Spatial Analysis for the Social Sciences

David Darmofal
University of South Carolina

Outline

Spatial Data

- All social science data are spatial data
- Behaviors, processes, and events occur at specific geographic locations
- Many of our theories incorporate a spatial dimension

Spatial Data

- Domino effect, waves of democratization, policy diffusion, contact vs. threat
- Too often we don't examine the spatial dimension of our theories in a rigorous manner
- Explosion of geocoded data, spatial methods, computational capability
- Time has never been better for modeling space in the social sciences

Galton's Problem

"[F]ull information should be given as to the degree in which the customs of the tribes and races which are compared together are independent. It might be, that some of the tribes had derived them from a common source, so that they were duplicate copies of the same original. . . . It would give a useful idea of the distribution of the several customs and of their relative prevalence in the world, if a map were so marked by shadings and colour as to present a picture of their geographical ranges."

Sir Francis Galton at The Royal Anthropological Institute, 1888

Spatial Dependence

Formally, spatial dependence is:

$$Cov(y_i, y_j) = E(y_i y_j) - E(y_i)E(y_j) \neq 0 \text{ for } i \neq j, \quad (1)$$

where the i, j locations have a spatial interpretation (Anselin and Bera 1998, 241-242).

Sources of Spatial Dependence

- There are three general classes of sources of spatial dependence:
- Spatial dependence may be produced by behavioral diffusion between "neighboring" units
- The behavior of unit i directly influences the behavior of unit j – and vice versa
- If this is the source of spatial dependence, we will wish to model it via a spatially lagged dependent variable

Sources of Spatial Dependence

- Alternatively, spatial dependence may be due simply to spatial dependence in the sources of behavior
- Thus, unit i and unit j may have no interaction with each other yet still exhibit spatial dependence
- If this is the source of spatial dependence, we will want to model it via covariates
- If the sources of spatial dependence are unknown or unmeasurable, this will produce spatial dependence in the errors
- We model this via spatially lagged errors

Sources of Spatial Dependence

- Finally, spatial dependence may be produced by spatial heterogeneity in the effects of covariates
- Covariates, for example, have a more positive effect on voter turnout in some locations, producing high (spatially autocorrelated) turnout in those locations
- Need to model this heterogeneity

