Visualizing Labor Migration Using Quantitative Data

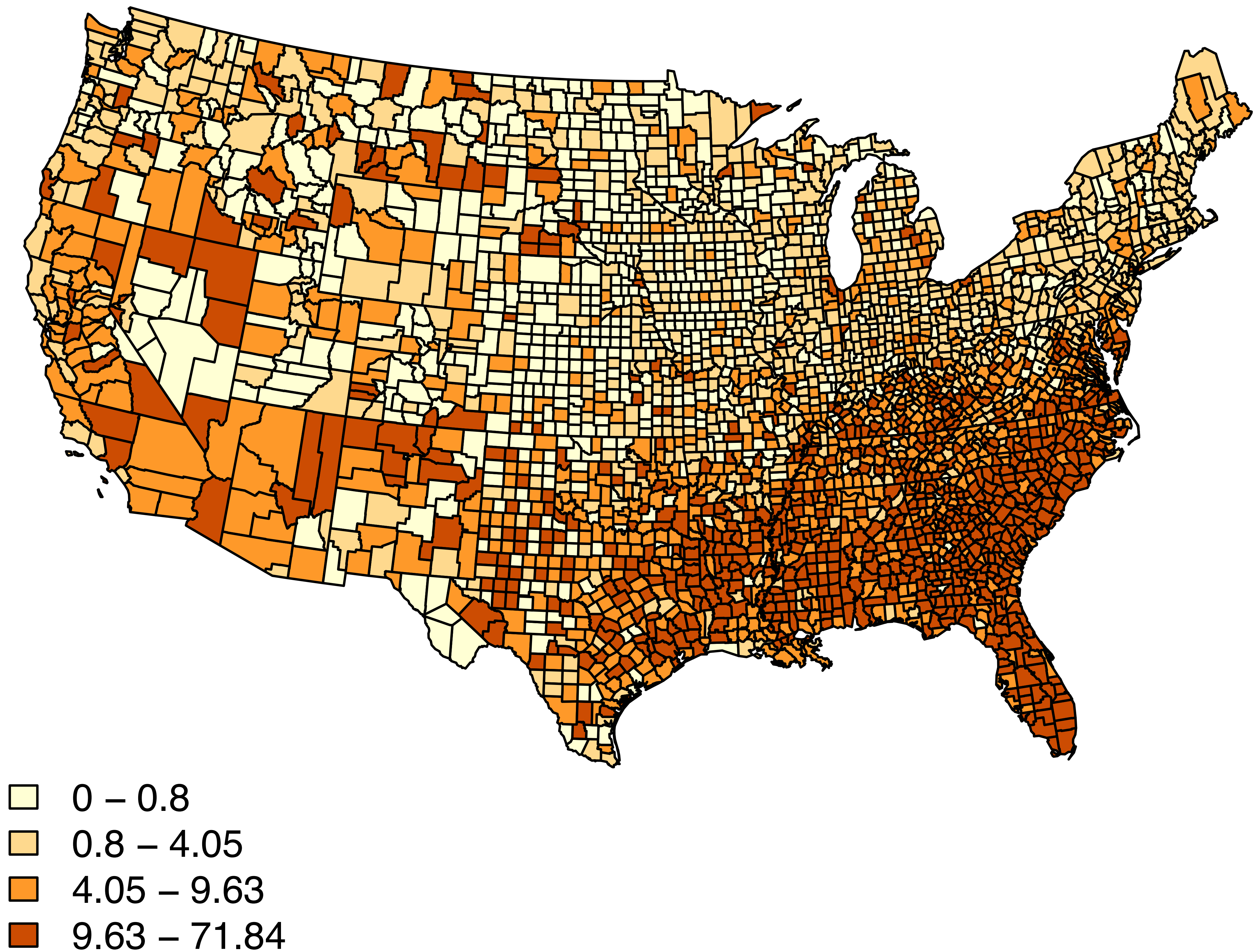
Spatial Regression

Outline

- Spatial processes and OLS diagnostics
- Spatial heterogeneity and spatial regimes
- Spatial regression models
  - Error vs. lag
  - Equilibrium estimates

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



How do we explain this particular arrangement of values?

