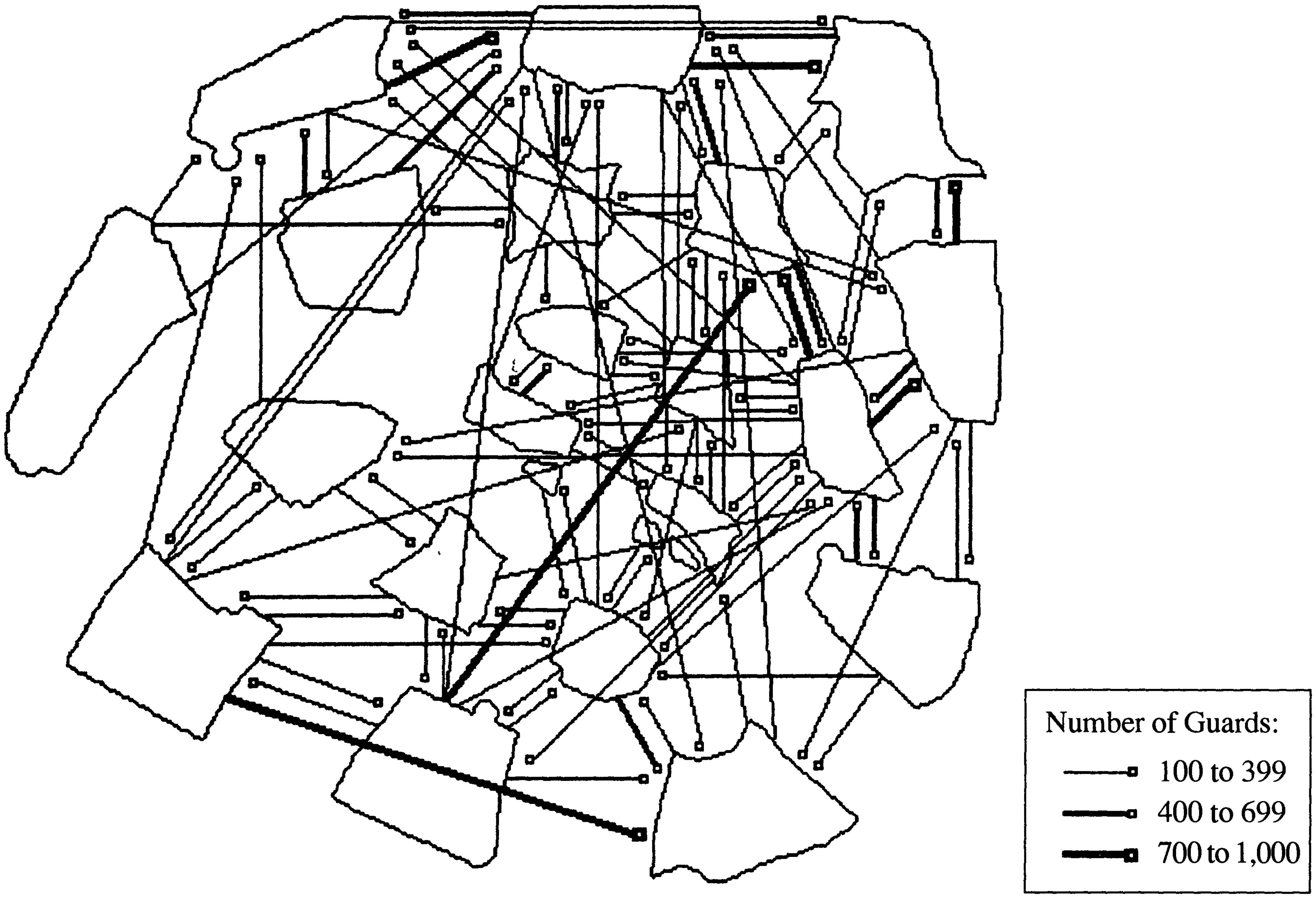


Figure 4. Numbers of National Guardsmen Serving in Legions Outside Their *Arrondissement* of Residence, by *Arrondissement* : Paris Commune, 1871

Note: Enlistment overlaps of fewer than 100 National Guardsmen are not shown; inclusion of such links would connect each *arrondissement* with nearly every other. Directionality is indicated by a hollow square, i.e., if 150 inhabitants of District A serve in a National Guard battalion from District B, this would appear as a thin line from A ending in a square near B.

Table 1. Coefficient Estimates for Average Battalion Size and Death Rate on Selected Independent Variables: Paris Commune, 1871

	Battalion Size					
	Early May		Late May		Death Rate, May 1871	
Independent Variable	Network Model (1)	Spatial Model (2)	Network Model (3)	Spatial Model (4)	Network Model (5)	Spatial Model (6)
Autocorrelation (ρ)	.289*	-.118	.477**	.038	.487*	.030
February military deaths	—	—	—	—	.076**	.068**
Poverty rate	2.217	2.419	2.217	2.320	16.818	18.103
Percent skilled workers	9.163*	9.311**	8.040*	8.164**	.064	.054
Percent unskilled workers	7.671	7.743	8.523	7.765	.081	.068
Percent white-collar employees	8.438	6.667	12.074	10.869	.066	.036
Constant	-148.918	180.656	-347.618	8.597	-4.650	-1.715
Fit[a]	.728	.722	.703	.674	.471	.441
Number of *arrondissements*	20	20	20	20	20	20

*$p < .05$ (one-tailed) **$p < .01$ (one-tailed)

[a] “Fit” is the square of the correlation between the observed and predicted values of the dependent variable. While it corresponds roughly to R^2 in standard regression analysis, it is not strictly comparable and should not be interpreted as the percentage of variance explained.

