

# Introduction to Spatial Regression Analysis

Paul Voss & Susan Ramsay

Day 2

Man U 2006

# Standard Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

In matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$(n \times 1)$   $(n \times k+1)$   $(k+1 \times 1)$   $(n \times 1)$

and where:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\hat{\sigma}^2 = \left[ \frac{1}{(n - k - 1)} \right] e^T e$$

# Violation of the OLS assumptions

- Non-linearity → biased estimates
- Mean dependence → biased estimates
- Heteroskedasticity → inefficient estimates (biased standard errors)
- Correlated disturbances → inefficient estimates
- Errors not normal → can't trust the  $p$ -values

# Violation of OLS assumptions: Uncorrelated disturbances

- Can be caused by:
  - Unmeasured variables common to a subset of observations
  - Spillover process; Diffusion process
- Test for violations:
  - Test visually
  - Formal tests:
    - Durbin-Watson
    - Moran
    - Geary
    - Others
- Consequences of violating:
  - Inefficiency (often)
  - Biased & inconsistent parameter estimates (if arising from spillover of dependent variable)

This brings us to the topic of

# Spatial Autocorrelation

Questions?

# Outline for today

- Sampling perspectives
- Global spatial autocorrelation & weights matrices
  - Moran's  $I$
  - Geary's  $c$
- Spatial dependence
- Spatial heterogeneity
- Introduction to ESDA
- Lab: Spatial autocorrelation in Geoda

# Global Spatial Autocorrelation

# To begin, some definitions...

- **Autocorrelation**: Pairwise correlation of *univariate* observations (values)
- “**Correlation**” retains its classical meaning of association
- “**Auto**” means “self”
- “**Spatial**” describes the manner in which the self-correlation arises
- “**Global**” means we are calculating the autocorrelation measure across the entire domain (all observations)

In other words...

Global spatial autocorrelation  
arises from the configurational  
arrangement of observations  
(attributes) in space

(with space commonly referring to  
a two-dimensional planar surface,  
i.e., a map)

# Spatial autocorrelation has been variously interpreted as:

(Acknowledgement to Daniel A. Griffith, 1993)

- Self-correlation attributable to the geographical ordering of the data
- A descriptor of the nature and degree of certain types of map pattern
- An index of information content latent in geographical data
- A diagnostic tool for spatial model misspecification
- A surrogate for unobserved/latent geographic variables
- A nuisance in applying conventional statistical methodology to spatial data
- An indicator of the appropriateness of, and possibly an artifact of, areal unit demarcation
- A spatial spillover effect
- A spatial process mechanism

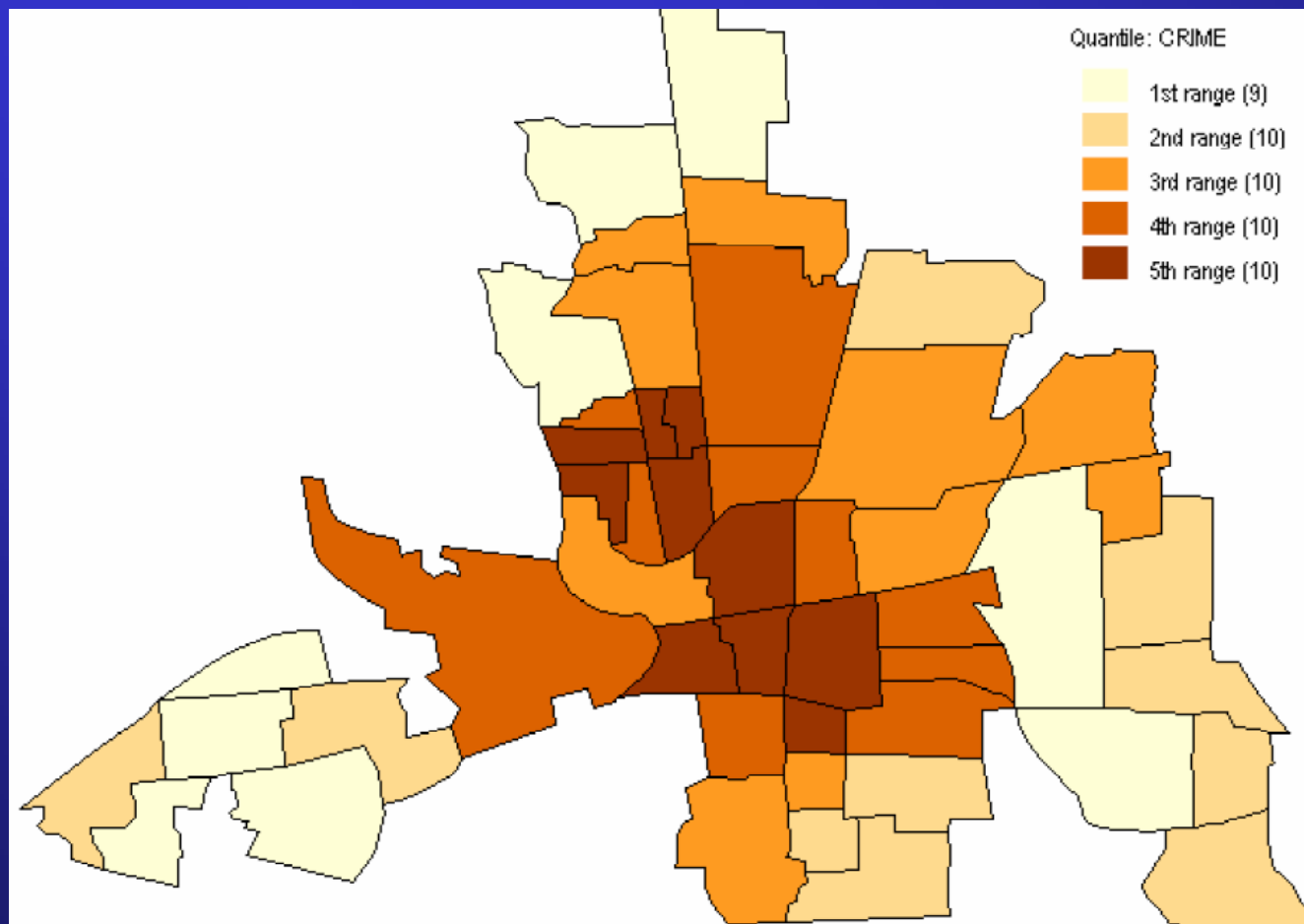
# Some summary observations

- These perspectives are not competing schools of thought. All are correct
- One or more of the individual perspectives will assume importance (over and above the others) depending on the particular analytical problem at hand
- Spatial autocorrelation concerns data values not data locations. The locations of the observations are taken as given.
- We are interested in if/how value similarity follows locational similarity.

