Spatial Analysis and Modeling (GIST 4302/5302)

Guofeng Cao
Department of Geosciences
Texas Tech University

Outline of This Week

- Last week, we learned:
  - spatial point pattern analysis (PPA)
  - focus on location distribution of ‘events’
  - Measure the cluster (spatial autocorrelation)in point pattern
- This week, we will learn:
  - How to measure and detect clusters/spatial autocorrelation in areal data (regional data)

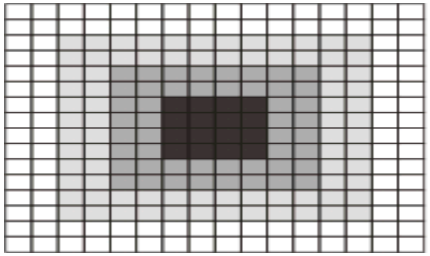
Spatial Autocorrelation

- Spatial autocorrelationship is everywhere
  - Spatial point pattern
    - K, G functions
    - Kernel functions
  - Areal/lattice (this topic)
  - Geostatistical data (next topic)

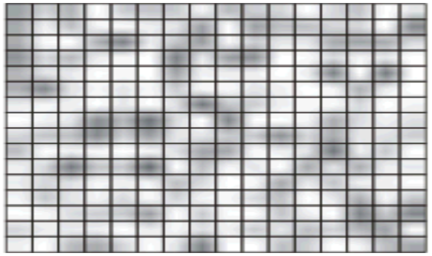
Spatial Autocorrelation of Areal Data

Spatial Autocorrelation

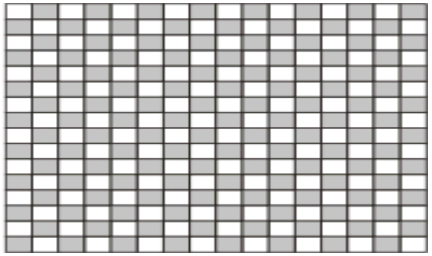
- Tobler's first law of geography
- Spatial auto/cross correlation



If like values tend to cluster together, then the field exhibits high **positive spatial autocorrelation**

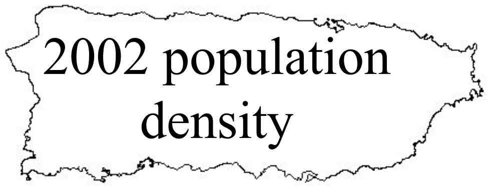


If there is no apparent relationship between attribute value and location then there is **zero spatial autocorrelation**

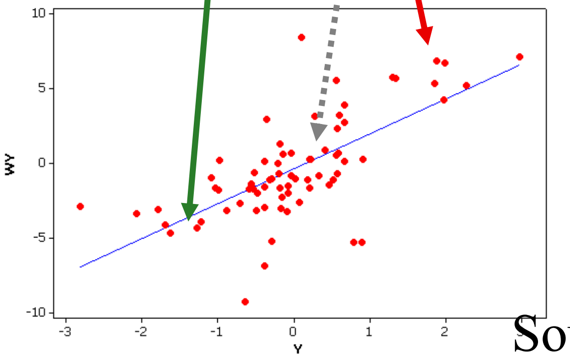


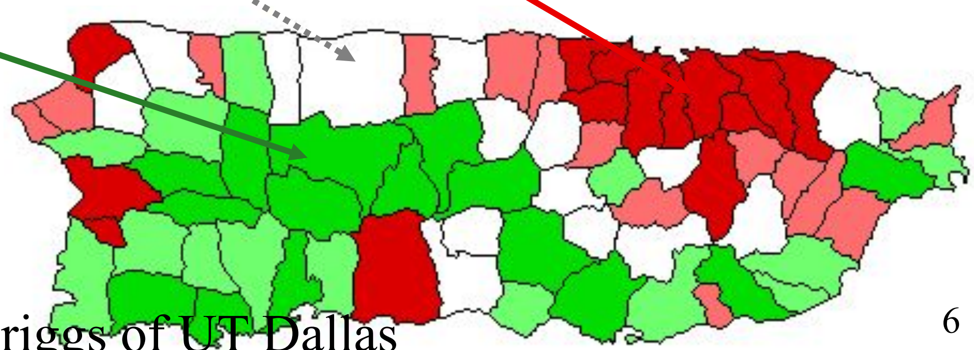
If like values tend to be located away from each other, then there is **negative spatial autocorrelation**

Positive spatial autocorrelation



- high values
  surrounded by nearby high values
- intermediate values surrounded
  by nearby intermediate values
- low values surrounded by
  nearby low values



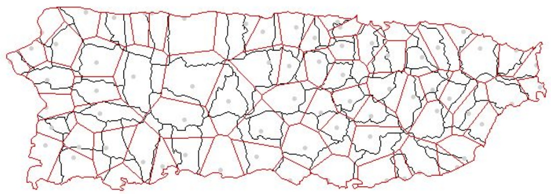


Source: Ron Briggs of UT Dallas

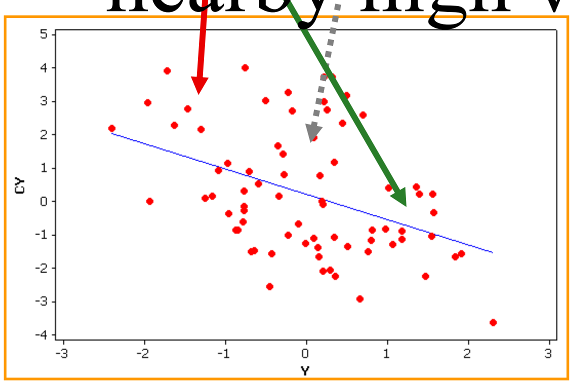
Negative spatial autocorrelation

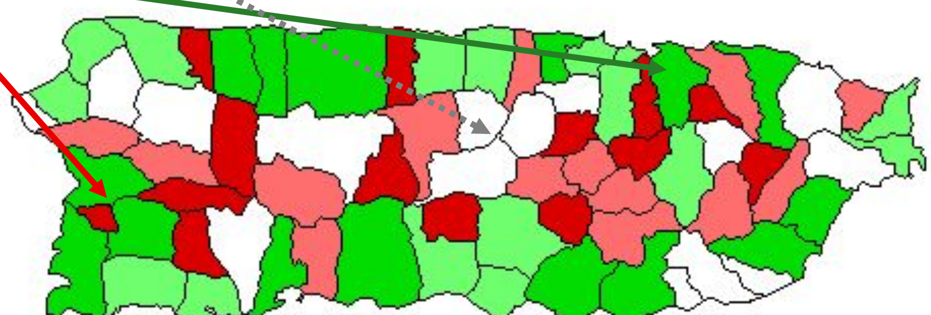
- high values surrounded by nearby low values
- intermediate values surrounded by nearby intermediate values
- low values surrounded by nearby high values

competition for space



Grocery store density





Source: Ron Briggs of UT Dallas

Measuring Spatial Autocorrelation:
the problem of measuring "nearness"