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

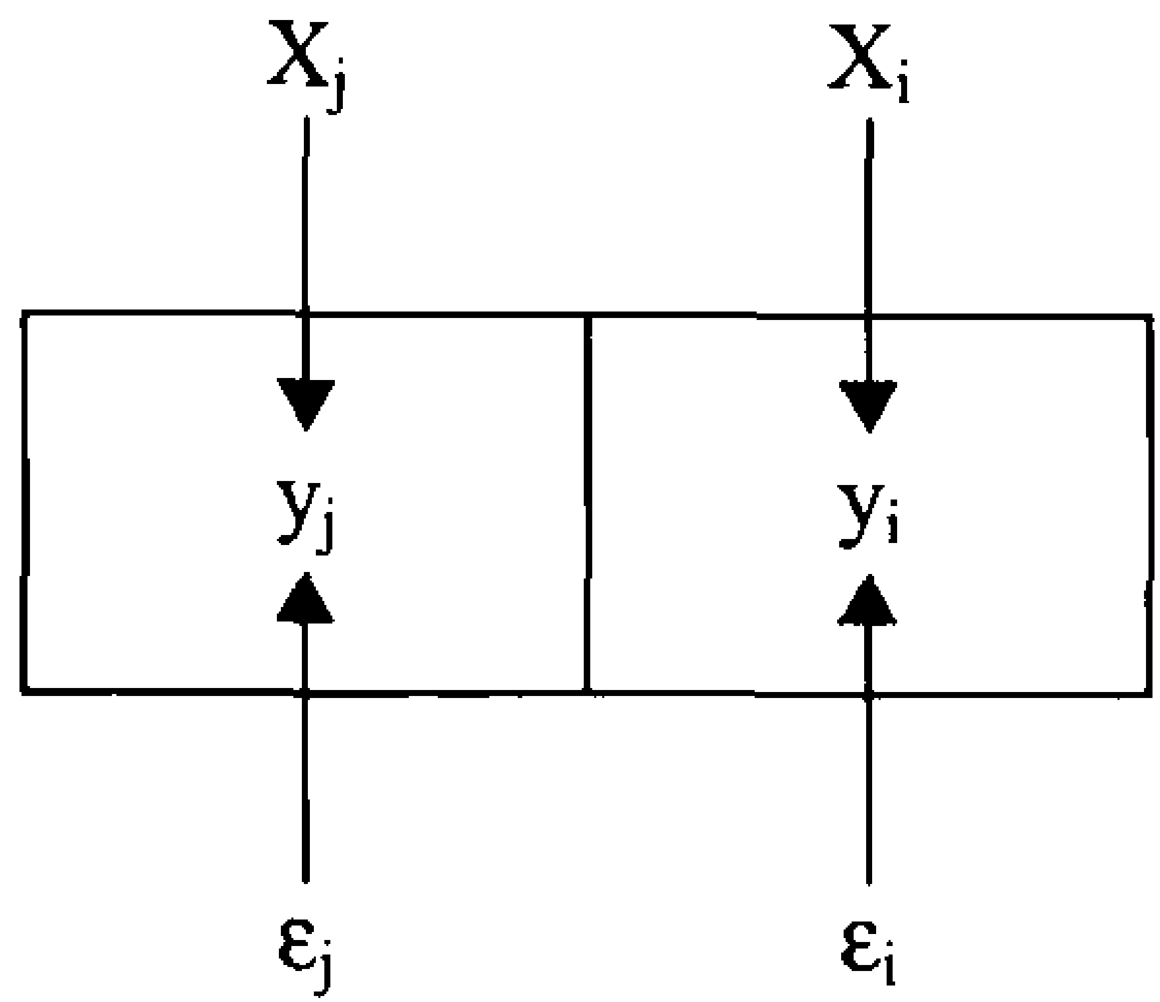
Spatial Processes

- 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…

Having looked at the degree of global and local autocorrelation in the outcome to figure out if there is a spatial problem to be solved, we want to start interrogating the suspects, starting with structural similarity.