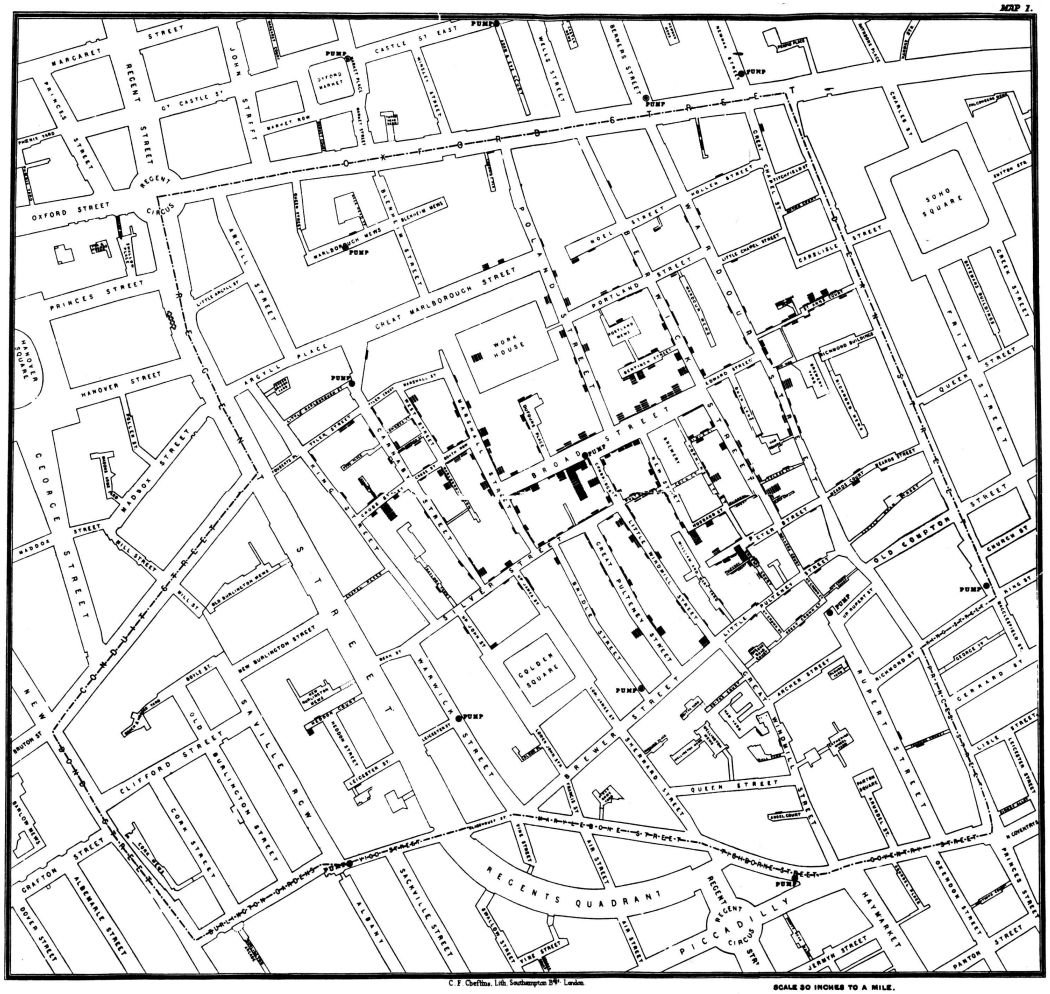
Geostatistical Data

- Geostatistical data are sample data from a continuous underlying surface
- Examples include fertility rates (Giulmoto and Rajan 2001), political campaign contributions (Cho and Gimpel 2007), and housing prices (Chica-Olmo 2007)
- Values are observed at sampled locations and unobserved at unsampled locations
- Researcher's interest is inferring information about values on the variable at unobserved locations from the sample data

Point Pattern Data

- In point pattern data, the observed spatial locations are the locations of discrete events
- For example, the locations of disease manifestation, civil wars, or terrorist acts
- Null often is complete spatial randomness (CSR)
- Interest is in whether there is spatial clustering of events (as in disease clustering) or spatial regularity (events are more dispersed than predicted under CSR)
- John Snow's analysis of the cholera epidemic was a primitive form of point pattern analysis that led the way for the modern discipline of epidemiology