To measure spatial autocorrelation, we must know the "nearness" of our observations as we did for point pattern case

- Which points or polygons are " near" or "next to" other points or polygons?
  - *Which states are near Texas?*
  - How to <u>measure</u> this?

Seems simple and obvious,

but it is not!

Spatial Weight Matrix

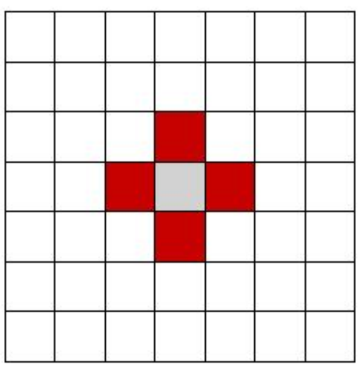
- **Core** concept in statistical analysis of areal data
- Two steps involved:
  - define which relationships between observations are to be given a nonzero weight, i.e., define spatial neighbors
  - assign weights to the neighbors

Spatial Neighbors

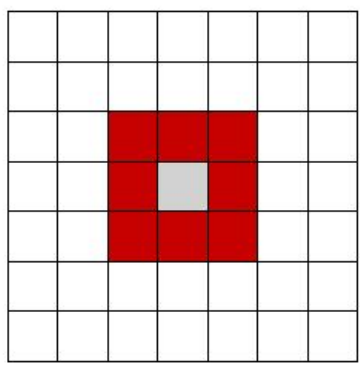
- **Contiguity-based neighbors**
  - Zone *i* and j are neighbors if zone i is contiguity or adjacent to zone *j*
  - But what constitutes contiguity?
- **Distance-based neighbors**
  - Zone i and j are neighbors if the distance between them are less than the threshold distance
  - But what distance do we use?

Contiguity-based Spatial Neighbors

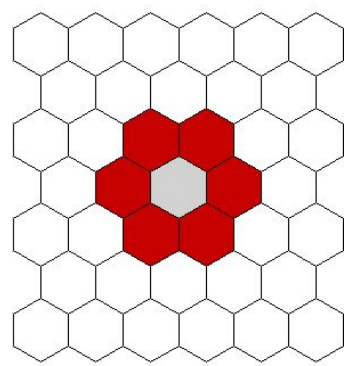
- Sharing a border or boundary
  - Rook: sharing a border
  - Queen: sharing a border <u>or</u> a point