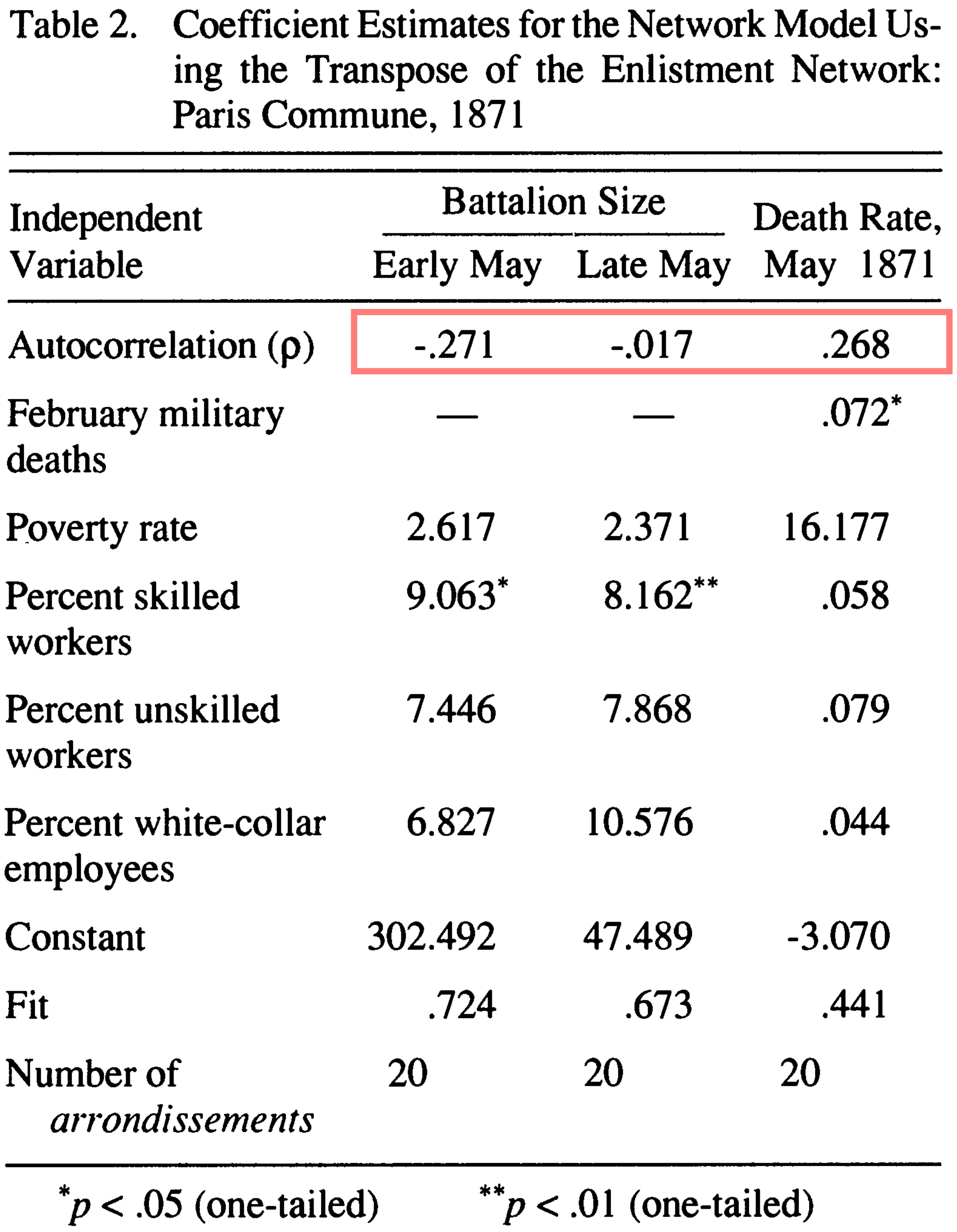


Table 2. Coefficient Estimates for the Network Model Using the Transpose of the Enlistment Network: Paris Commune, 1871

Independent Variable	Battalion Size		Death Rate,
	Early May	Late May	May 1871
Autocorrelation (ρ)	-.271	-.017	.268
February military deaths	—	—	.072*
Poverty rate	2.617	2.371	16.177
Percent skilled workers	9.063*	8.162**	.058
Percent unskilled workers	7.446	7.868	.079
Percent white-collar employees	6.827	10.576	.044
Constant	302.492	47.489	-3.070
Fit	.724	.673	.441
Number of *arrondissements*	20	20	20

*$p < .05$ (one-tailed) **$p < .01$ (one-tailed)

Let’s find out how to do this all in R!