Structural Similarity



$$y_i = \alpha + \beta x_i + \varepsilon_i$$

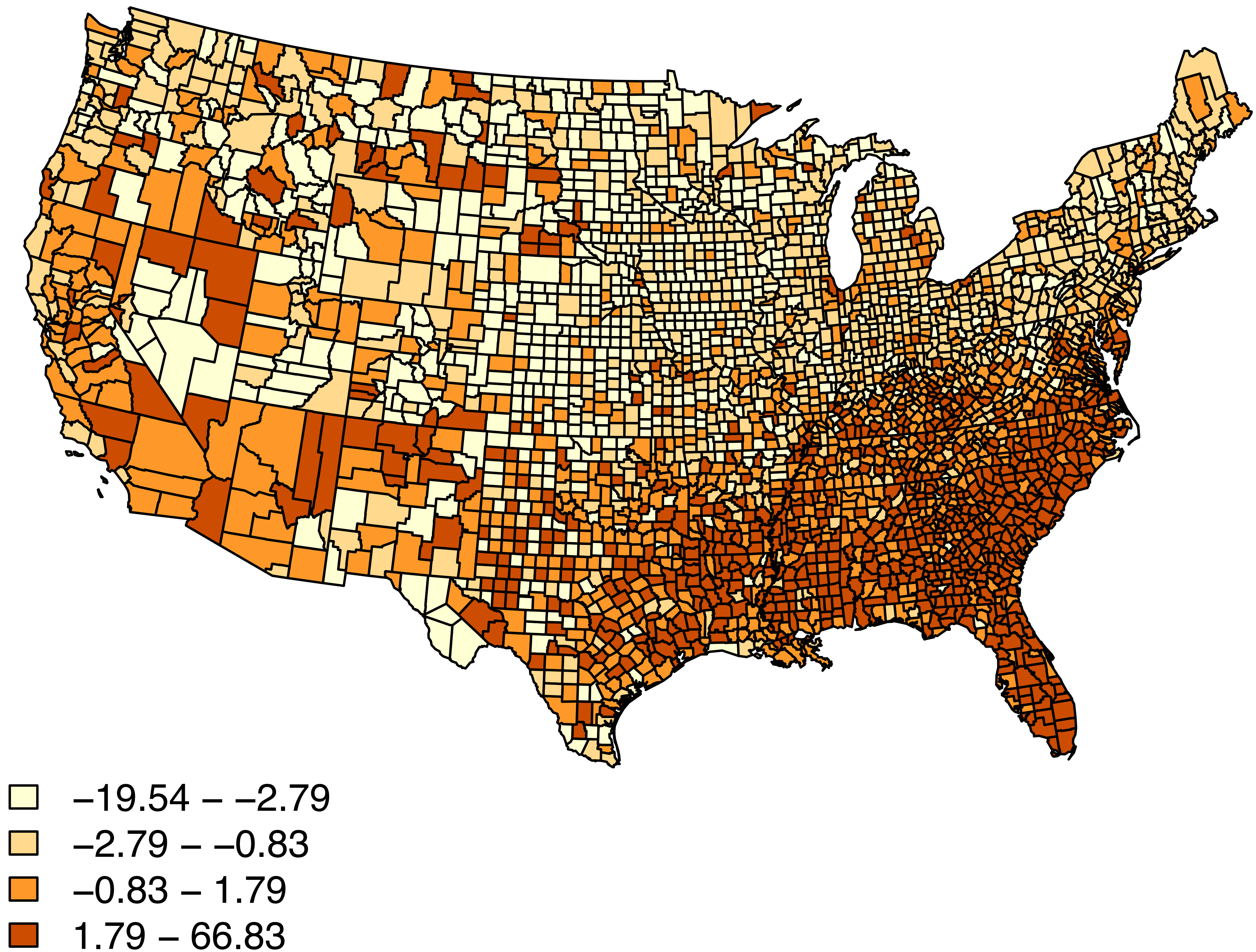
To do this, all we need is a regular regression.

Table 1. Ordinary Least-Squares Regression of County Homicide Rates 1960–1990[a]