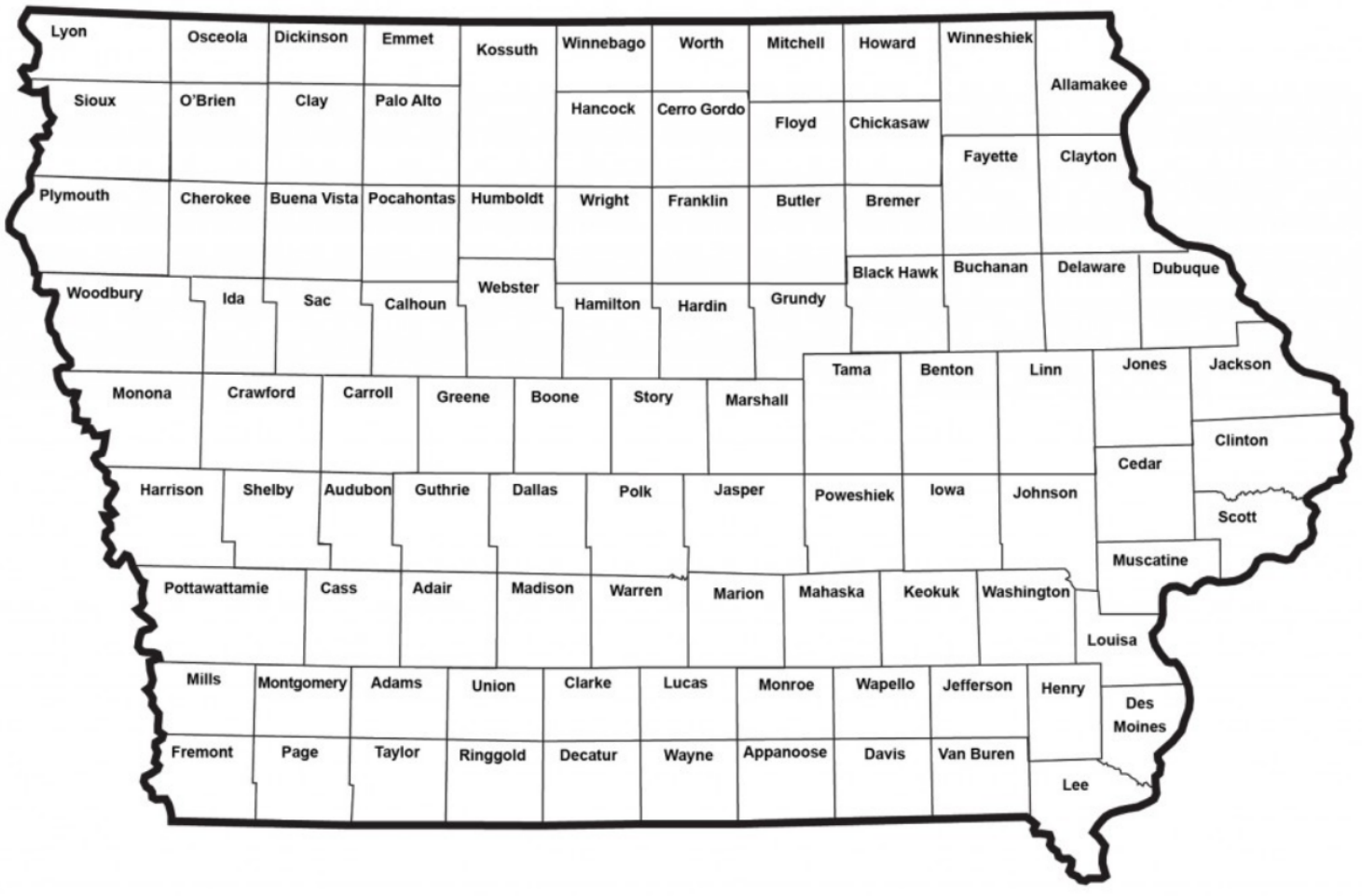
John Snow's Cholera Analysis



Lattice Data

- Most of the spatial data of interest to social scientists takes the form of areal, or lattice, data
- Lattice data: spatial plane is partitioned into a finite number of areal objects, or polygons
- Regular lattice data: grid of square or rectangular objects
- Irregular lattice data: block groups, legislative districts, states, countries
- Irregular lattice data are frequently encountered in the social sciences

Irregular Lattice Data



Parameterizing Spatial Dependence in Areal Data

- First step in modeling spatial dependence in areal data is diagnosing it in absence of covariates
- Do the data exhibit spatial autocorrelation?
- Or are the data spatially independent?

- There is insufficient information in cross-sectional areal data to estimate all of the separate covariances
- Therefore, we need to parameterize spatial dependence using a limited number of spatial parameters
- This constraint imposed via spatial neighbor definitions in a spatial weights matrix
- Only a unit's neighbors are allowed to exhibit first-order spatial dependence with unit i

Oracle of Bacon

- *Six Degrees of Kevin Bacon*

Spatial Weights Matrix

- Reflects neighbor definition
- n x n matrix, W
- $w_{ij} \neq 0$ if i and j are neighbors
- $w_{ij} = 0$ if i and j are not neighbors
- By convention, $w_{ii} = 0$
- Typically, the weights matrix is row-standardized, so that the sum of the values for the neighbors of i equal 1