Moran's *I*

- Global
  - Measures the overall level of autocorrelation
  - Equal to the slope of the line associated with the regression of $\mathbf{Wy}$ on $\mathbf{y}$
  - Depicted using a Moran scatterplot
- Local
  - Identifies clustering or instability
  - Equal to the set of case-level contributions to the global measure
  - Depicted using a LISA map
- Parametric and non-parametric significance tests for both

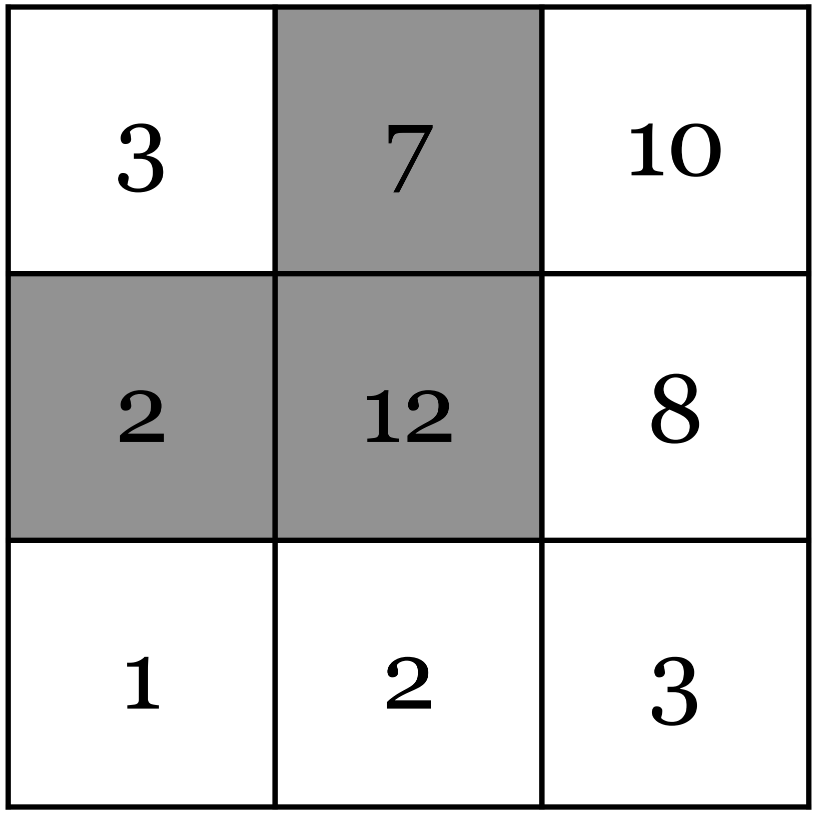
3	7	10
2	12	8
1	2	3

3	7	10
2	12	8
1	2	3

y
3
7
10
2
12
8
1
2
3

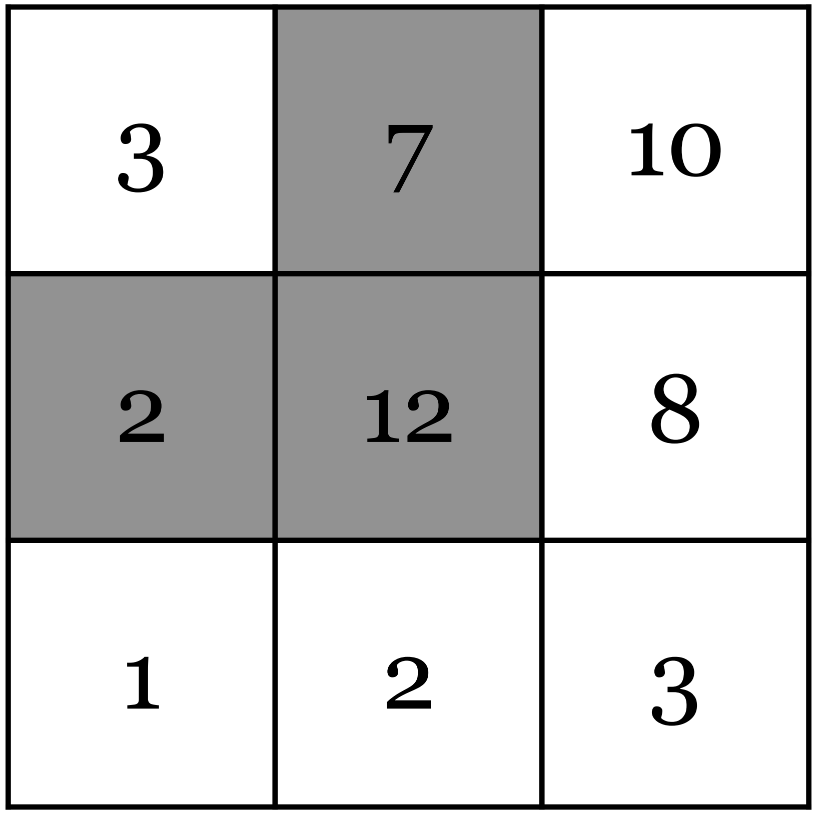
3	7	10
2	12	8
1	2	3

y	y*
3	
7	
10	
2	
12	
8	
1	
2	
3	



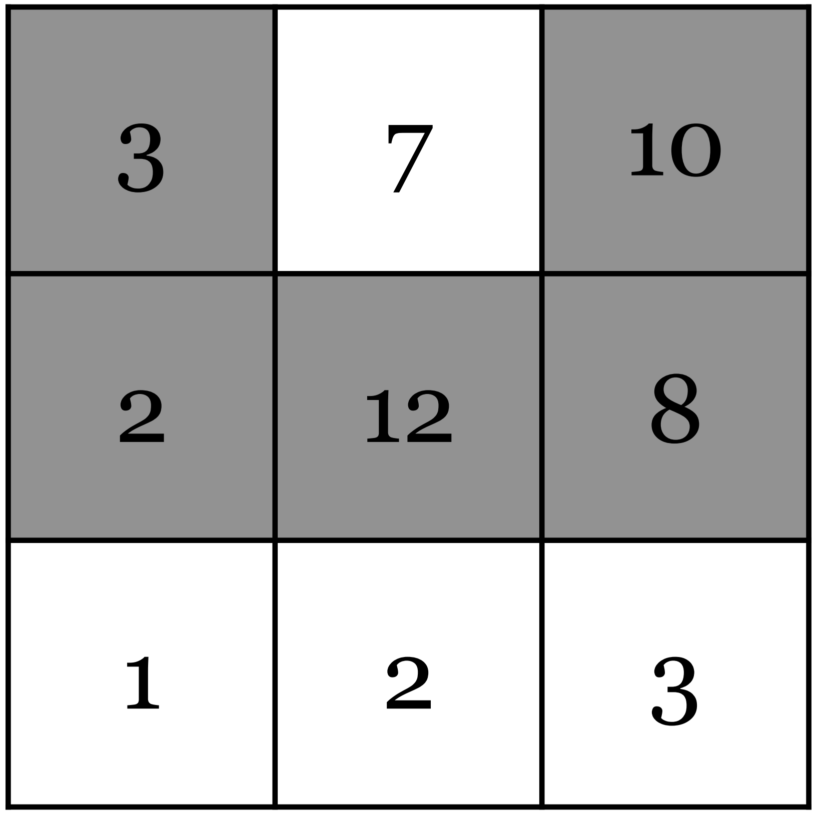
3	7	10
2	12	8
1	2	3

y	y*
3	
7	
10	
2	
12	
8	
1	
2	
3	