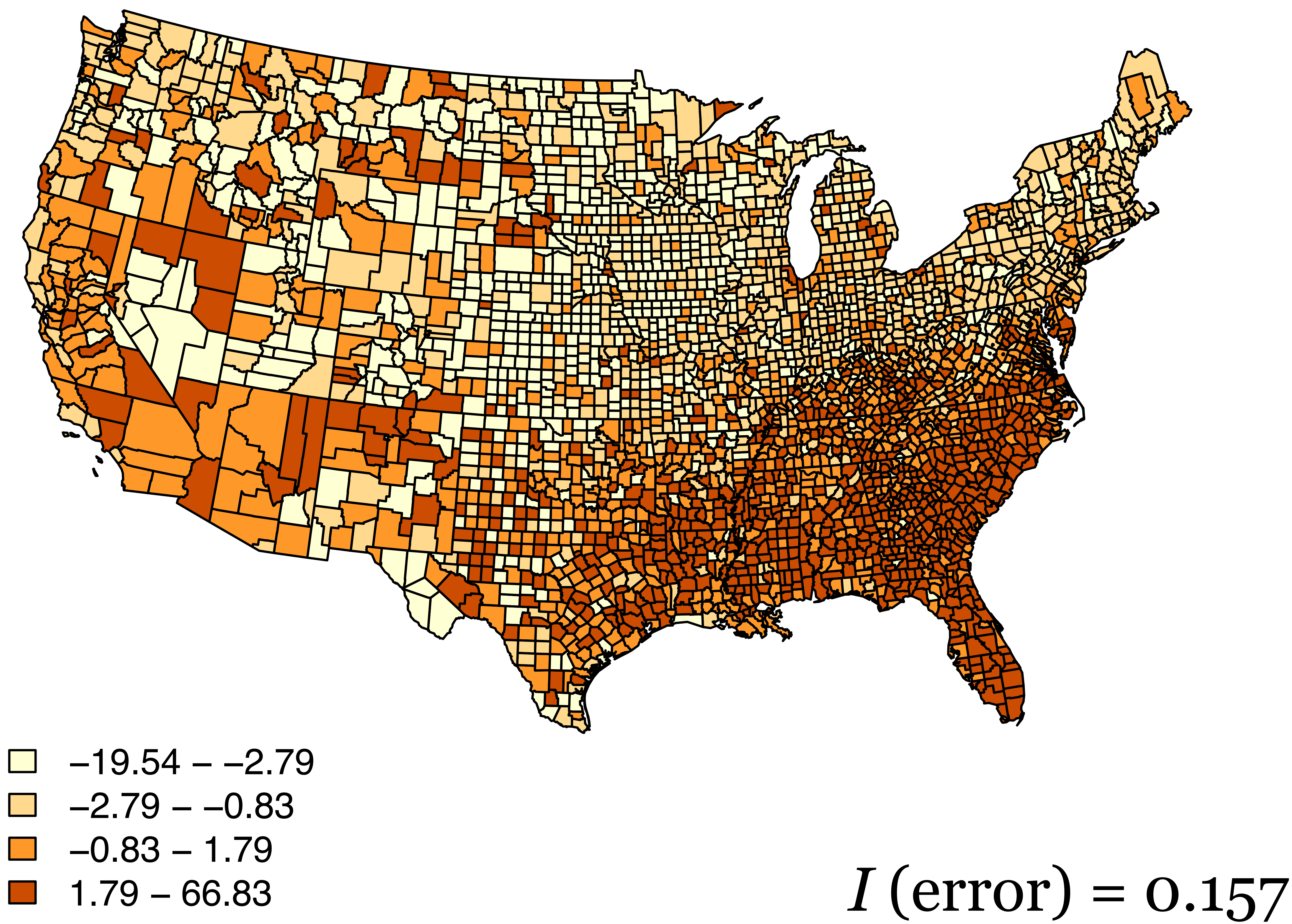
Independent Variables	1960	1970	1980	1990
Resource	1.798**	2.913**	3.412**	3.872**
Dep./Aff. Comp.	[0.318]	[0.396]	[0.500]	[0.583]
	(14.571)	(19.511)	(28.268)	(27.133)
Pop. Struct.	0.359**	0.812**	0.747**	1.353**
Comp.	[0.064]	[0.111]	[0.109]	[0.204]
	(3.892)	(6.959)	(7.315)	(13.491)
Median Age	−0.231**	−0.191**	−0.242**	−0.101**
	[−0.192]	[−0.130]	[−0.137]	[−0.055]
	(−11.931)	(−8.394)	(−9.671)	(−3.691)
Divorce	1.160**	1.264**	1.250**	0.583**
	[0.205]	[0.184]	[0.266]	[0.152]
	(12.233)	(12.109)	(18.586)	(10.690)
Unemployment	−0.062	−0.278**	−0.122**	−0.306**
	[−0.028]	[−0.087]	[−0.059]	[−0.141]
	(−1.762)	(−5.562)	(−3.965)	(−7.472)
South	2.639**	3.589**	2.113**	2.194**
	[0.233]	[0.243]	[0.154]	[0.165]
	(11.312)	(12.557)	(9.129)	(9.952)
Intercept	8.126**	8.653**	8.541**	6.517**
	(12.804)	(11.275)	(9.720)	(6.364)
Adj. R-Squared	0.295	0.360	0.431	0.435
N	3085	3085	3085	3085

[a] Unstandardized regression coefficients, standardized regression coefficients in brackets, and t-ratios in parentheses.
* $p \leq .05$; ** $p \leq .01$ (two-tailed tests).

OLS Regression Residuals, 1970



OLS Regression Residuals, 1970



In addition to testing for residual autocorrelation, we can also use the residuals to help us figure out what kind of spatial dependence we are dealing with.