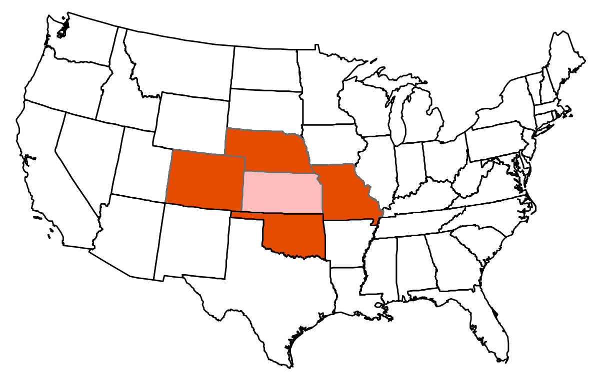
rook



queen



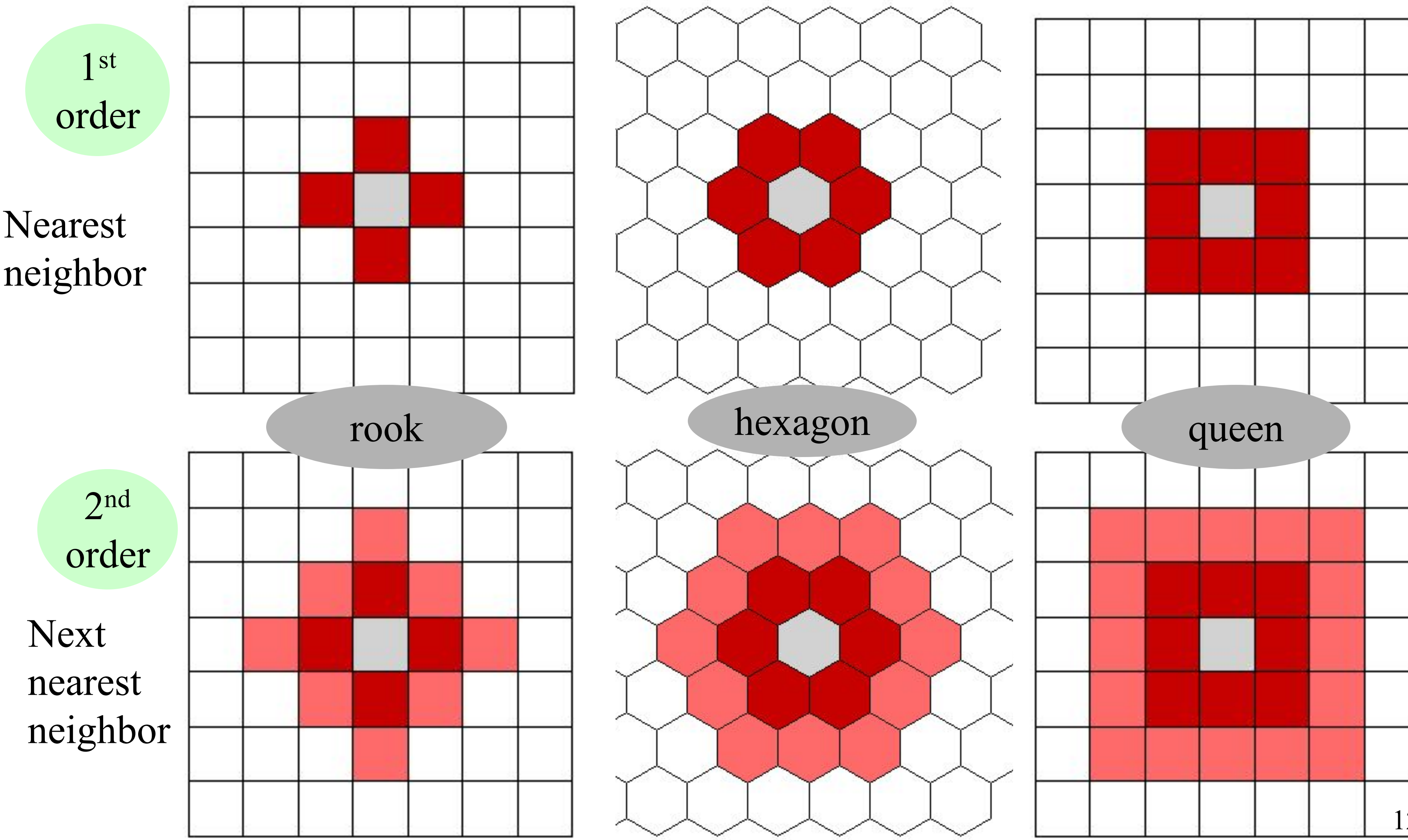
Hexagons



Irregular

Which use?

Higher-Order Contiguity



Distance-based Neighbors

- How to measure distance between polygons?
- Distance metrics
  - 2D Cartesian distance (projected data)
  - 3D spherical distance/great-circle distance (lat/long data)
    - Haversine formula

Haversine formula:

$$a = \sin^2(\Delta\varphi/2) + \cos(\varphi_1).\cos(\varphi_2).\sin^2(\Delta\lambda/2)$$

$$c = 2.\mathrm{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R.c$$

where φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km)

Distance-based Neighbors

- k-nearest neighbors

Fig. 9.5. (**a**) $k = 1$ neighbours; (**b**) $k = 2$ neighbours; (**c**) $k = 4$ neighbours

Source: Bivand and Pebesma and Gomez-Rubio

Distance-based Neighbors

- thresh-hold distance (buffer)

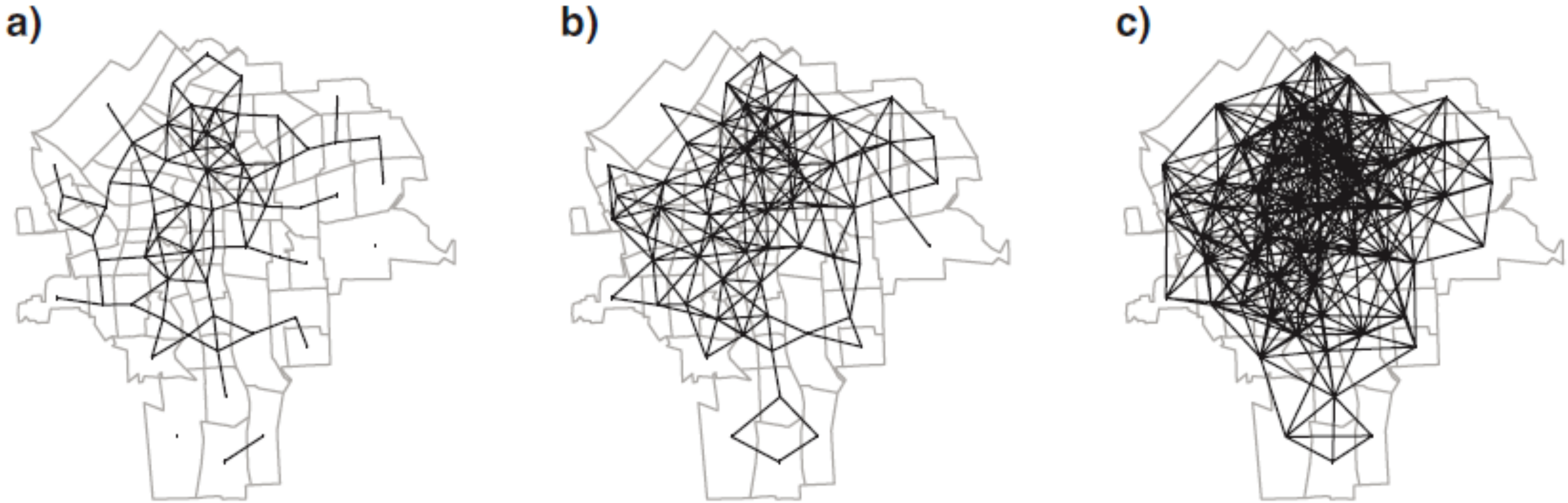
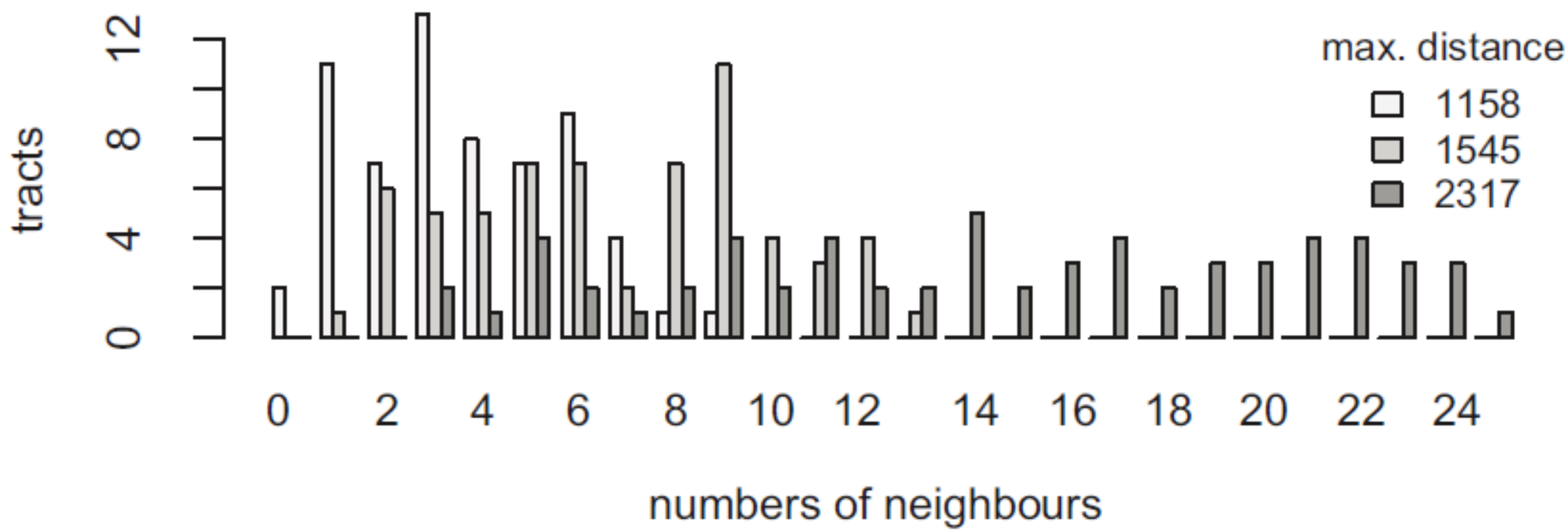