# So, where do we go from here?

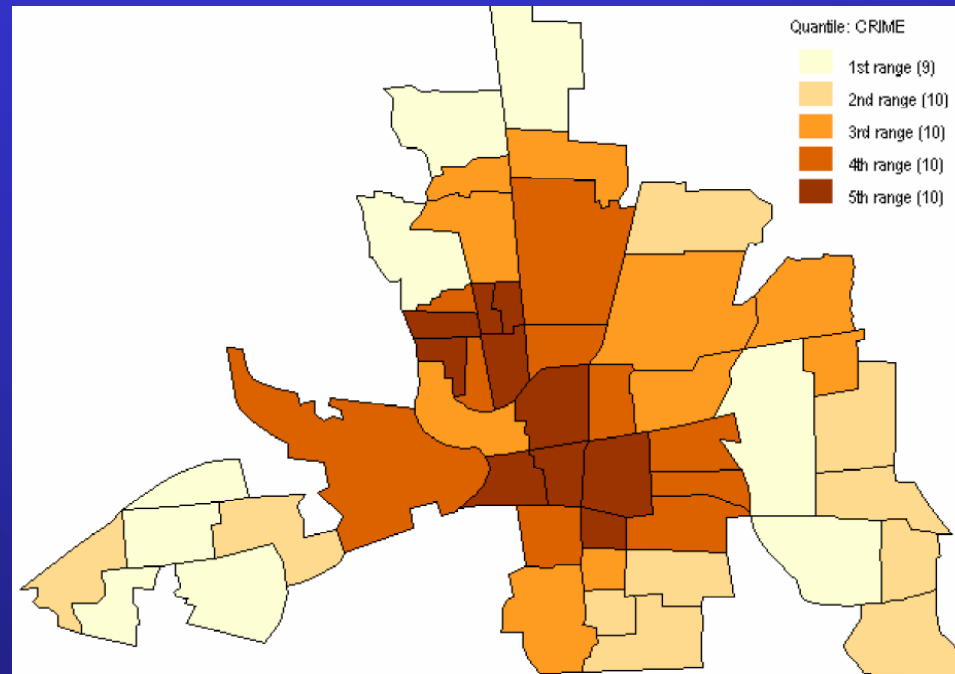
- How do we develop a means to (statistically) differentiate among different kinds of maps?
- That is, can we *quantify* different kinds of map patterns?
- And once we develop a statistic for describing (quantifying) different kinds of map patterns, can we derive the sampling distribution for this statistic and thus also quantify the expected value and standard error for the statistic?

Suppose we observe the following map of crime rates (in quintiles) from 49 census tracts in Columbus, Ohio



# The question for us then is this:

If the set of crime rates  $\{z_i\}$  had been allocated in a random fashion to the set of 49 census tracts, would this observed spatial distribution be a likely outcome of such an allocation procedure?



But what does this question mean? Does it make sense?

# What does it mean to ask about crime rates as if these observations are the result from some type of sampling experiment?

- The data are... well, the data. Right?
- We've got all the tracts, not just a sample of them.
- The data for each tract (crime rates) are based on complete count data, not on a sample.
- So what can it possibly mean to ask whether these rates are allocated in a random fashion (or not)?
- We have the "universe of observations" not some sample. They are what they are. Where's the sample?

There are a number of formal  
perspectives on this topic

and a terrific quote...

*“[Our data often render] the idea that one is working with a (spatial) sample somewhat remote. Great imagination has gone into turning what appears to be a population into a sample, thereby making statistical theory relevant...”*

Graham J. G. Upton and Bernard Fingleton  
*Spatial Data Analysis by Example, Vol. I*  
(Wiley & Sons, 1985:325)

# Sampling Perspectives

Generally there are four spatial sampling perspectives discussed in the literature based on the sampling design:

with replacement?                      Yes    No

order Important?                      Yes    No

# We'll focus on two of these sampling perspectives

Sampling with replacement, order is important  
("free sampling"): "Normalization Perspective"

Sampling without replacement, order is important  
("nonfree sampling"): "Randomization Perspective"

Consider this simple “map” with four areas for which we have attribute values A, B, C, & D

A	B
C	D

# Nonfree Sampling Approach: Without Replacement Order is Important

The number of possible spatial distributions is  $n! / \prod_{i=1}^{i=k} (n_i!)$ , where  $k$  denotes the number of distinct values of distinct values

For example, consider the set of values  $\{0,2,3,3\}$  for our “map.”

The possible number of different spatial samples is given by:

$$4! / [(1!)(1!)(2!)] = 12$$

Under the map set  $\{0,2,3,3\}$ , the sample space would be:

0	2
3	3

0	3
2	3

0	3
3	2

2	0
3	3

2	3
0	3

2	3
3	0

3	3
2	0

3	0
2	3

3	0
3	2

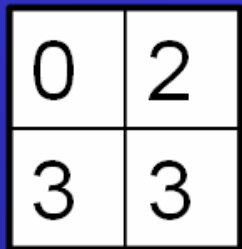
3	3
0	2

3	2
0	3

3	2
3	0

# Nonfree sampling is relatively easy to understand

Let's say the  
map you observe  
is this one



0	2
3	3

# Nonfree sampling (cont.)

Let's say the map you observe is this one.



0	2
3	3

0	3
2	3

0	3
3	2

2	0
3	3

2	3
0	3

2	3
3	0

3	3
2	0

3	0
2	3

3	0
3	2

3	3
0	2

3	2
0	3

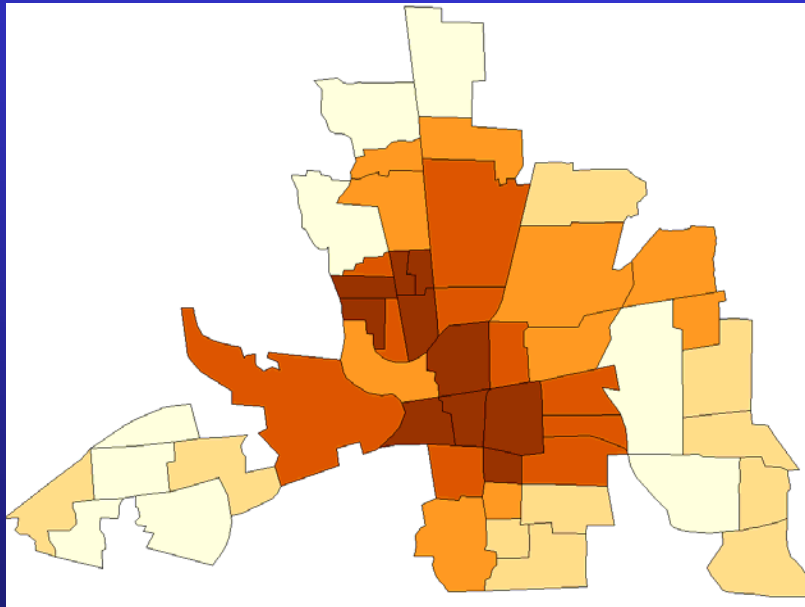
3	2
3	0

Under nonfree sampling this outcome is but one of twelve possible outcomes.