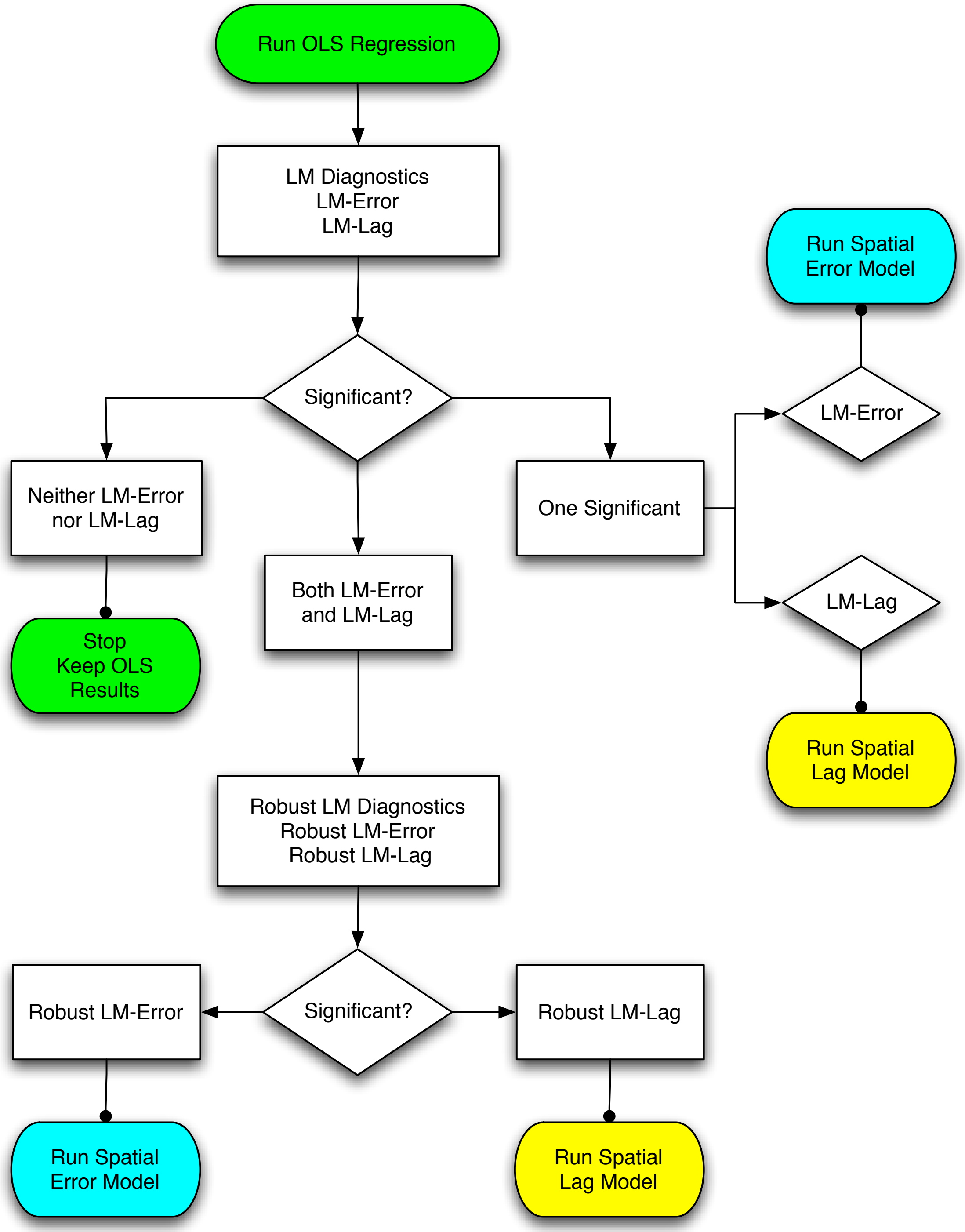


Figure 23.24: Spatial regression decision process.

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

Outline

- Spatial processes and OLS diagnostics
- Spatial heterogeneity and spatial regimes
- Spatial regression models
  - Error vs. lag
  - Equilibrium estimates

Before addressing the question of spatial dependence in more depth, we first need to examine the role played by our second suspect: spatial heterogeneity.

Spatial heterogeneity is a first-order process characterized by variation in the model parameters.

$$
\begin{aligned}
y_i &= \alpha + \beta x_i + \varepsilon_i \\
\varepsilon &\sim N(0, \sigma^2)
\end{aligned}
$$

$$
\begin{aligned}
y_i &= \alpha_i + \beta_i x_i + \varepsilon_i \\
\varepsilon &\sim N(0, \sigma_i^2)
\end{aligned}
$$

varying intercept

$$
\begin{aligned}
y_i &= \alpha_i + \beta_i x_i + \varepsilon_i \\
\varepsilon &\sim N(0, \sigma_i^2)
\end{aligned}
$$

varying intercept   varying slope

$$
\begin{aligned}
y_i &= \alpha_i + \beta_i x_i + \varepsilon_i \\
\varepsilon &\sim N(0, \sigma_i^2)
\end{aligned}
$$

varying intercept

varying slope

$$\begin{aligned} y_i &= \alpha_i + \beta_i x_i + \varepsilon_i \\ \varepsilon &\sim N(0, \sigma_i^2) \end{aligned}$$

heteroskedasticity

Spatial Effects: Heterogeneity

- Discrete
  - Spatial regime models
  - Multilevel models
- Continuous
  - Spatial expansion models
  - Geographically weighted regression (GWR)
  - Bayesian spatially varying coefficient models
  - Random effects eigenvector spatial filtering

Spatial Regimes

- Discrete heterogeneity
  - Model parameters vary across discretely defined spaces (i.e. groupwise variation)
    - e.g., North vs. South
- Detected and modeled in the usual way
  - Tests for parametric stability
    - including tests for heteroskedasticity
  - Group-specific models

Table 2. Stability of Regression Coefficients by Spatial Regime—County Homicide Rates 1960–1990