Fig. 9.6. (**a**) Neighbours within 1,158 m; (**b**) neighbours within 1,545 m; (**c**) neighbours within 2,317 m

Source: Bivand and Pebesma and Gomez-Rubio

Neighbor/Connectivity Histogram



Source: Bivand and Pebesma and Gomez-Rubio

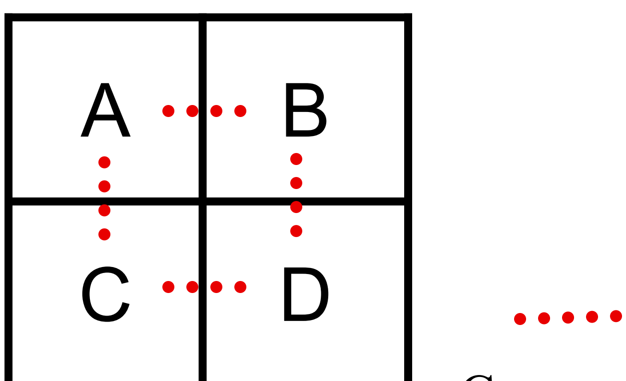
Spatial Weight Matrix

- Spatial weights can be seen as a list of weights indexed by a list of neighbors
- If zone j is not a neighbor of zone i, weights Wij will set to zero
  - The weight matrix can be illustrated as an image
  - Sparse matrix

A Simple Example for Rook case

- Matrix contains a:
  - 1 if share a border
  - 0 if do not share a border

4 areal units



Common border

4x4 matrix

$$
\mathbf{W} = \begin{array}{c|cccc}
 & A & B & C & D \\
\hline
A & 0 & 1 & 1 & 0 \\
B & 1 & 0 & 0 & 1 \\
C & 1 & 0 & 0 & 1 \\
D & 0 & 1 & 1 & 0
\end{array}
$$

1 Washington	1				1																																														
2 Oregon	1		1		1	1																																												
3 California		1		1	1																																													
4 Arizona			1		1				1	1			1																																					
5 Nevada		1	1	1		1			1																																									
6 Idaho	1	1			1		1	1	1																																										
7 Montana							1		1									1	1																																		
8 Wyoming						1	1		1					1		1	1																																	
9 Utah				1	1	1		1		1			1																																						
10 New Mexico				1					1		1	1	1																																						
11 Texas										1		1									1	1																														
12 Oklahoma										1	1		1	1						1	1																														
13 Colorado				1				1	1	1		1		1	1																																			
14 Kansas											1	1		1					1																														
15 Nebraska								1				1	1		1			1	1																														
16 South Dakota							1	1						1		1	1	1																															
17 North Dakota							1								1		1																																
18 Minnesota																	1	1		1									1																						
19 Iowa															1	1		1		1						1	1																					
20 Missouri												1		1	1				1		1			1	1	1																						
21 Arkansas											1	1								1		1	1	1																									
22 Louisiana											1										1		1																										
23 Mississippi																					1	1		1									1																
24 Tennessee																				1	1		1		1								1	1		1	1												
25 Kentucky																				1				1		1			1	1	1						1												
26 Illinois																			1	1					1		1		1																					
27 Wisconsin																		1	1								1		1																					
28 Michigan																											1		1	1																				
29 Indiana																									1	1		1		1																				
30 Ohio																									1			1	1		1											1							
31 West Virginia																									1					1							1	1				1							
32 Florida																																1	1																
33 Alabama																						1	1							1		1																
34 Georgia																							1							1	1		1	1														
35 South Carolina																																	1		1														
36 North Carolina																							1									1	1		1												
37 Virginia																							1	1						1					1		1		1										
38 Maryland																														1						1		1	1		1							
39 Delaware																																					1			1	1							
40 District of Columbia																																				1	1												
41 New Jersey																																						1			1	1						
42 Pennsylvania																													1	1							1	1		1		1						
43 New York																																								1	1		1		1		1	
44 Connecticut																																										1		1	1			
45 Rhode Island																																											1		1			
46 Massachussets																																										1	1	1		1	1	
47 New Hampshire																																													1		1	1
48 Vermont																																										1			1	1		
49 Maine																																														1		