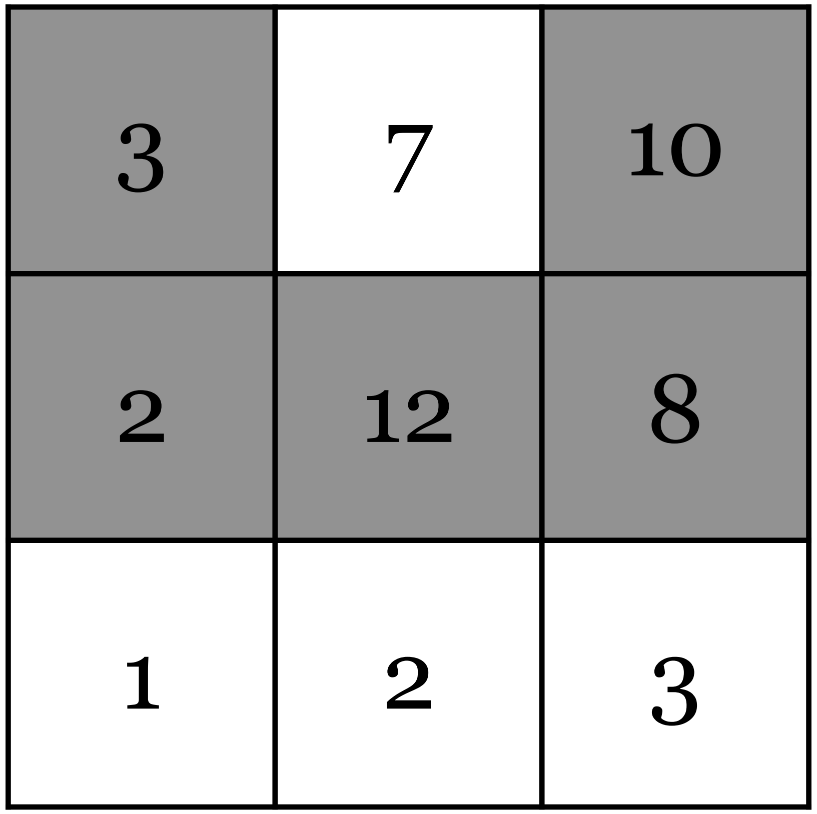
3	7	10
2	12	8
1	2	3

y	y*
3	7
7	
10	
2	
12	
8	
1	
2	
3	



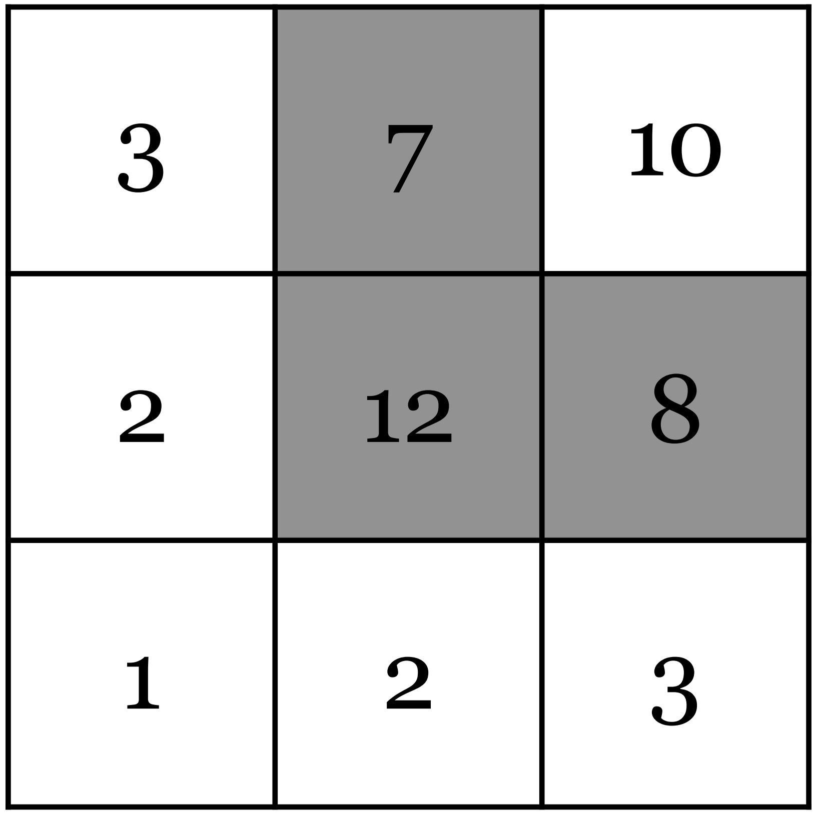
3	7	10
2	12	8
1	2	3

y	y*
3	7
7	
10	
2	
12	
8	
1	
2	
3	



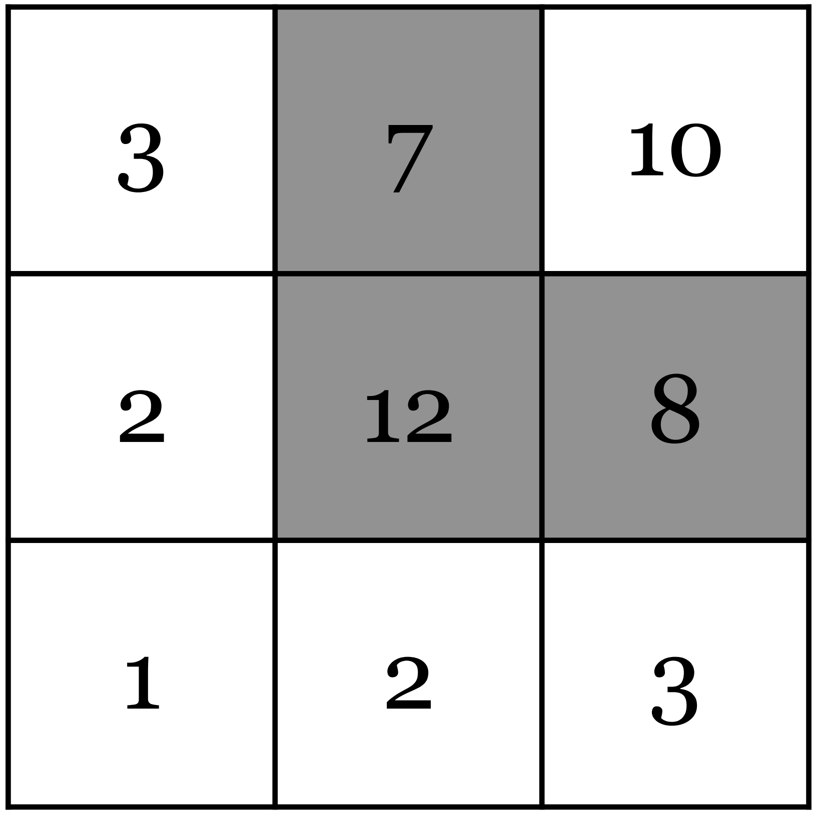
3	7	10
2	12	8
1	2	3

y	y*
3	7
7	7
10	
2	
12	
8	
1	
2	
3	



3	7	10
2	12	8
1	2	3

y	y*
3	7
7	
10	
2	
12	
8	
1	
2	
3	



3	7	10
2	12	8
1	2	3

y	y*
3	7