	1960	1970	1980	1990
I. Spatial Chow Test on Overall Stability[a]:				
	150.527**	227.468**	162.712**	168.438**
II. Stability of Individual Coefficients (non-South versus South)[b]:				
Resource Dep. Comp.	0.135	0.868	7.303**	36.065**
Pop. Struct Comp.	0.118	0.286	32.490**	18.758**
Median Age	3.480	0.036	7.352**	0.982
Divorce	0.057	11.088**	15.822**	0.641
Unemployment	24.849**	45.870**	12.922**	28.150**
III. Heteroscedastic Coefficients:				
Non-south	9.776	16.016	21.750	16.209
South	36.930	54.544	30.451	34.204
IV. Test on Heteroscedasticity[b]:				
	360.392**	328.375**	40.296**	164.284**
N (*N* of South)	3085 (1412)	3085 (1412)	3085 (1412)	3085 (1412)

[a] distributed as χ^2 with 6 degrees of freedom.
[b] distributed as χ^2 with 1 degree of freedom.
* $p \leq .05$; ** $p \leq .01$ (two-tailed tests).

When faced with evidence of spatial regimes, you should rerun your residual diagnostics based on the regime-specific models, allowing for the possibility that form of spatial dependence varies across regimes.

Table 3. Spatial Lag Models of Southern Homicide Rates 1960–1990[a]

Independent Variables	1960	1970	1980	1990
Resource	0.832**	1.792**	3.026**	4.028**
Dep./Aff. Comp.	[0.121]	[0.218]	[0.478]	[0.602]
	(3.386)	(5.820)	(13.994)	(14.814)
Pop. Struct.	−0.057	0.401	1.551**	1.747**
Comp.	[−0.007]	[0.041]	[0.198]	[0.209]
	(−0.265)	(1.497)	(7.637)	(8.247)
Median Age	−0.129**	−0.060	−0.150**	−0.018
	[−0.099]	[−0.039]	[−0.093]	[−0.009]
	(−2.942)	(−1.378)	(−3.736)	(−0.368)
Divorce	0.786**	0.642**	0.775**	0.482**
	[0.092]	[0.075]	[0.149]	[0.097]
	(3.241)	(3.060)	(6.302)	(4.251)
Unemployment	−0.070	−0.353**	−0.244**	−0.438**
	[−0.026]	[−0.092]	[−0.108]	[−0.191]
	(−0.897)	(−3.023)	(−4.145)	(−5.928)
Spatial Lag (ρ)	0.713**	0.651**	0.182*	0.230**
	[0.379]	[0.359]	[0.100]	[0.125]
	(6.005)	(6.905)	(2.431)	(3.261)
Intercept	4.108*	4.153*	9.101**	5.249*
	(2.207)	(2.042)	(5.364)	(2.513)
Sq. Corr.	0.178	0.239	0.311	0.333
N	1412	1412	1412	1412

[a] Unstandardized regression coefficients, standardized regression coefficients in brackets, and t-ratios in parentheses.
* $p \le .05$; ** $p \le .01$ (two-tailed tests).

Table 4. Spatial Regression Models of Non-Southern Homicide Rates, 1960–1990[a]

Independent Variables	1960	1970	1980	1990
Resource	1.571**	3.007**	4.143**	2.875**
Dep./Aff. Comp.	[0.275]	[0.389]	[0.467]	[0.405]
	(9.395)	(14.626)	(19.837)	(13.435)
Pop. Struct.	0.386**	0.859**	0.290*	0.962**
Comp.	[0.126]	[0.211]	[0.056]	[0.229]
	(5.011)	(7.795)	(2.132)	(8.299)
Median Age	−0.156**	−0.157**	−0.304**	−0.066*
	[−0.191]	[−0.163]	[−0.197]	[−0.050]
	(−7.336)	(−6.452)	(−8.607)	(−2.034)
Divorce	0.833**	1.403**	1.318**	0.572**
	[0.276]	[0.359]	[0.366]	[0.239]
	(8.552)	(13.980)	(14.560)	(9.156)
Unemployment	0.079**	−0.024	0.008	−0.045
	[0.061]	[−0.013]	[0.005]	[−0.029]
	(2.622)	(−0.502)	(0.196)	(−0.888)
Spatial Lag (ρ)	0.415**	NI	NI	NI
	[.197]			
	(4.645)			
Spatial Error (λ)	NI	0.243	0.329	0.268
Intercept	4.832**	6.164**	9.622**	3.261**
	(6.544)	(7.309)	(7.588)	(2.621)
Sq. Corr.	0.199	0.234	0.348	0.258
N	1673	1673	1673	1673

[a] Unstandardized regression coefficients, standardized regression coefficients in brackets, and t-ratios in parentheses.
* $p \le .05$; ** $p \le .01$ (two-tailed tests).