Sparse Contiguity Matrix for US States -- obtained from Anselin's web site (see powerpoint for link)										
Name	**Fips**	**Ncount**	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**
Alabama	1	4	28	13	12	47				
Arizona	4	5	35	8	49	6	32			
Arkansas	5	6	22	28	48	47	40	29		
California	6	3	4	32	41					
Colorado	8	7	35	4	20	40	31	49	56	
Connecticut	9	3	44	36	25					
Delaware	10	3	24	42	34					
District of Columbia	11	2	51	24						
Florida	12	2	13	1						
Georgia	13	5	12	45	37	1	47			
Idaho	16	6	32	41	56	49	30	53		
Illinois	17	5	29	21	18	55	19			
Indiana	18	4	26	21	17	39				
Iowa	19	6	29	31	17	55	27	46		
Kansas	20	4	40	29	31	8				
Kentucky	21	7	47	29	18	39	54	51	17	
Louisiana	22	3	28	48	5					
Maine	23	1	33							
Maryland	24	5	51	10	54	42	11			
Massachusetts	25	5	44	9	36	50	33			
Michigan	26	3	18	39	55					
Minnesota	27	4	19	55	46	38				
Mississippi	28	4	22	5	1	47				
Missouri	29	8	5	40	17	21	47	20	19	31
Montana	30	4	16	56	38	46				
Nebraska	31	6	29	20	8	19	56	46		
Nevada	32	5	6	4	49	16	41			
New Hampshire	33	3	25	23	50					
New Jersey	34	3	10	36	42					
New Mexico	35	5	48	40	8	4	49			
New York	36	5	34	9	42	50	25			
North Carolina	37	4	45	13	47	51				
North Dakota	38	3	46	27	30					
Ohio	39	5	26	21	54	42	18			
Oklahoma	40	6	5	35	48	29	20	8		
Oregon	41	4	6	32	16	53				
Pennsylvania	42	6	24	54	10	39	36	34		
Rhode Island	44	2	25	9						
South Carolina	45	2	13	37						
South Dakota	46	6	56	27	19	31	38	30		
Tennessee	47	8	5	28	1	37	13	51	21	29
Texas	48	4	22	5	35	40				
Utah	49	6	4	8	35	56	32	16		
Vermont	50	3	36	25	33					
Virginia	51	6	47	37	24	54	11	21		
Washington	53	2	41	16						
West Virginia	54	5	51	21	24	39	42			
Wisconsin	55	4	26	17	19	27				
Wyoming	56	6	49	16	31	8	46	30		

Style of Spatial Weight Matrix

- Row
  - a weight of unity for each neighbor relationship
- Row standardization
  - Symmetry not guaranteed
  - can be interpreted as allowing the calculation of average values across neighbors
- General spatial weights based on distances

Row vs. Row standardization

A	B	C
D	E	F

Divide each number by the **row sum**

Total number of neighbors
--some have more than others

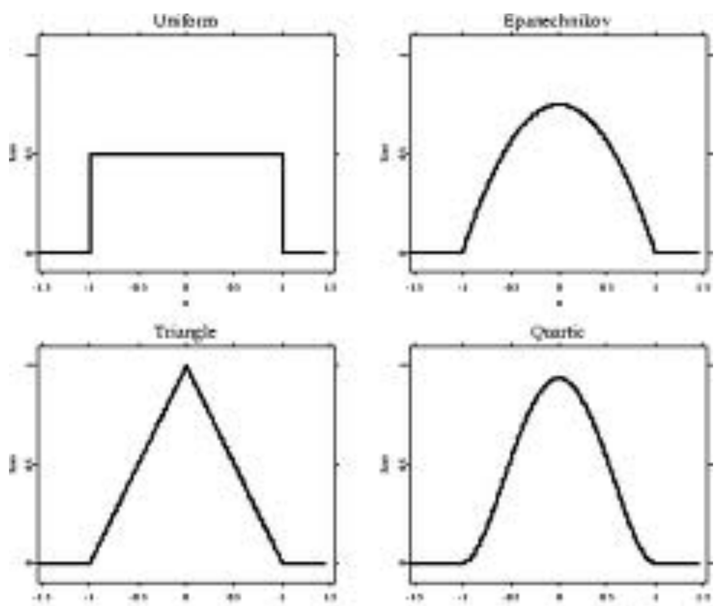
	A	**B**	**C**	**D**	**E**	**F**	**Row Sum**
A	0	1	0	1	0	0	2
B	1	0	1	0	1	0	***3***
C	0	1	0	0	0	1	2
D	1	0	0	0	1	0	2
E	0	1	0	1	0	1	***3***
F	0	0	1	0	1	0	2

Row standardized
--usually use this

	A	**B**	**C**	**D**	**E**	**F**	**Row Sum**
A	0.0	0.5	0.0	0.5	0.0	0.0	1
B	0.3	0.0	0.3	0.0	0.3	0.0	1
C	0.0	0.5	0.0	0.0	0.0	0.5	1
D	0.5	0.0	0.0	0.0	0.5	0.0	1
E	0.0	0.3	0.0	0.3	0.0	0.3	1
F	0.0	0.0	0.5	0.0	0.5	0.0	1

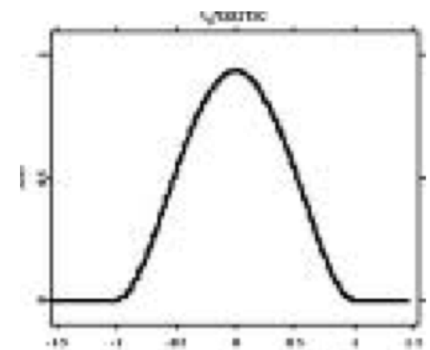
General Spatial Weights Based on Distance

- Decay functions of distance
  - Most common choice is the inverse (reciprocal) of the distance between locations i and j ($w_{ij} = 1/d_{ij}$)
  - Other functions also used
    - inverse of <u>squared</u> distance ($w_{ij} = 1/d_{ij}^2$), or
    - negative exponential ($w_{ij} = e^{-d}$ *or* $w_{ij} = e^{-d^2}$)



Distance-based Spatial Weight Matrix

A	B	C
D	E	F



	A	B	C	D	E	F
A	0	2	0	2	1	0
B	2	0.0	2	1	2	1
C	0	2	0	0	1	2
D	2	1	0	0	2	0
E	1	2	1	2	0	2
F	0	1	2	0	2	0

Measure of Spatial Autocorrelation

Global Measures and Local Measures

- Global Measures
  - A single value which applies to the entire data set
    - The same pattern or process occurs over the entire geographic area
    - An average for the entire area
- Local Measures
  - A value calculated for <u>each</u> observation unit
    - Different patterns or processes may occur in different parts of the region
    - A unique number for each location
- Global measures usually can be decomposed into a combination of local measures

Global Measures and Local Measures

- Global Measures
  - Moran's I
- Local Measures
  - Local Moran's I

Moran's I

- The most common measure of Spatial Autocorrelation
- Use for points <u>or</u> polygons