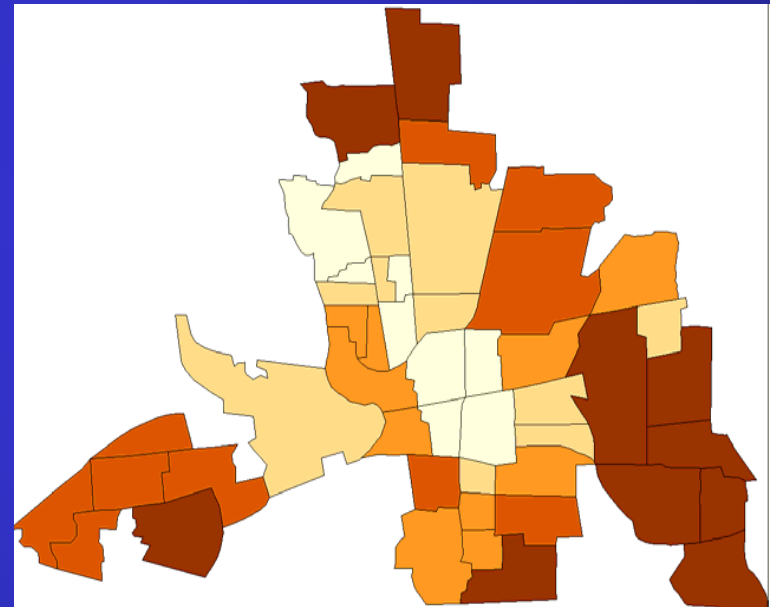
We now have one observed outcome from the sampling distribution

# Back to the Columbus Crime Data

This was our  
“observation”



But here is another  
under nonfree sampling



The question becomes: How unusual is map 1 given the  $49!$  (approx.  $6 \times 10^{62}$ ) possible permutations of these results under nonfree sampling?

But even if you subscribe to  
this approach...

To operationalize this, all (or many of) the  
different arrangements that are possible  
need to be identified in order to construct  
(approximate) the appropriate sampling  
distribution for our statistic

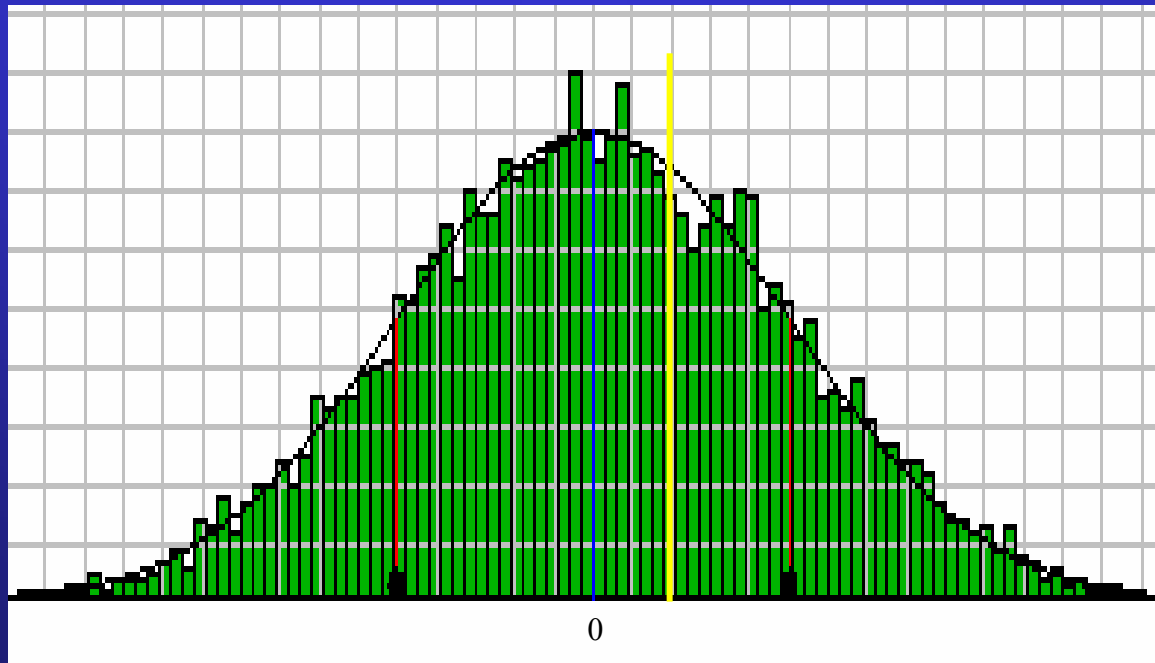
When the sampling distribution is simply  
too large to construct, we can estimate it.  
Some statisticians prefer to call such an  
approximate sampling distribution a  
“reference distribution”

# So... again (for Columbus):

- There are  $n = 49$  different values of the crime rate (i.e.,  $\{z_i\}$ ,  $i = 1, \dots, n$ ) and  $n = 49$  tracts
- Once we allocate a value to a tract, it is not available for allocation to any remaining tract
- To begin, there are  $n$  choices for the first tract
- That leaves  $(n-1)$  choices for the second tract
- $(n-2)$  choices for the third tract, and so on
- Since there are  $n$  possibilities for the first choice and  $(n-1)$  possibilities for the second choice, then there are  $n(n-1)$  possibilities for the first and second choices together
- Extending this counting principle for all  $n$  tracts results in  $n!$  (i.e.,  $49!$ ) possible maps in the sampling distribution
- Note: this counting principle assumes that the set  $\{z_i\}$  consists of  $n$  distinct values, and are the only values available to us in the assignment process
- Analogous to sampling without replacement; order important
- Again... sometimes called nonfree sampling or the randomization sampling perspective