Outline

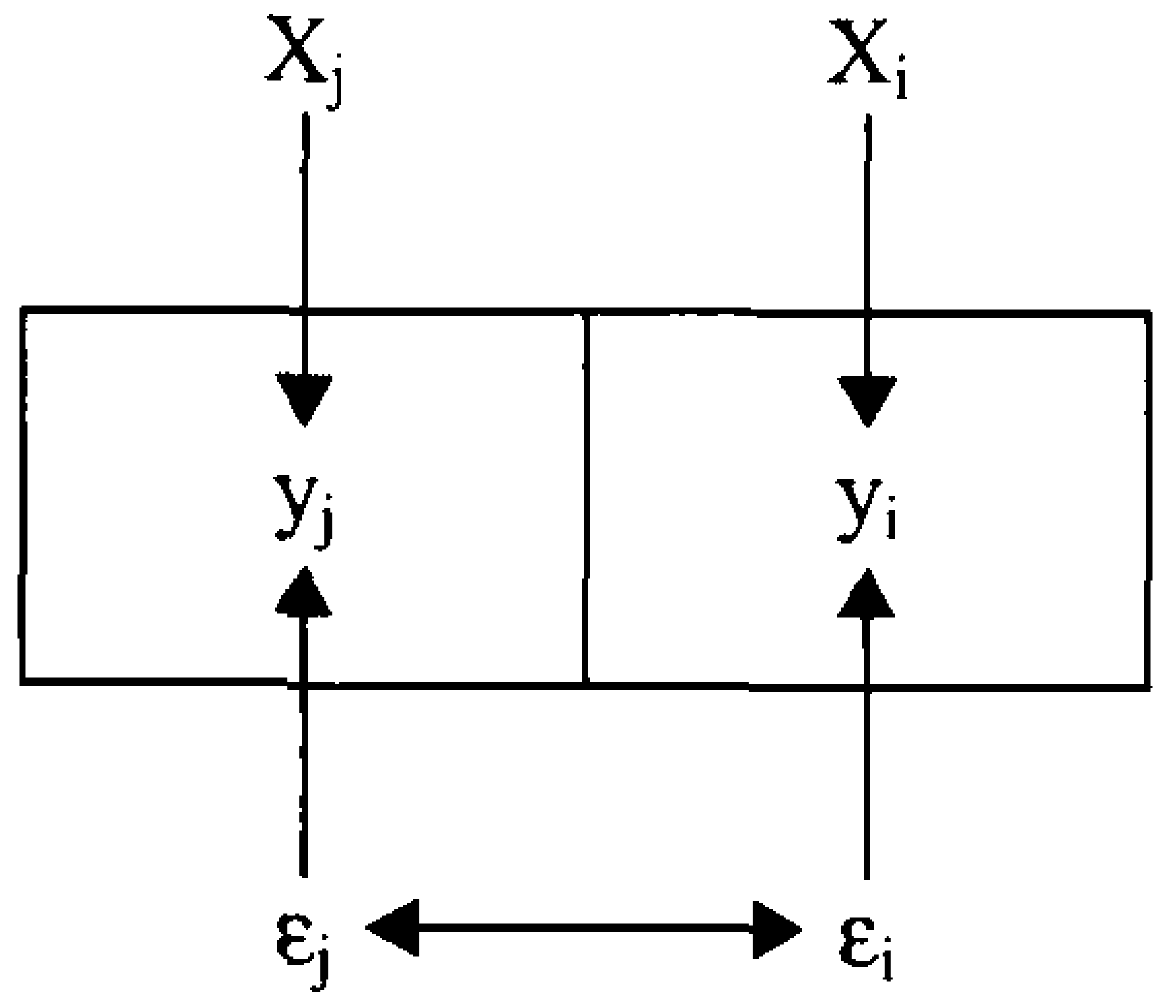
- Spatial processes and OLS diagnostics
- Spatial heterogeneity and spatial regimes
- Spatial regression models
  - Error vs. lag
  - Equilibrium estimates

Spatial dependence is a second-order process characterized by the interaction between observations.