Patrick Alfred Pierce Moran (1917-1988)

Formula for Moran's I

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

- Where:

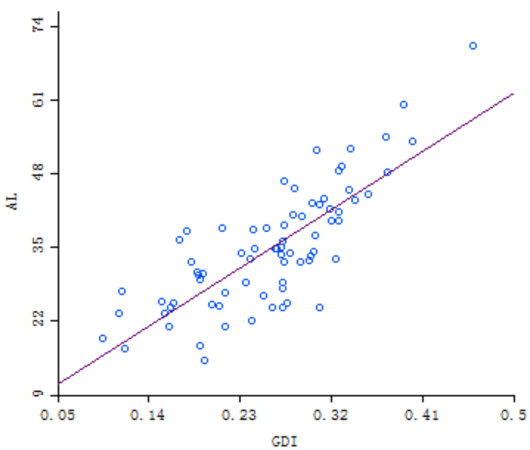
N is the number of observations (points or polygons)
$\bar{x}$ is the mean of the variable
X_i is the variable value at a particular location
X_j is the variable value at another location
W_{ij} is a weight indexing location of *i* relative to *j*

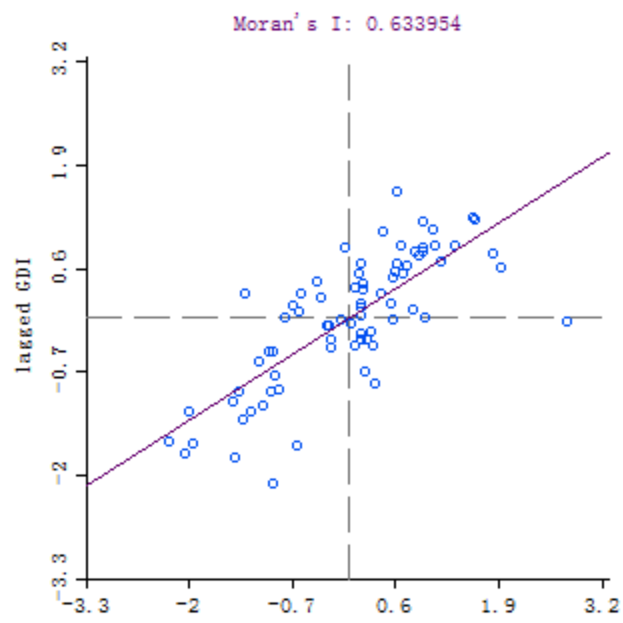
Moran's *I* and Correlation Coefficient

- **Correlation Coefficient [-1, 1]**
  - Relationship between <u>two</u> different variables
- **Moran's I [-1, 1]**

– Spatial autocorrelation and often involves <u>one</u> (spatially indexed) variable only

– Correlation between observations of a spatial variable at location X and "spatial lag" of X formed by averaging all the observation at neighbors of X





Correlation Coefficient

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

Note the similarity of the numerator (top) to the measures of spatial association discussed earlier if we view Yi as being the Xi for the neighboring polygon

(see next slide)

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

Spatial auto-correlation

Source: Ron Briggs of UT Dallas

Correlation Coefficient

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

Spatial weights

Yi is the Xi for the neighboring polygon

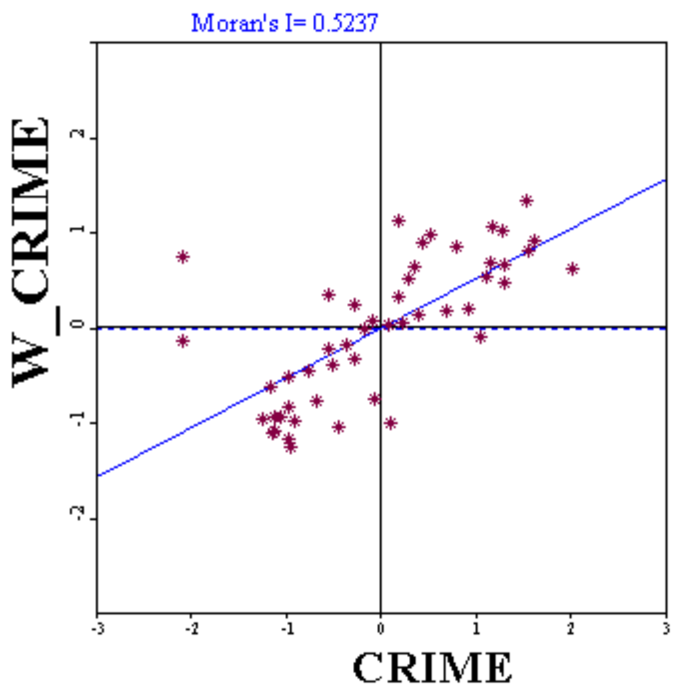
$$\frac{N\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})/\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}(x_i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