7	7
10	9
2	
12	
8	
1	
2	
3	

3	7	10
2	12	8
1	2	3

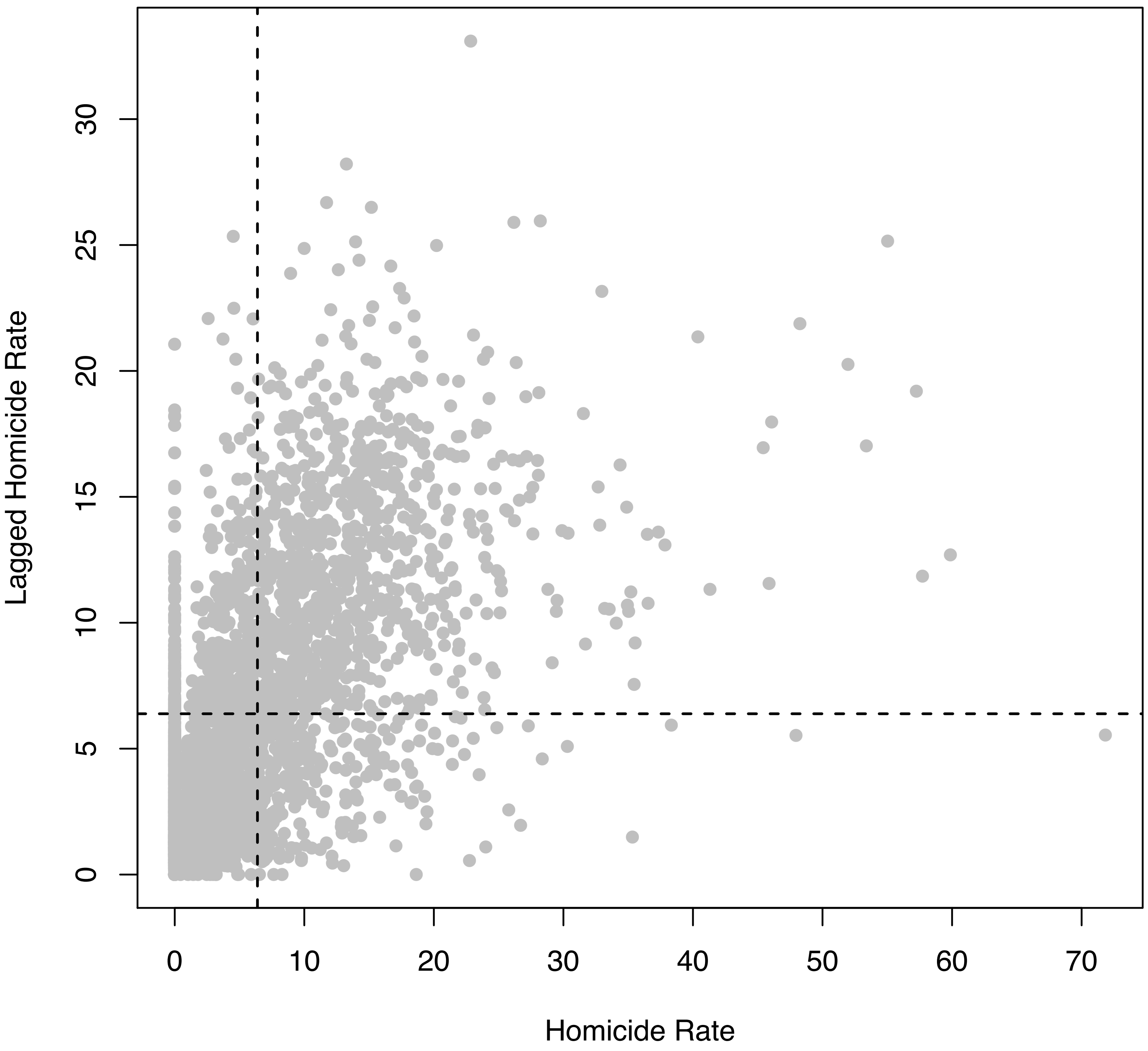
y	y*
3	7
7	7
10	9
2	5
12	4
8	6.8
1	5·33
2	5·2
3	7·33

				W					y	y*
0	1	0	1	1	0	0	0	0	3	7
1	0	1	1	1	1	0	0	0	7	7
0	1	0	0	1	1	0	0	0	10	9
1	1	0	0	1	0	1	1	0	2	5
1	1	1	1	0	1	1	1	1	12	4
0	1	1	0	1	0	0	1	1	8	6.8
0	0	0	1	1	0	0	1	0	1	5·33
0	0	0	1	1	1	1	0	1	2	5·2
0	0	0	0	1	1	0	1	0	3	7·33