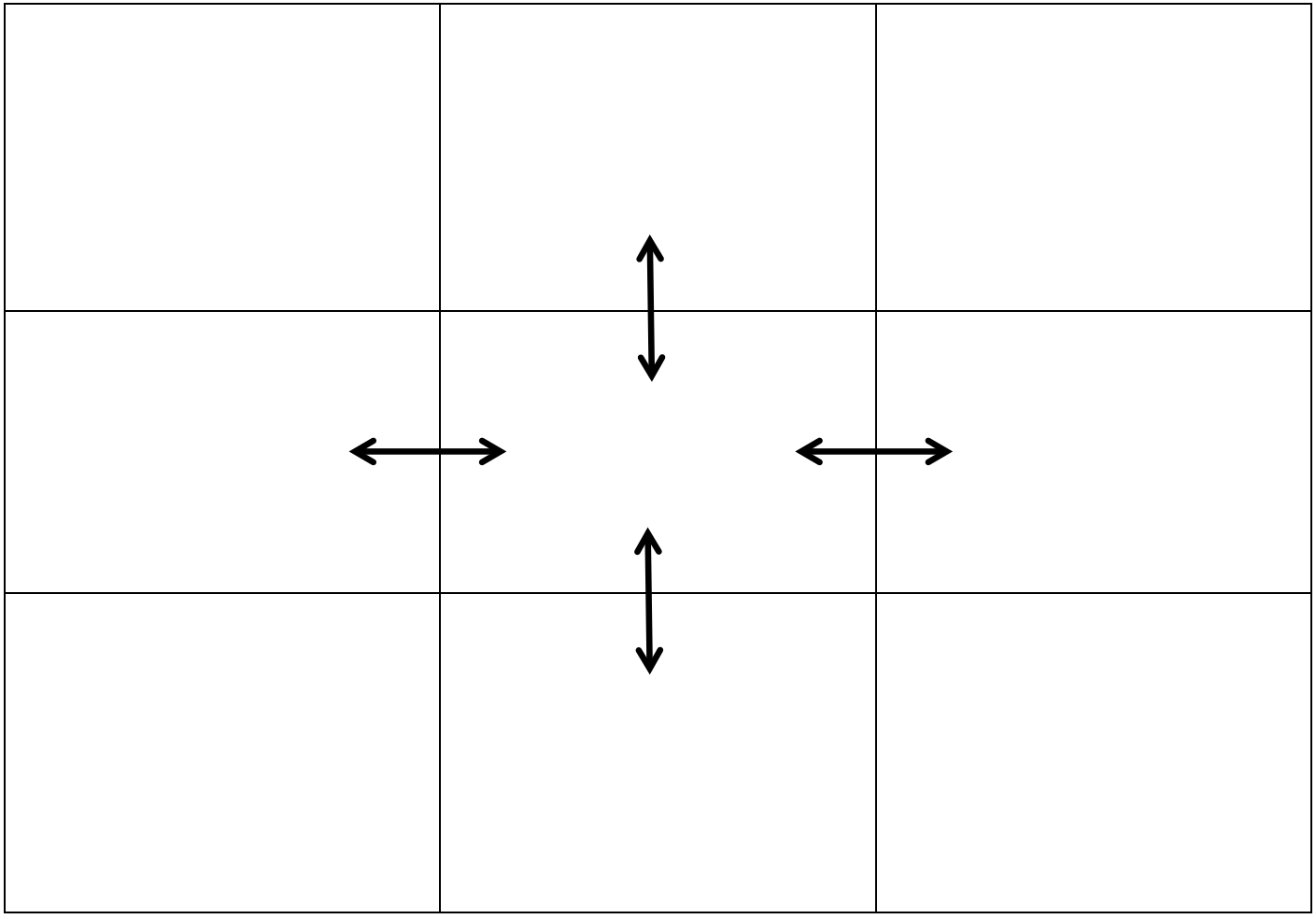
Neighbor Definitions

- There are many different ways that we can define neighbors (the non-zero cells) in a spatial weights matrix
- This definition of neighbors should be driven by theory:
- How do we believe spatial diffusion or spatial attributional dependence operates?