Moran's I

Source: Ron Briggs of UT Dallas

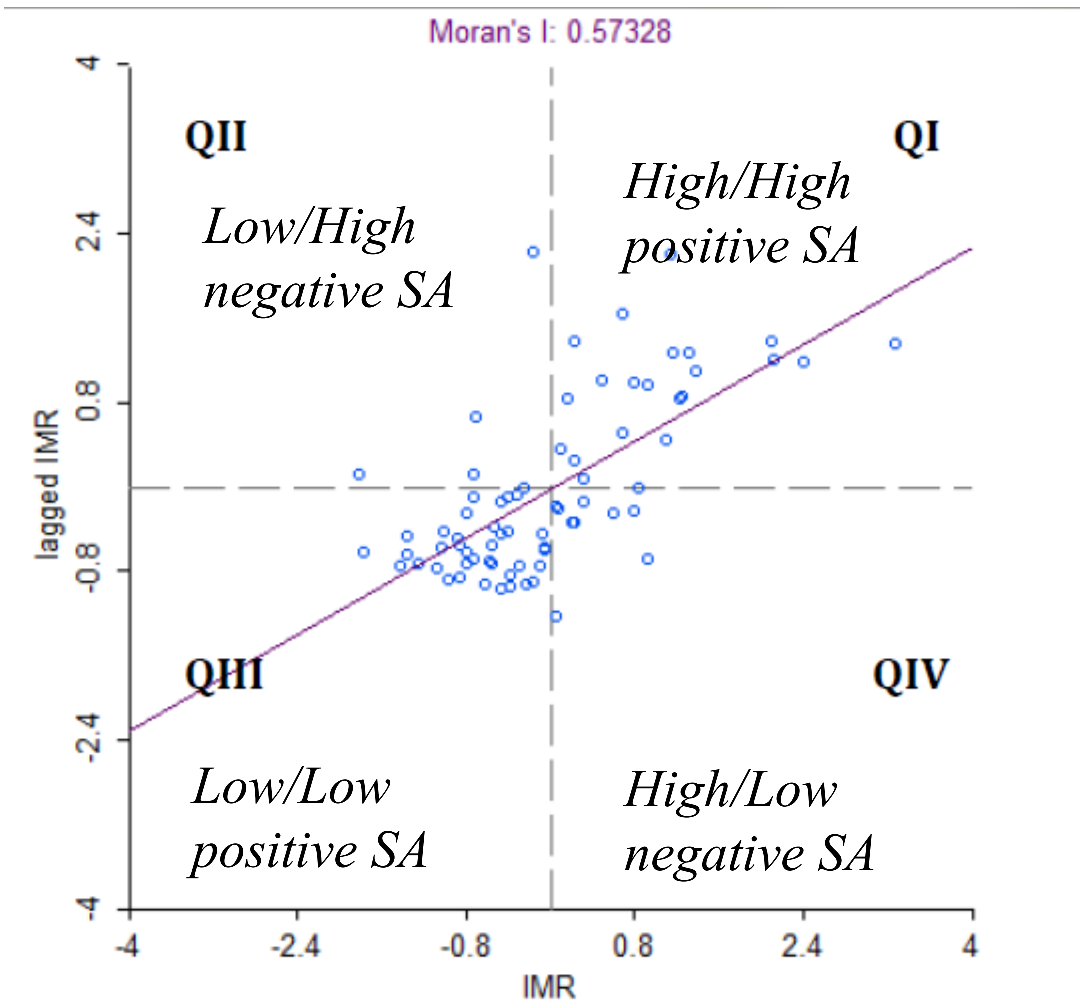
Moran Scatter Plots

We can draw a scatter diagram between these two variables (in standardized form): **X** and **lag-X** (or W_X)