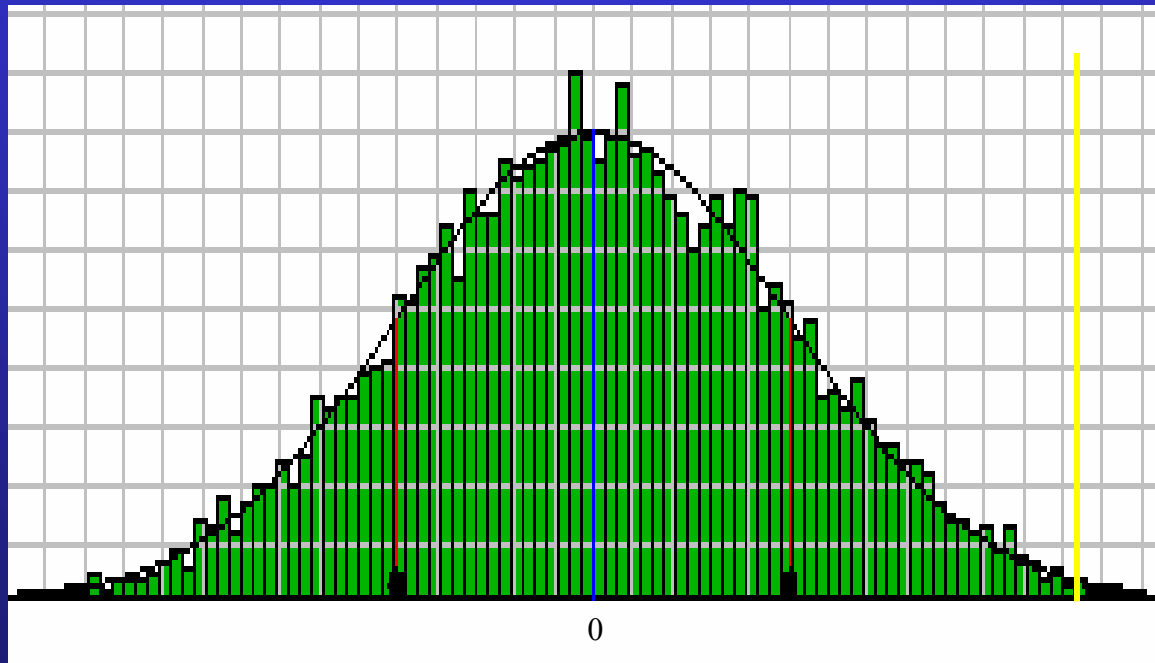
# Thus...

If we create  $49!$  maps (or a large sample from this huge number) and derive our spatial autocorrelation statistic for each of these maps, then we have a reference distribution against which to weigh the one we actually observed. It might look something like this:



# And...

If our observed spatial autocorrelation statistic lies in a tail of the sampling distribution, then we would have a statistical basis for arguing that the observed spatial distribution of crime rates in Columbus probably do not come from a random allocation process

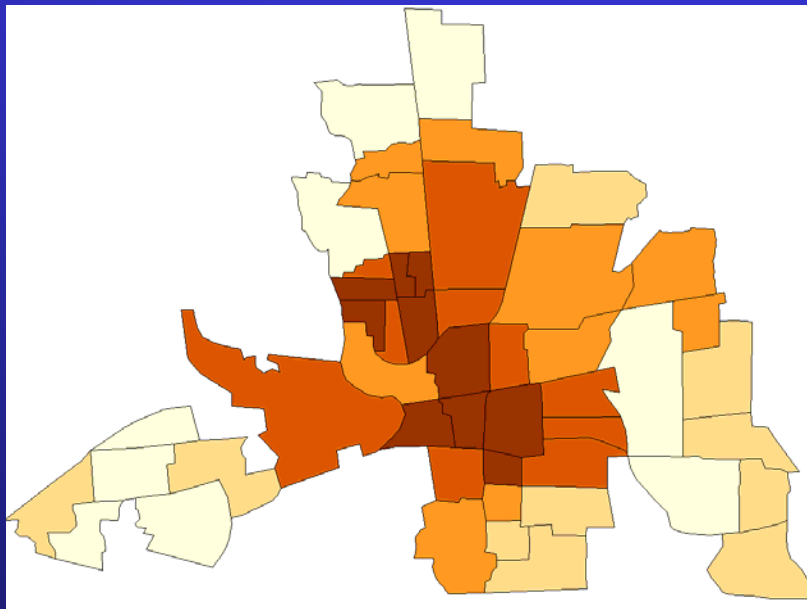


# The other common sampling perspective in spatial data analysis is that of free sampling

- With replacement; order is important
- Consider the map pattern as the joint realization of  $n$  independent random draws  $z_i$  from some multivariate joint distribution  $Z_i$  (“Superpopulation”)
- To proceed, the theoretical distribution (the probability density function) of  $Z_i$  must be known
- Assume that the observations each represent an independent draw from a multivariate normal distribution (Gaussian Random Field)
  - because of the many advantageous features of the multivariate normal distribution, this is handy
  - also justified under the Central Limit Theorem

Under the normalization perspective  
(or free sampling perspective)

This was our  
“observation”



But there are almost  
an infinite number of  
different maps that  
might be observed  
under a free  
sampling perspective

The question becomes: How unusual is our map given the number of maps that might have been observed under an assumption of free sampling?

# Questions about this?

This will only become important when we have our spatial autocorrelation measures and wish to know if they are large enough to reject the null hypothesis of no spatial autocorrelation

Some software packages ask you whether you wish to compute the standard error of your spatial autocorrelation statistic from a perspective of free sampling or non-free sampling

Now, back to the matter of  
**spatial autocorrelation**

# Global Spatial Autocorrelation in Two Dimensions, Common Measures:

Moran's  $I$

Geary's  $c$

But... before we can proceed,  
one more interruption.  
we need to understand what we  
mean by a...

- Spatial “neighbor”
- Spatial “neighborhood”
- “Proximity matrix”
- “Contiguity matrix”
- “Weights matrix”
- “Spatial lag operator”