Spatial Effects: Dependence

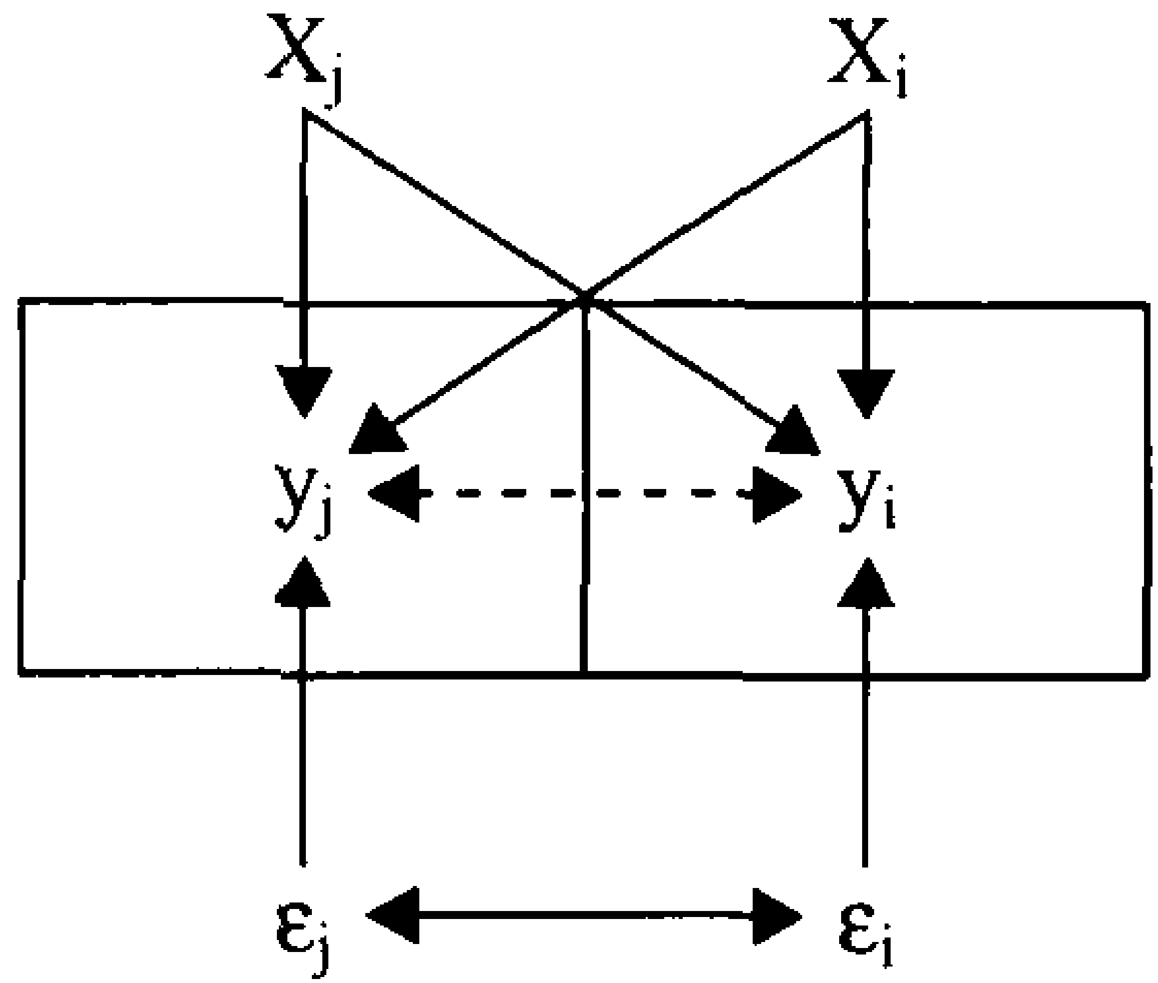
- Conditional autoregressive models
  - Implies local interdependence in the errors
- Simultaneous autoregressive models
  - Implies a fully interdependent system
    - Error vs. lag

Spatial Error
Effects



$$
\begin{aligned}
y_i &= \alpha + \beta x_i + \varepsilon_i \\
\varepsilon_i &= \lambda \sum_j w_{ij} \varepsilon_j + u_i
\end{aligned}
$$

Spatial Lag
Effects



$$y_i = \alpha + \rho \sum_j w_{ij} y_j + \beta x_i + \varepsilon_i$$

Estimating Spatial Effects Models

- Simultaneity bias undermines the use of OLS
- Two general alternatives
  - Maximum likelihood
  - GMM/IV
- Maximum likelihood is preferable
  - Computationally difficult with large matrices
  - Estimation routines vary in terms of how they compute the Jacobian

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

Unfortunately, the coefficients associated with the spatial lag model can't be interpreted in the usual way…

$$
\begin{aligned}
\mathbf{y} &= \rho \mathbf{W} \mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \varepsilon \\
(\mathbf{y} - \rho \mathbf{W} \mathbf{y}) &= \mathbf{X}\boldsymbol{\beta} + \varepsilon \\
\mathbf{y}\left(\mathbf{I} - \rho \mathbf{W}\right) &= \mathbf{X}\boldsymbol{\beta} + \varepsilon \\
\mathbf{y} &= \left(\mathbf{I} - \rho \mathbf{W}\right)^{-1} \mathbf{X}\boldsymbol{\beta} + \left(\mathbf{I} - \rho \mathbf{W}\right)^{-1} \varepsilon
\end{aligned}
$$

$$\left(\mathbf{I} - \rho\mathbf{W}\right)^{-1} \beta_k = \left(\mathbf{I} + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \rho^3\mathbf{W}^4 + \cdots + \rho^\infty\mathbf{W}^\infty\right) \beta_k$$

Impacts in the Spatial Lag Model

- A one-unit change in x_k produces an $n \times n$ network of effects
  - Average of diagonal entries represents the average direct impact
  - Average of row or column entries represents the average total impact
    - Rows are impacts to an observation
    - Columns are impacts from an observation
  - The average indirect impact is given by the difference between the average total and average direct impact
- We use simulation to produce standard errors and tests

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