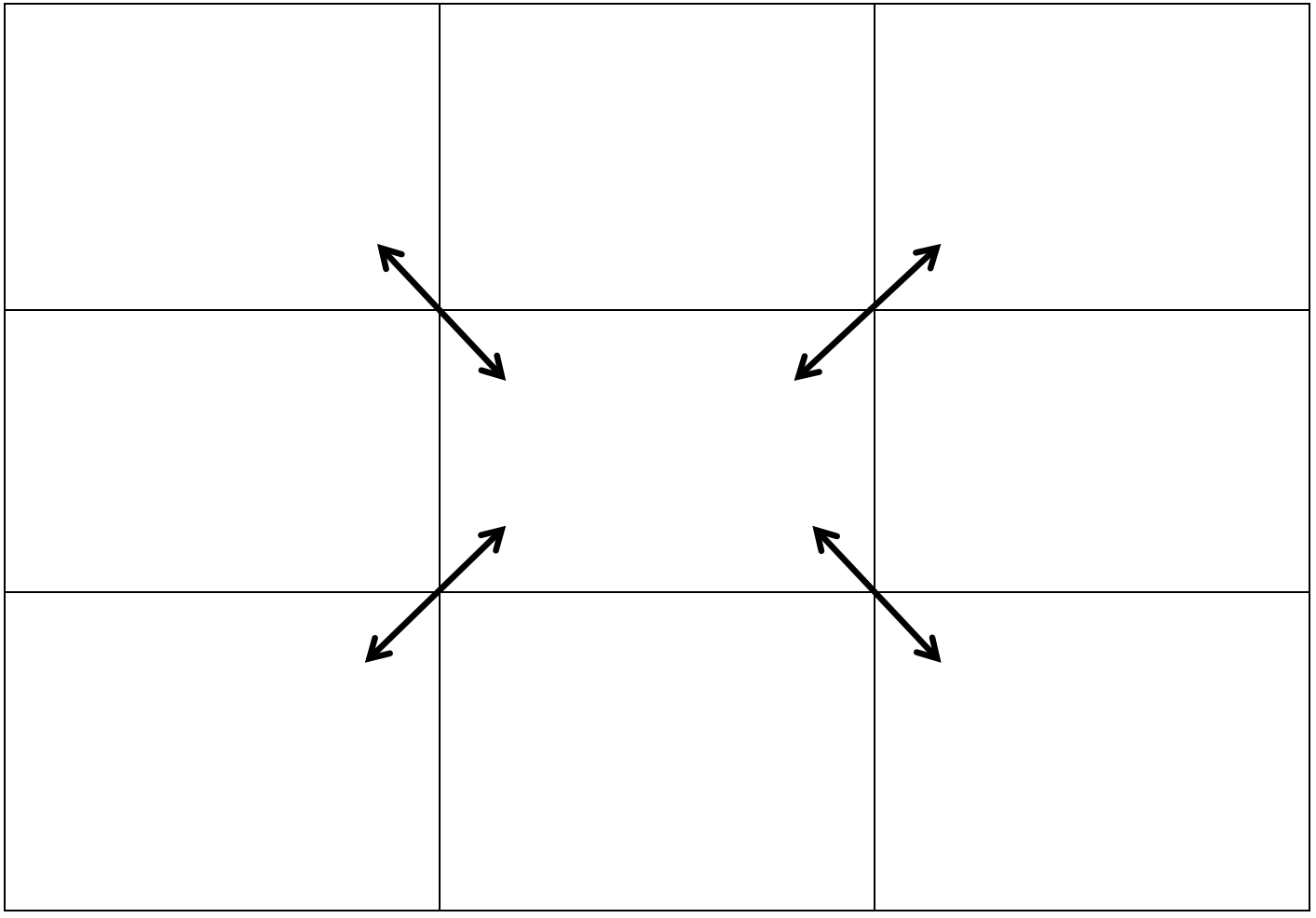
Neighbor Definitions

- Contiguous neighbors

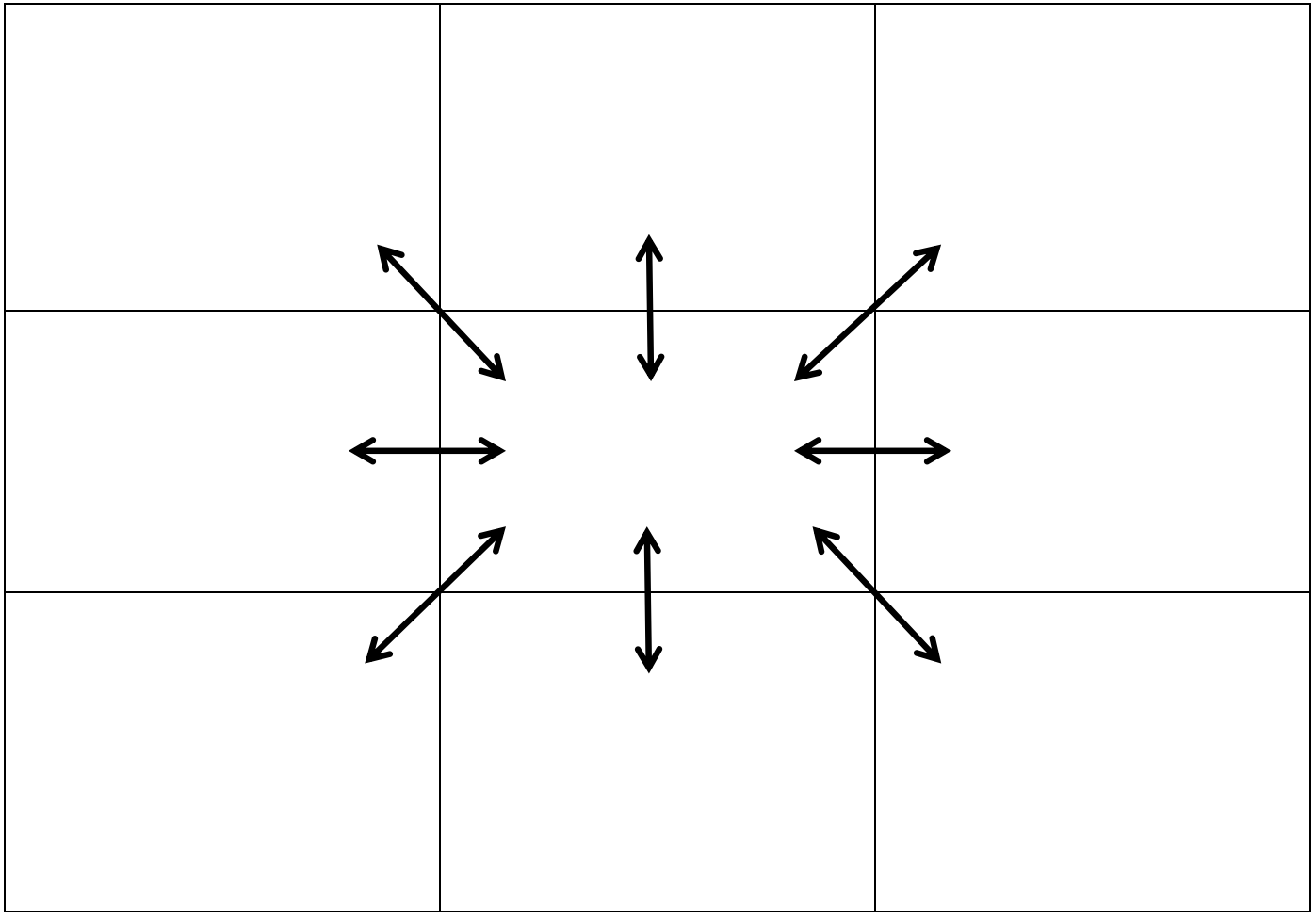
Rook Neighbor Definition



Bishop Neighbor Definition



Queen Neighbor Definition



Neighbor Definitions

- K-nearest neighbors
- Distance band neighbor definition
- Distance-decay neighbor definition
- Non-spatial neighbor definition (e.g., trade flows)

Diagnosing Spatial Dependence

- The first step in diagnosing and modeling spatial dependence is diagnosing it in the absence of covariates
- We can diagnose spatial dependence at either the global or local levels
- Global measures diagnose spatial dependence in the data as a whole
- Local measures diagnose which units exhibit spatial dependence with their neighbors

Global Moran's *I*

$$I = \frac{N}{S} \frac{\Sigma_i \Sigma_j w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\Sigma_i (y_i - \overline{y})^2} \tag{2}$$

where N is the number of observations, S is the sum of the weights, and $\overline{y}$ is the mean on y.