# Consider this lattice

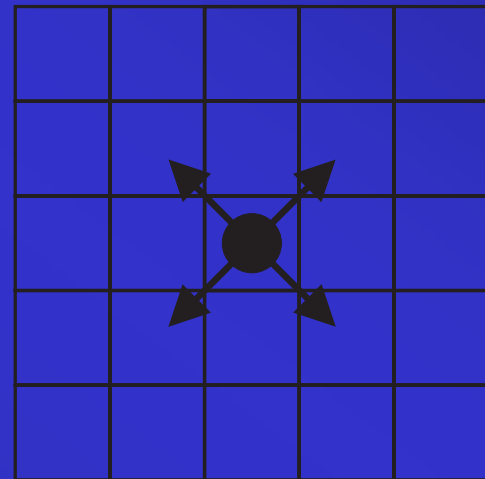
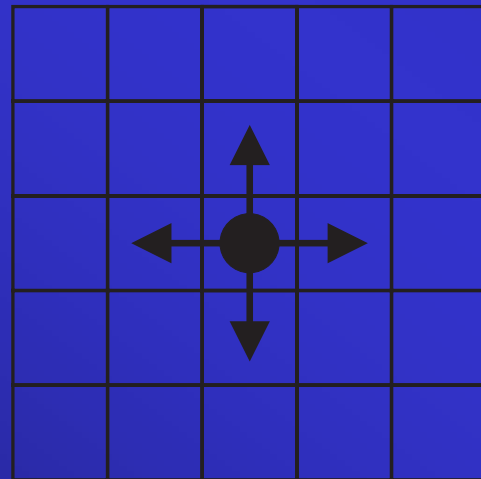
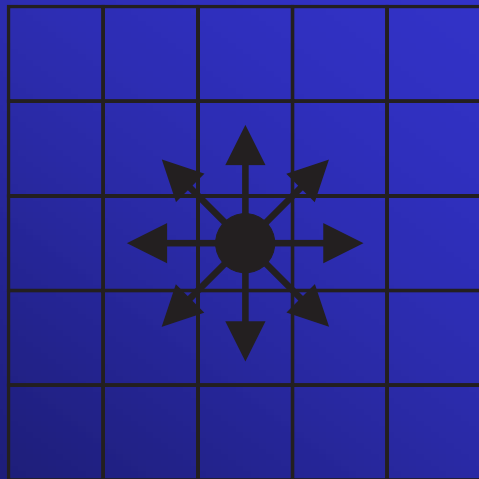
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Let's say we're interested in cell  $i = 6$

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

# Queens and Rooks (and—occasionally—Bishops)

These terms are self explanatory,  
referring to which types of adjacent cells  
we choose to include as “neighbors”



# Under a (1<sup>st</sup> order) “queen” criterion

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

# Neighbors $j$

1 2 ...

16

1

2

⋮

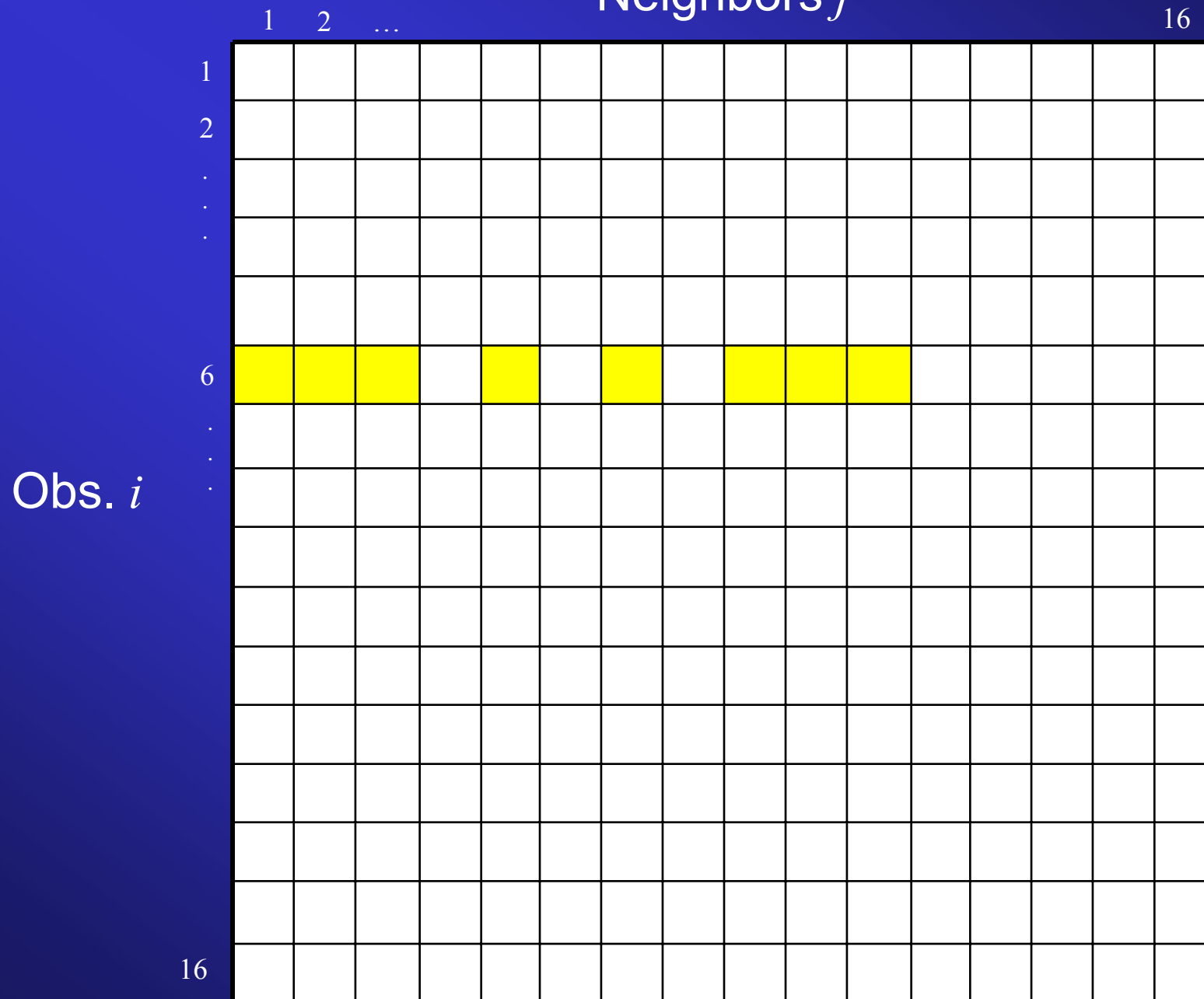
⋮

⋮

Obs.  $i$

16

# Neighbors $j$



Now let's be a little more precise concerning what we're talking about

- “Proximity matrix” – a general term that can refer either to a contiguity matrix or a spatial weights matrix
- “Contiguity matrix” – refers to a matrix that structures the notion of proximity or adjacency as 1 or 0
- “Weights matrix” – I will almost always reserve this term to refer to a row-standardized contiguity matrix

*i*

*j*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	$\Sigma$
1		1			1	1											3
2	1		1		1	1	1										5
3		1		1		1	1	1									5
4			1				1	1									3
5	1	1				1			1	1							5
6	1	1	1		1		1		1	1	1						8
7		1	1	1		1		1		1	1	1					8
8			1	1			1				1	1					5
9					1	1				1			1	1			5
10					1	1	1		1		1		1	1	1		8
11						1	1	1		1		1		1	1	1	8
12							1	1			1				1	1	5
13									1	1				1			3
14									1	1	1		1		1		5
15										1	1	1		1		1	5
16											1	1			1		3