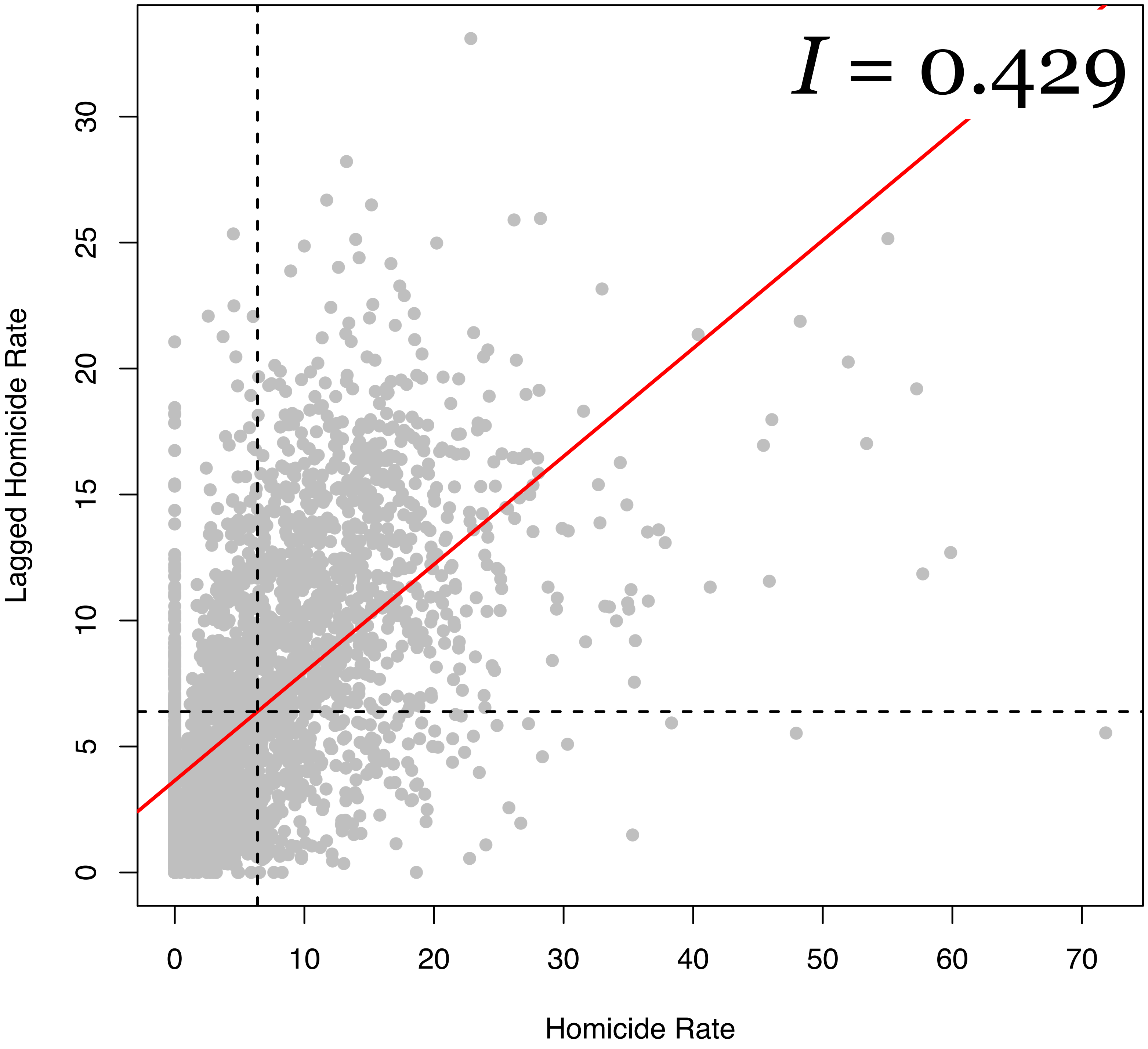
W									y	y*
0	0.33	0	0.33	0.33	0	0	0	0	3	7
0.2	0	0.2	0.2	0.2	0.2	0	0	0	7	7
0	0.33	0	0	0.33	0.33	0	0	0	10	9
0.2	0.2	0	0	0.2	0	0.2	0.2	0	2	5
0.125	0.125	0.125	0.125	0	0.125	0.125	0.125	0.125	12	4
0	0.2	0.2	0	0.2	0	0	0.2	0.2	8	6.8
0	0	0	0.33	0.33	0	0	0.33	0	1	5.33
0	0	0	0.2	0.2	0.2	0.2	0	0.2	2	5.2
0	0	0	0	0.33	0.33	0	0.33	0	3	7.33

$$\mathbf{y}^* = \mathbf{W}\mathbf{y}$$

Moran Scatterplot of County Homicide Rates, 1970



Moran Scatterplot of County Homicide Rates, 1970



Let's find out how to do this all in R!