Global Geary's c

$$c = \frac{N-1}{2S} \frac{\Sigma_i \Sigma_j w_{ij}(y_i - y_j)^2}{\Sigma_i (y_i - \overline{y})^2}, \tag{3}$$

Local Indicators of Spatial Association

- Local indicators of spatial association (LISA) statistics satisfy two conditions:
- Diagnose which units exhibit spatial dependence with their neighbors
- The sum of LISAs for all observations is proportional to a corresponding global diagnostic
- LISAs can be used to identify which units are producing the global pattern and which run counter to it

Local Moran's *I*

$$I_i = \frac{\sum_j w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{(y_i - \overline{y})^2} \tag{4}$$

Local Geary's c

$$c_i = \frac{\sum_j w_{ij}(y_i - y_j)^2}{(y_i - \overline{y})^2} \tag{5}$$

Join Count Statistics

Dichotomous variable

$$BB = \frac{1}{2}\Sigma_i\Sigma_j w_{ij} y_i y_j, \tag{6}$$

$$BW = \frac{1}{2}\Sigma_i\Sigma_j w_{ij} (y_i y_j)^2, \tag{7}$$

$$WW = \frac{1}{2}\Sigma_i\Sigma_j w_{ij} (1 - y_i)(1 - y_j) \tag{8}$$

where $B = 1$ on the random variable, $W = 0$, w_{ij} is an element of the spatial weights matrix W, and y_i and y_j are the values on the variable at locations i and j

Inference on Spatial Diagnostics

- Assume normality
- Permutation: Randomly permute observed values across all locations

Spatial Models

- If spatial dependence is diagnosed, the next step is to model this dependence
- Two principal alternative models
- Behavioral diffusion: spatial lag model
- Shared attributes: spatial error model

Spatial Lag Dependence

- Spatial dependence bears surface similarities to temporal dependence
- Spatial diffusion is modeled via a lagged dependent variable, in a spatial lag model:

$$y = \rho Wy + \varepsilon$$

where y is an N by 1 vector of observations on the dependent variable, Wy is a spatially lagged dependent variable with spatial weights matrix W, ρ is the spatial autoregressive parameter for the spatially lagged dependent variable, and ε is an N by 1 vector of error terms.

Spatial Lag and Temporal Lag

- The spatial lag model bears some similarity to a time series model with a lagged dependent variable:

$$y_t = \rho y_{t-1} + \varepsilon_t, \tag{9}$$

- In the time series case, OLS is a biased but consistent estimator of ρ in the absence of serial correlation and other misspecification errors
- The OLS estimator is a consistent estimator only because y_{t-1} is uncorrelated with ε_t when there is no serial correlation in the errors
- This does not hold in the multidimensional spatial case

Spatial Lag and Temporal Lag

- Due to the simultaneity of spatial dependence, Wy_i is correlated not only with ε_i, but also with the errors at all other locations
- As a consequence, the OLS estimator $\hat{\rho}$ is inconsistent, regardless of whether or not there is dependence in the errors

$$\text{plim } N^{-1}(y'W'\varepsilon) = \text{plim } N^{-1}\varepsilon'W(I - \rho W)^{-1}\varepsilon, \qquad (10)$$

- Only when $\rho = 0$ does the probability limit equal zero

Omitted Spatial Lag

- If a diffusion process exists in the DGP and a spatially lagged dependent variable is omitted, the result is biased and inconsistent parameter estimates for the covariates in the model
- Omitted variable bias
- Need to estimate the spatial lag effect via maximum likelihood or instrumental variables

Monte Carlo Analysis

- Data generating process:

$$y = \rho W y + \beta_0 + \beta_1 x_1 + \varepsilon, \tag{11}$$

  where $\varepsilon \sim N(0, \sigma^2 I)$.
- x_1 is normally distributed with a mean of 0 and a standard deviation of 3, $\beta_0 = \beta_1 = 1$
- OLS estimates reflect a standard OLS specification that ignores the spatial lag dependence

Monte Carlo Analysis

- Ten values of ρ: -.9, -.7, -.5, -.3, -.1, .1, .3, .5, .7, .9
- Four different square lattice structures: a 5 by 5 lattice ($n = 25$), a 10 by 10 lattice ($n = 100$), a 20 by 20 lattice ($n = 400$), and a 30 by 30 lattice ($n = 900$)
- In each case, a queen contiguity definition of neighbors is employed
- For each combination of lattice size and ρ value, 1000 replications were performed