**Simple  
Contiguity  
Matrix  
{*c<sub>ij</sub>*}  
Queen  
Criterion**

	1	2	3	4	5	6	7	8	<i>j</i>	9	10	11	12	13	14	15	16	Σ
1		1/3			1/3	1/3												1
2	1/5		1/5		1/5	1/5	1/5											1
3		1/5		1/5		1/5	1/5	1/5										1
4			1/3				1/3	1/3										1
5	1/5	1/5				1/5			1/5	1/5								1
6	1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8							1
7		1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8						1
8			1/5	1/5			1/5			1/5	1/5							1
9					1/5	1/5			1/5			1/5	1/5					1
10					1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8			1
11						1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8		1
12							1/5	1/5			1/5					1/5	1/5	1
13									1/3	1/3					1/3			1
14									1/5	1/5	1/5		1/5		1/5			1
15										1/5	1/5	1/5		1/5			1/5	1
16											1/3	1/3				1/3		1

Row  
Standardized  
Weights  
Matrix

$$w_{ij} = \frac{c_{ij}}{\sum_j c_{ij}}$$

Now consider these  $y_i$  values,  $i = 1, \dots, 16$

1 7	2 6	3 4	4 5
5 4	6 5	7 4	8 4
9 5	10 6	11 3	12 4
13 3	14 4	15 1	16 2

For  $i = 6$ , the *spatial lag operator*  $w_{6j}y_j$  is given by:

$$\begin{aligned}w_{6j}y_j &= \sum_{j=1}^{j=16} w_{6j}y_j \\ &= \frac{1}{8}7 + \frac{1}{8}6 + \frac{1}{8}4 + \frac{1}{8}4 + \frac{1}{8}4 + \frac{1}{8}5 + \frac{1}{8}6 + \frac{1}{8}3 \\ &= 4.9 \quad (\text{zeros not shown})\end{aligned}$$

In general, the spatial lag is expressed (in matrix notation) as:

$$Wy = \sum_{i=1}^{i=16} \sum_{j=1}^{j=16} w_{ij} y_j$$

where  $W$  is a (16 x 16) weights matrix and  $y$  is a (16 x 1) column vector

Finally, we're ready to  
take a look at the  
Moran statistic

# Global Moran's $I$

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

covariance term

normalization term to scale  $I$  to the overall variance in the dataset

# Moran's $I$ Coefficient as a Measure of Spatial Autocorrelation

Pearson product-moment correlation

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Moran's  $I$  coefficient

$$I_x = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n \sum_{j=1}^n w_{ij} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{j=1}^n (x_j - \bar{x})^2}}$$

# Expected Value of Moran's $I$ Under Hypothesis of No Spatial Autocorrelation

$$E(I) = -\frac{1}{n-1}$$

# Variance of Moran's $I$ Under Hypothesis of No Spatial Autocorrelation (Normalization Perspective)

$$\text{Var}(I) = \frac{n^2 S_1 - n S_2 + 3(\sum_i \sum_j w_{ij})^2}{(\sum_i \sum_j w_{ij})^2 (n^2 - 1)}$$

where  $S_1 = \frac{\sum_i \sum_j (w_{ij} + w_{ji})^2}{2}$

and  $S_2 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2$

# Variance of Moran's $I$ Under Hypothesis of No Spatial Autocorrelation (Randomization Perspective)

$$\text{Var}(I) = \frac{nS_4 - S_3S_5}{(n-1)(n-2)(n-3)\left(\sum_i \sum_j w_{ij}\right)^2}$$

where

$$S_3 = \frac{n^{-1} \sum_i (y_i - \bar{y})^4}{\left(n^{-1} \sum_i (y_i - \bar{y})^2\right)^2}$$

$$S_4 = (n^2 - 3n + 3)S_1 - nS_2 + 3\left(\sum_i \sum_j w_{ij}\right)^2$$

$$S_5 = S_1 - 2nS_1 + 6\left(\sum_i \sum_j w_{ij}\right)^2$$

If  $n$  is large...

$$Z = \frac{I - E(I)}{\sqrt{\text{Var}(I)}}$$

# Interpretation of Moran's $I$

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

- **Positive spatial autocorrelation**
  - $I > -1/(n-1)$ , and  $z > 0$
  - spatial clustering of high and/or low values
- **Negative spatial autocorrelation**
  - $I < -1/(n-1)$ , and  $z < 0$
  - checkerboard pattern, “competition”

# Global Geary's $c$

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

# Expected Value of Geary's $c$ Under Hypothesis of No Spatial Autocorrelation

$$E(c) = 1$$

# Variance of Geary's $c$ Under Hypothesis of No Spatial Autocorrelation (Normalization Perspective)

$$\text{Var}(c) = \frac{(2S_1 + S_2)(n-1) - 4\left(\sum_i \sum_j w_{ij}\right)^2}{2(n+1)\left(\sum_i \sum_j w_{ij}\right)^2}$$

where

$$S_1 = \frac{\sum_i \sum_j (w_{ij} + w_{ji})^2}{2}$$

and

$$S_2 = \sum_i \left( \sum_j w_{ij} + \sum_j w_{ji} \right)^2$$

# Variance of Geary's $c$ Under Hypothesis of No Spatial Autocorrelation (Randomization Perspective)

$$\begin{aligned} \text{Var}(c) = & \{(n-1)S_1[n^2 - 3n + 3 - (n-1)S_3] \\ & - (1/4)(n-1)S_2[n^2 + 3n - 6 - (n^2 - n + 2)S_3] \\ & + (\sum_i \sum_j w_{ij})^2 [n^2 - 3 - (n-1)^2 S_3]\} / n(n-1)(n-2)(\sum_i \sum_j w_{ij})^2 \end{aligned}$$

where  $S_3 = \frac{n^{-1} \sum_i (y_i - \bar{y})^4}{(n^{-1} \sum_i (y_i - \bar{y})^2)^2}$

As with Moran's  $I$ , if  
 $n$  is large...

$$Z = \frac{c - E(c)}{\sqrt{Var(c)}}$$

# Interpretation of Geary's $c$

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

- **Positive spatial autocorrelation**
  - $0 < c < 1, z < 0$
  - spatial clustering of high and/or low values
- **Negative spatial autocorrelation**
  - $1 < c < 2, z > 0$
  - checkerboard pattern, “competition”