Moran's *I*

- Global
  - Measures the overall level of autocorrelation
  - Equal to the slope of the line associated with the regression of $\mathbf{Wy}$ on $\mathbf{y}$
  - Depicted using a Moran scatterplot
- Local
  - Identifies clustering or instability
  - Equal to the set of case-level contributions to the global measure
  - Depicted using a LISA map
- Parametric and non-parametric significance tests for both

Local Moran's *I* as a LISA Statistic

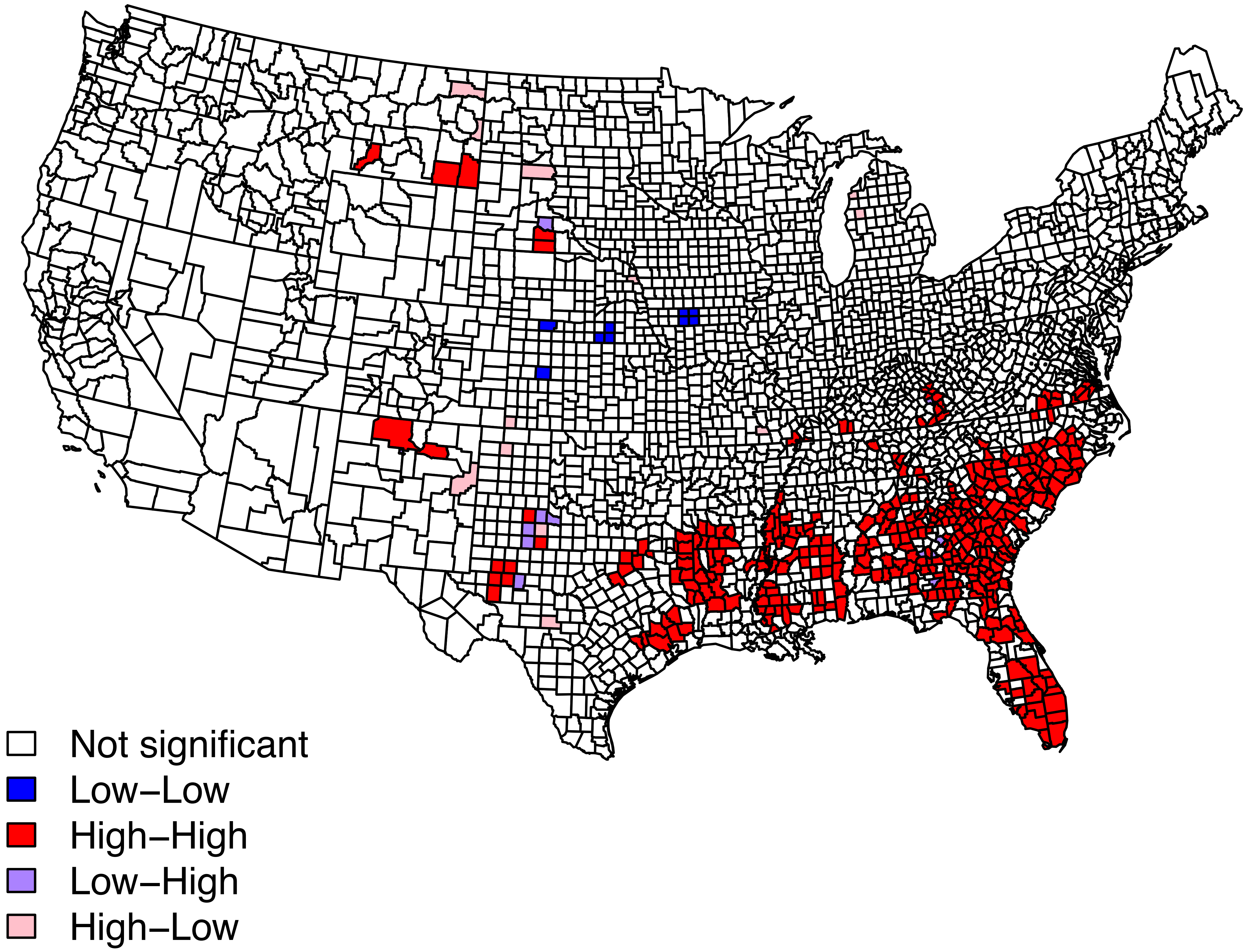
- A LISA statistic is any statistic that meets the following requirements:
  - The LISA for any given observation measures the degree of clustering around the observation in question
  - The sum of the LISAs is proportional to a global measure of spatial association
- To construct the LISA statistic associated with the global Moran's *I* we start with $z_i z_i^*$ as our measure of local clustering
  - $\mathbf{z} = \mathbf{y} - \mathbf{1}\bar{y}$
  - $\mathbf{z}^* = \mathbf{Wz}$

Assuming that $\mathbf{W}$ is row-standardized, the following relationships hold…

$$
\begin{aligned}
\sum_i L_i &= \gamma \Lambda \\
\sum_i I_i &= N I \\
\frac{\sum_i I_i}{N} &= I
\end{aligned}
$$

$$
\begin{aligned}
I &= \left(\frac{\sum_i z_i z_i^*}{\sum_i z_i^2}\right) \\
I_i &= \left(\frac{z_i z_i^*}{m_2}\right) \\
m_2 &= \left(\frac{\sum_i z_i^2}{N}\right)
\end{aligned}
$$

LISA Map of County Homicide Rates, 1970



Let’s find out how to do this all in R!