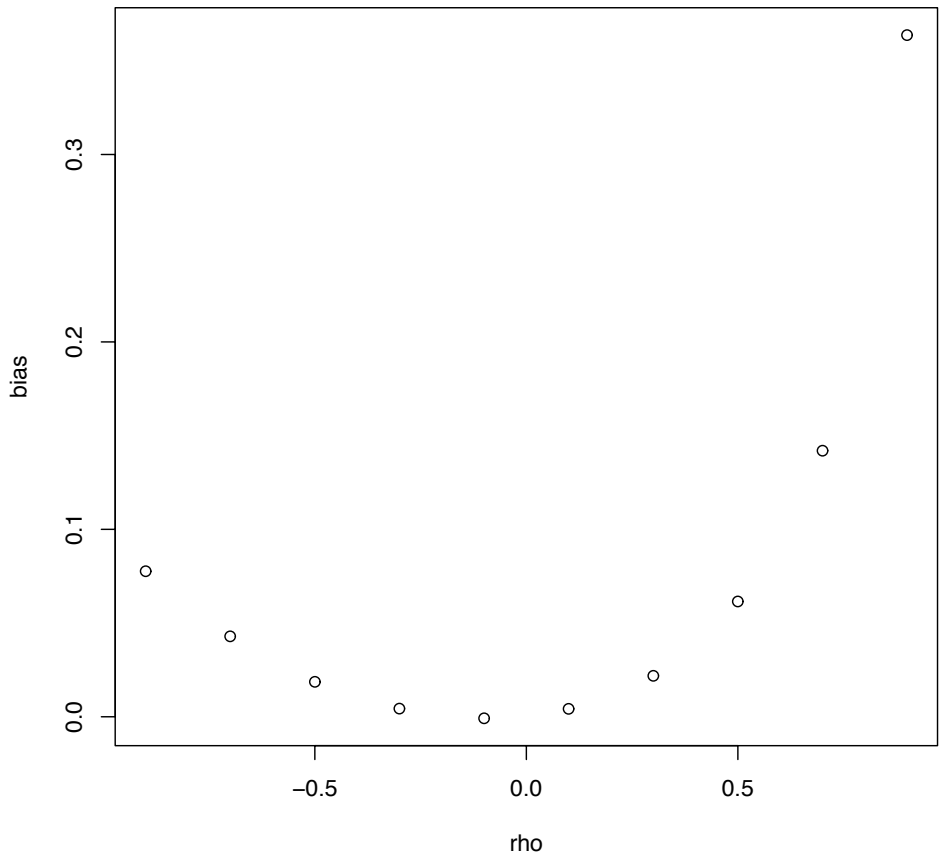
Monte Carlo Results

Table: Bias of the OLS Estimator with Omitted Spatially Lagged Dependent Variable

					ρ					
N	-.9	-.7	-.5	-.3	-.1	.1	.3	.5	.7	.9
25	.09	.05	.02	.01	.00	.01	.03	.07	.15	.35
100	.12	.07	.04	.01	.00	.00	.01	.04	.11	.29
400	.11	.06	.04	.02	.00	.00	.01	.04	.09	.25
900	.08	.04	.02	.00	.00	.00	.02	.06	.14	.36

Bias of the OLS Estimator with Omitted Spatially Lagged Dependent Variable

- $N = 900$



Spatial Error Model

- If diffusion is absent, shared attributes can be modeled via OLS
- If this attributional dependence is not modeled, produces spatial dependence in errors
- Spatial error model:

$$\varepsilon = \lambda W \varepsilon + \xi \tag{12}$$

  where λ is the spatial autoregressive parameter for the spatial error dependence and ξ is an i.i.d. error term.
- Simultaneous error dependence produces a non-zero covariance between the errors at all locations
- OLS estimates of λ are not consistent
- Employ maximum likelihood estimation

Omitting Spatial Error Dependence

- Consequences of ignoring spatial error dependence:
- OLS parameter estimates remain unbiased, but are no longer efficient
- Type I errors

Modeling Spatial Dependence

- If only attributional dependence exists, OLS may be employed
- If the shared attributes can be modeled
- Researchers should have theoretical reasons for expecting lag vs. error dependence
- However, researchers need to diagnose this to avoid model misspecification

Diagnosing Spatial Dependence in the Presence of Covariates

- The first step in modeling spatial dependence for a continuous DV is to estimate an OLS model
- Then examine diagnostics for spatial dependence in the presence of covariates
- Two types of diagnostics: focused and unfocused diagnostics

Unfocused Spatial Diagnostics

- Unfocused diagnostics do not have a clear alternative hypothesis of spatial lag or spatial error dependence
- The null is simply the absence of spatial dependence
- Moran's I diagnostic:

$$I = \frac{N}{S}\frac{e'We}{e'e}, \tag{13}$$

  where e are the residuals from an OLS regression.
- Similar to the Durbin-Watson statistic
- Kelejian-Robinson diagnostic

Focused Spatial Diagnostics

- In contrast to unfocused diagnostics, focused diagnostics have a clear alternative hypothesis of spatial lag or spatial error dependence
- Lagrange multiplier diagnostic for spatial lag dependence
- Lagrange multiplier diagnostic for spatial error dependence
- Problem: These Lagrange multiplier diagnostics are not robust against the presence of the alternative form of spatial dependence
- LM diagnostic for spatial error dependence may diagnose spatial dependence if only spatial lag dependence is present
- LM diagnostic for spatial lag dependence may diagnose spatial dependence if only spatial error dependence is present

Robust Lagrange Multiplier Diagnostics

- When the alternative form of spatial dependence is present, the LM diagnostics converge to a noncentral χ^2 distribution with an additional noncentrality parameter
- Bera and Yoon (1993) developed modified Lagrange multiplier tests
- Account for the noncentrality parameter and are robust to the presence of the alternative form of dependence
- Robust Lagrange multiplier diagnostic for spatial lag dependence
- Robust Lagrange multiplier diagnostic for spatial error dependence