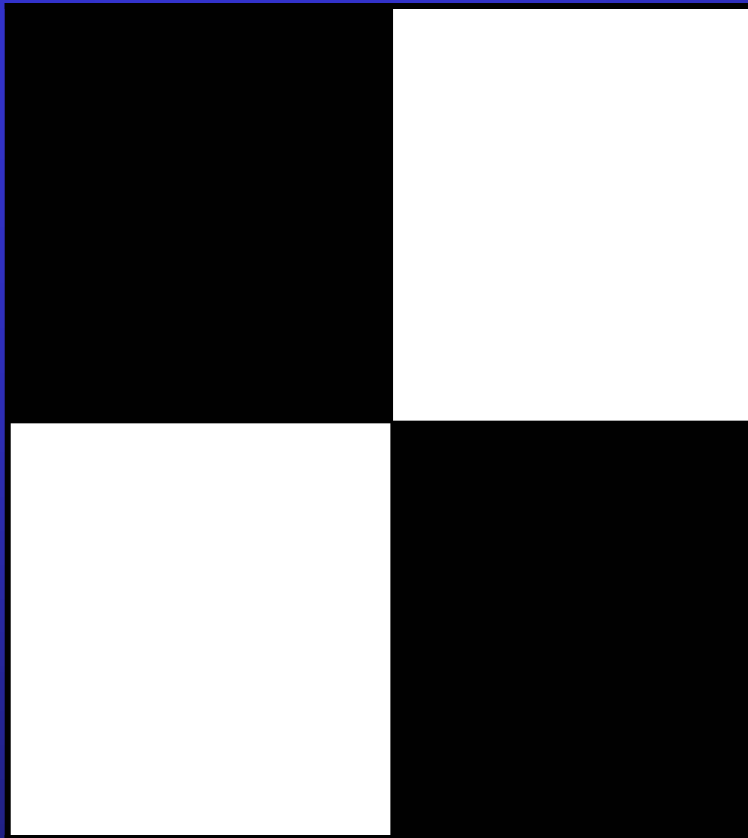
# Values of $I$ and $c$

- When  $I$  approaches  $+1$ : Strong pos. SA
- When  $I$  approaches  $-1$ : Strong neg. SA
- When  $I$  approaches  $0$ : No SA
  
- When  $c$  approaches  $0$ : Strong pos. SA
- When  $c$  approaches  $2$ : Strong neg. SA
- When  $c$  approaches  $1$ : No SA

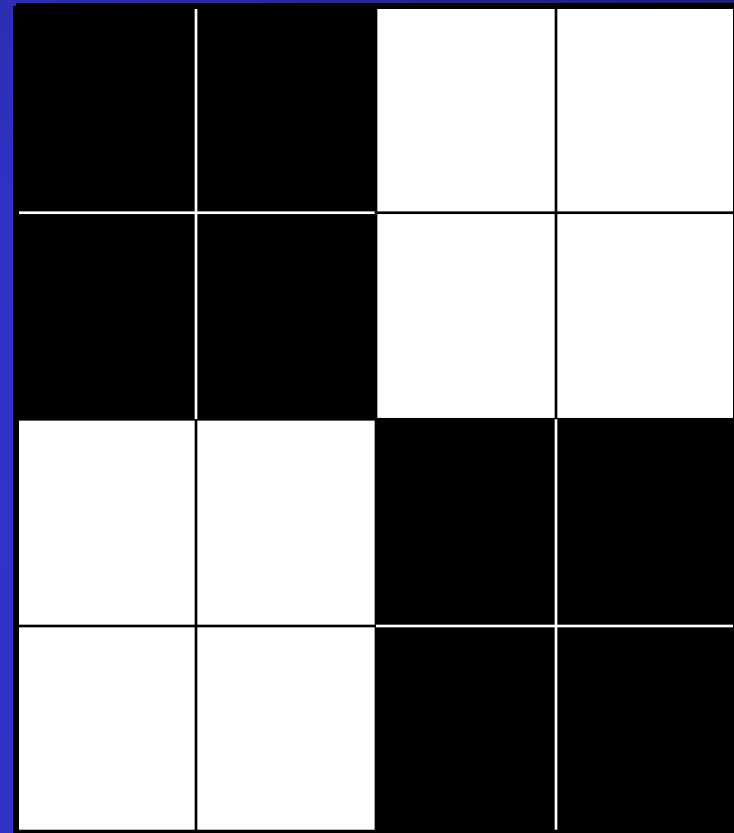
The Geary statistic is a bit awkward, but...

Can transform Geary's  $c$  to a more familiar scale (-1,+1) by subtracting the observed  $c$  from +1

Measures of spatial  
autocorrelation are  
scale dependent



Moran's  $I = -1.00$



Moran's  $I = +0.33$

Testing for spatial autocorrelation in your data is important

Unfortunately, identifying and quantifying the extent of spatial autocorrelation doesn't tell you *what's causing it*

It does alert you to the presence of Spatial “Effects” (or Spatial “Processes”) at work in your data

Spatial dependence

Spatial heterogeneity

- Conceptually, these are very different processes and thus are modeled in very different ways
- Each precludes a straightforward application of standard econometric models
- Each is indicated by spatial autocorrelation

# When spatial autocorrelation in our data is indicated...

- At least one assumption of the standard linear regression model is violated (the classical independence assumption)
- The latent information content in the data is diminished
- We need to do something about it:
  - get rid of it; model it away
  - take advantage of it; bring it into the model
- Either spatial dependence or spatial heterogeneity (or both) should be entertained as potential data-generating models

# Spatial Dependence

# Spatial Dependence

... the existence of a *functional relationship* between what happens at one point in space and what happens elsewhere (Anselin, 1988:11)

- Sounds a lot like spatial autocorrelation... but I do not use the terms interchangeably
- Means a lack of independence among observations (by definition)
- “Functional relationship” is the key
- Sometimes is called a “2<sup>nd</sup> order spatial process,” one operating as a small-scale, localized, short-distance spatial process

# Spatial Dependence...

- In the study of data on a lattice, this spatial process generally is handled through the exogenous declaration of a “neighborhood” for each observation (and operationalized by a “weights matrix”)
- Follows *informally* from the so-called “First Law of Geography”

# Spatial Dependence... (cont.)

- *More formally*, we impose restrictions on the spatial random process that derive from the “ergodicity” property, an invoked assumption of asymptotic independence (non-correlation)
- Structure must be imposed on the random field or we have no traction whatever in terms of parameter estimation. After all, with only  $n$  observations but possibly  $n$  variances and  $n(n-1)/2$  possible covariance parameters defining the spatial random process (plus  $k$  regression parameters), restrictions *must* be imposed so as to reduce the number of parameters to the point where their values can be inferred from the “single realization” at hand. For data on a lattice, this is done by invoking an assumption of ergodicity – i.e., observations are influenced by their neighbors only within a specified limited distance.