Decision Rule for Spatial Models

- Estimate OLS model
- Estimate standard and robust Lagrange multiplier error and lag diagnostics
- If neither standard LM diagnostic is significant, keep OLS results
- If one of the standard LM diagnostics is significant and the other is not, estimate the appropriate spatial model
- If both standard LM diagnostics are significant, consult the robust LM diagnostics

Decision Rule for Spatial Models

- If one of the robust LM diagnostics is significant and the other is not, estimate the appropriate spatial model
- If both of the robust LM diagnostics are significant, estimate the appropriate spatial model as indicated by the larger statistic

Spatial Lag Model

- The mixed regressive, spatial autoregressive model modifies the pure spatial lag model to include a set of covariates and associated parameters:

$$y = \rho W y + X\beta + \varepsilon, \tag{14}$$

  where X is an N by K matrix of observations on the covariates, and β is a K by 1 vector of parameters
- Can employ maximum likelihood estimation or instrumental variables.

Spatial Error Model

- The regressive model with autoregressive error dependence:

$$y = X\beta + \varepsilon$$

$$\varepsilon = \lambda W\varepsilon + \xi, \tag{15}$$

- Employ maximum likelihood estimation

Two Issues in Spatial Modeling

- When modeling spatial data, scholars should be conscious of two particular issues
- Modifiable areal unit problem (MAUP)
- Boundary value problem

Modifiable Areal Unit Problem

- Spatial autocorrelation estimates depend on lattice units
- "Million or so correlation coefficients"
- Scale problem: dependence on the number of lattice units
- Aggregation problem: dependence on division into particular number of polygons
- Want theoretical match with chosen units

Boundary Value Problem

- Spatial dependence transcends the observed data
- Observed units are spatially autocorrelated with unobserved units
- Particularly likely to be observed at boundaries of the observed data (e.g., Kentucky counties influenced by Tennessee counties and vice versa)
- "Solutions" are suboptimal: e.g., empirical buffer zones, wrapping onto a torus

Data

- The data for *Spatial Analysis for the Social Sciences* are at: https://dataverse.harvard.edu/dataverse/SpatialAnalysis

Software

- R features extensive spatial modeling capabilities
- For cross-sectional models on areal data, spdep is the principal R package
- For panel data models on areal data, splm is the principal R package

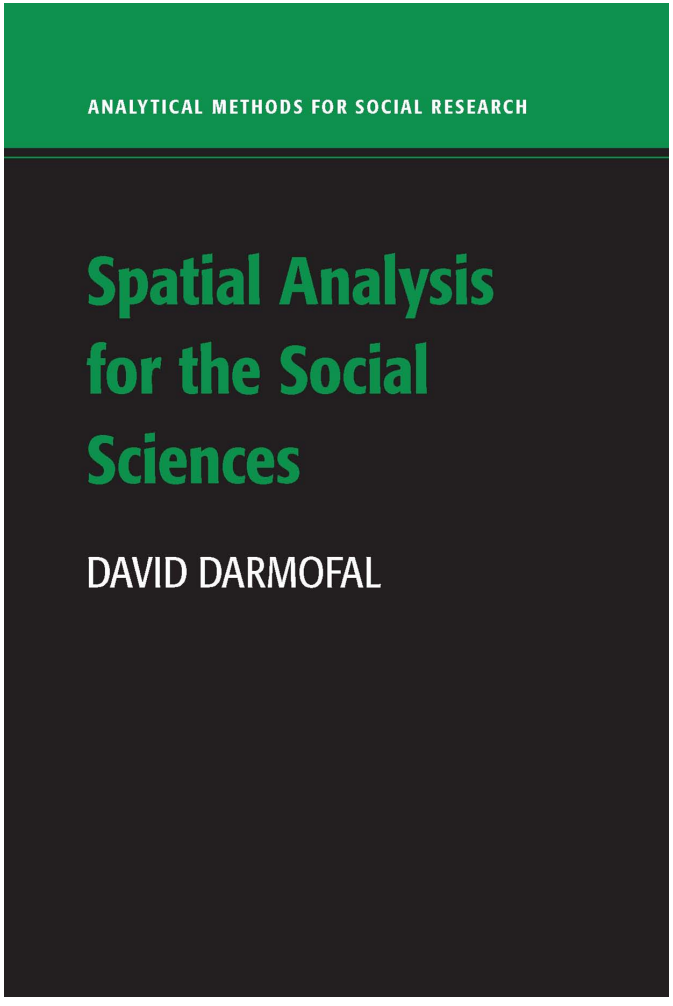
Software

- Stata has now included a suite of functions in Stata 15
- These can be found in the Sp manual

Software

- GeoDa is free open source software developed by Luc Anselin and his team
- Allows for data visualization, weights creation, and spatial modeling
- Available at https://geodacenter.github.io/

Thank You!


ANALYTICAL METHODS FOR SOCIAL RESEARCH
Spatial Analysis for the Social Sciences
DAVID DARMOFAL