# Spatial Heterogeneity

# Spatial Heterogeneity

... exists when the mean, and/or variance, and/or covariance structure “drifts” over a mapped process

- Typified by regional differentiation. Sometimes is called a “1<sup>st</sup> order spatial process,” one operating across the entire region under study; a large-scale, non-localized, long-distance spatial process
- Reflects the “spatial continuities” of social processes which, “taken together help bind social space into recognizable structures” – a “mosaic of homogeneous (or nearly homogeneous)” areas in which each is different from its neighbors (Haining, 1990:22)

## Spatial Heterogeneity... (cont.)

- But no *spatial interaction* is assumed in the process generating spatial heterogeneity. Follows from the “intrinsic uniqueness of each location” (Anselin, 1996:112)
- An undesirable property, because an assumption of spatial *homogeneity* (stationarity) is needed to reduce the number of parameters to an estimable set
- Note: the definition also includes drift in *covariance structure*, and this is the central issue for many analysts...

“The term spatial heterogeneity  
refers to variation in  
*relationships over space.*”

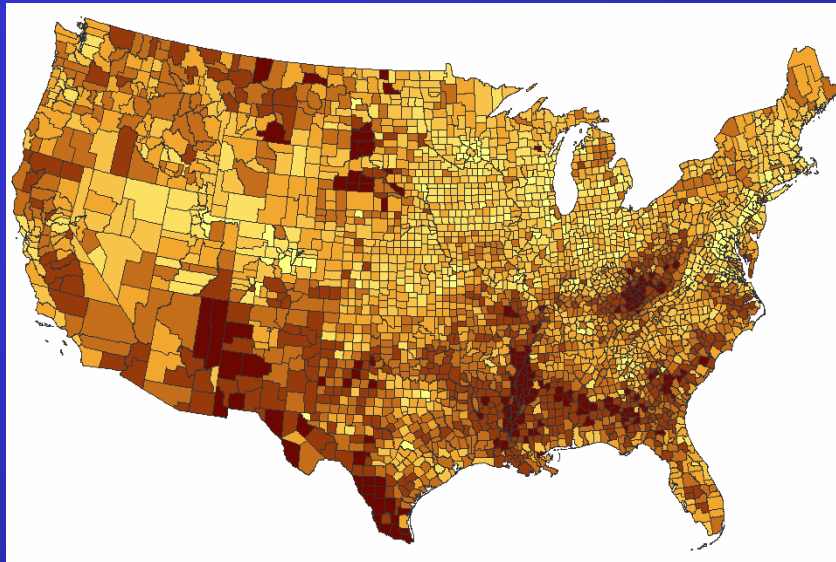
James P. LeSage  
*Spatial Econometrics*  
December, 1998, p. 6  
(emphasis added)  
(book manuscript online at  
<http://www.spatial-econometrics.com/html/wbook.pdf>)

# Spatial Heterogeneity... (cont.)

- Formally is addressed by the concept(s) of spatial stationarity (note here the similarity to time-series analysis)
- Under so-called 2<sup>nd</sup>-order or weak stationarity:
  - $E[X(s)] = E[X(s+\delta)] = \mu$
  - $E[X(s)]^2 = E[X(s+\delta)]^2 = \sigma^2$
  - $E[X(s_i)X(s_j)] = \gamma(d_{ij})$

# Spatial Heterogeneity... (cont.)

- Like the ergodicity assumption, stationarity is assumed in order to help make manageable the number of model parameters to be estimated
- Note: stationarity is an assumed property of our model (the data-generating process). We do not require it as a property of our data (one realization of the data generating process)
- Often generates spatial autocorrelation as a “nuisance”



So, which is it...  
spatial dependence or  
spatial heterogeneity?

**It's not always easy to know**

The data (in cross-section) usually don't help.

Requires our theory to do some heavy lifting.

The process may result from a combination of both.

**Our goal is to understand how we should  
proceed to model our data, recognizing...**

“All models are wrong, some  
models are useful”

G.E.P.Box

“Robustness in the Strategy of Scientific Model Building”  
pp. 201-236 in Lanner and Wilkerson (eds.)

*Robustness in Statistics*  
Academic Press, 1979

So, how do we proceed?

There's no formal roadmap for how to conduct a spatial data analysis, but certainly some steps must precede other

# Recommended Steps in Spatial Data Analysis (1)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
  - put your theory hat on, consider possible structural covariates of dependent variable
  - transform variables as necessary; outliers?
  - visually inspect your maps; outliers?
  - test different weights matrices
  - global and local tests for spatial autocorrelation
  - examine Moran scatterplot; outliers?
  - decisions about outliers
  - look for extent of, and possible amelioration of, spatial heterogeneity

# Recommended Steps in Spatial Data Analysis (2)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
- **OLS baseline model and accompanying diagnostics**
  - Specify model and run in OLS; iterate this for other specifications
  - map residuals & be on lookout for such things as geographic clustering, variance nonstationarity, possible spatial regimes; outliers?
  - examine the diagnostics; where are your problems?
  - What do the LM diagnostics suggest wrt spatial dependence modeling?
  - run model in GWR to further understand spatial structural variance

# Recommended Steps in Spatial Data Analysis (3)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
- OLS baseline model and accompanying diagnostics
- **Correct for spatial heterogeneity if indicated**
  - kernel smoothing
  - surface trend fitting
  - dummy spatial regimes

# Recommended Steps in Spatial Data Analysis (4)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
- OLS baseline model and accompanying diagnostics
- Correct for spatial heterogeneity if indicated
- **With possible controls for spatial heterogeneity, estimate and contrast spatial error and spatial lag model results**
  - Spatial lag model?
  - Spatial error model?
  - What's your theory?

# Recommended Steps in Spatial Data Analysis (5)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
- OLS baseline model and accompanying diagnostics
- Correct for spatial heterogeneity if indicated
- With possible controls for spatial heterogeneity, estimate and contrast spatial error and spatial lag model results
- **Iterate these steps as necessary**

We'll begin here tomorrow morning to explore each of these steps more fully, with examples

Questions?

# Afternoon Lab

ESDA

Spatial Autocorrelation

# Readings for today

- Anselin, Luc. 1996. “The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association.” Pp. 111-125 in Fischer, Manfred, Henk J. Scholten, and David Unwin (eds.) *Spatial Analytical Perspectives on GIS: GISDATA 4* (London: Taylor & Francis).
- Anselin, Luc. 1995. “Local Indicators of Spatial Association – LISA.” *Geographical Analysis* 27(2):93-115. Goodchild, Michael F., Luc Anselin, Richard P. Applebaum, and Barbara Herr Harthorn. 2000. “Toward Spatially Integrated Social Science.” *International Regional Science Review* 23:139-159.