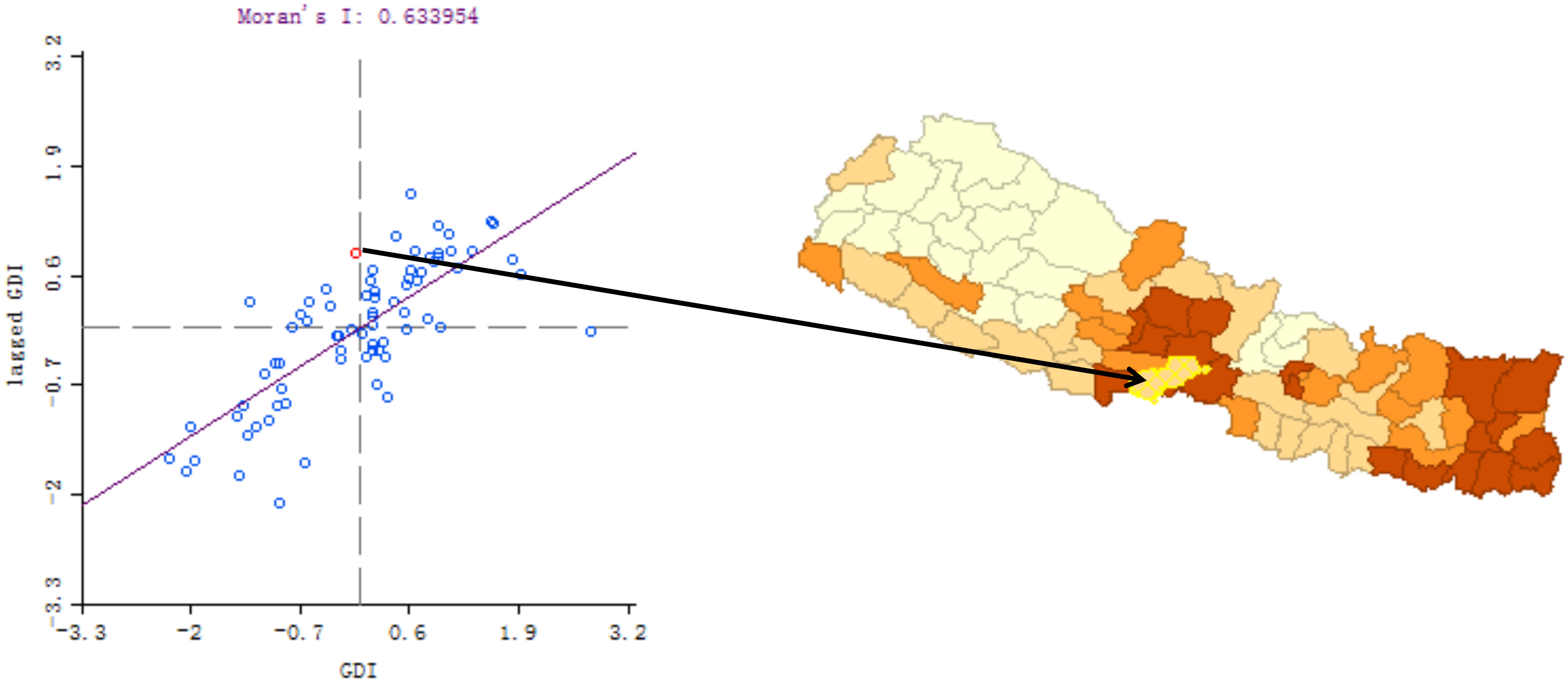


The <u>slope</u> of this *regression line* is Moran's I

Moran Scatter Plots



Moran Scatterplot: Example



Statistical Significance Tests for Moran's I

- Based on the normal frequency distribution with

$$Z = \frac{I - E(I)}{S_{error(I)}}$$

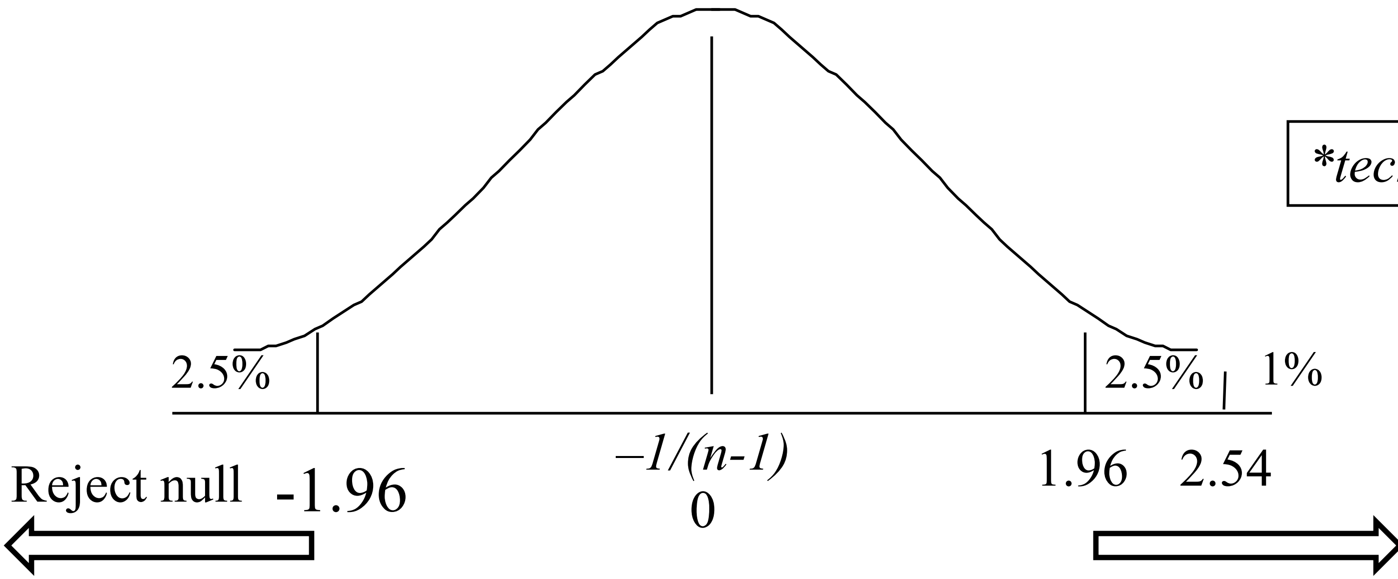
Where: I is the calculated value for Moran's I from the sample

E(I) is the expected value if random

S is the standard error

- Statistical significance test
  - Monte Carlo test, as we did for spatial pattern analysis
  - Permutation test
    - Non-parametric
    - Data-driven, no assumption of the data
    - Implemented in GeoDa

Test Statistic for Normal Frequency Distribution



technically –1/(n-1)

Null Hypothesis: no spatial autocorrelation

*Moran's I = 0

Alternative Hypothesis: spatial autocorrelation exists

*Moran's I > 0

Reject *Null Hypothesis* if Z test statistic > 1.96 (or < -1.96)

---less than a 5% chance that, in the population, there is no spatial autocorrelation

---95% confident that spatial auto correlation exits

Null Hypothesis: no spatial autocorrelation

*Moran's I = 0

Alternative Hypothesis: spatial autocorrelation exists

*Moran's I > 0

Reject *Null Hypothesis* if Z test statistic > 1.96 (or < -1.96)

---less than a 5% chance that, in the population, there is no spatial autocorrelation

---95% confident that spatial auto correlation exits

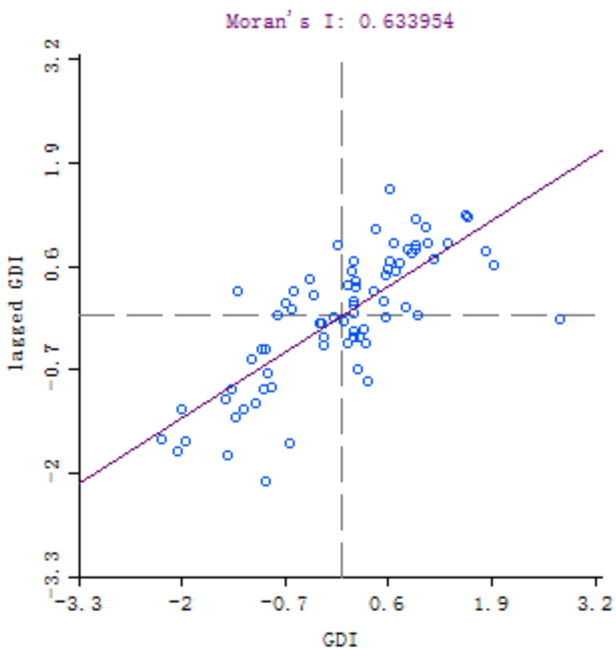
Spatial Autocorrelation vs Correlation

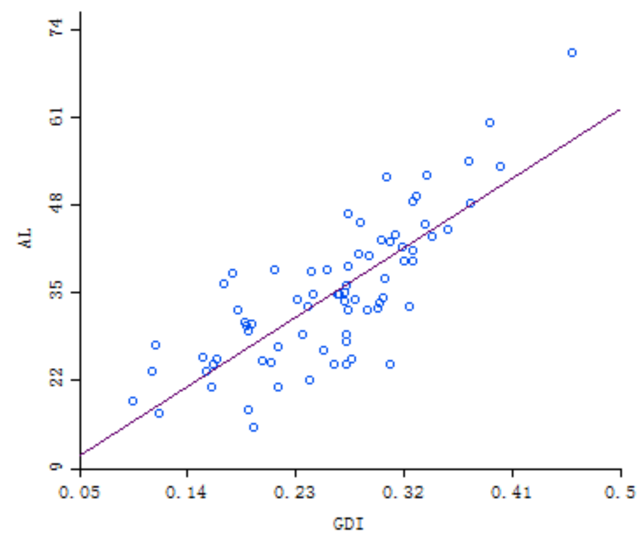
Spatial Autocorrelation:

shows the association or relationship between the <u>same</u> variable in “near-by” areas.

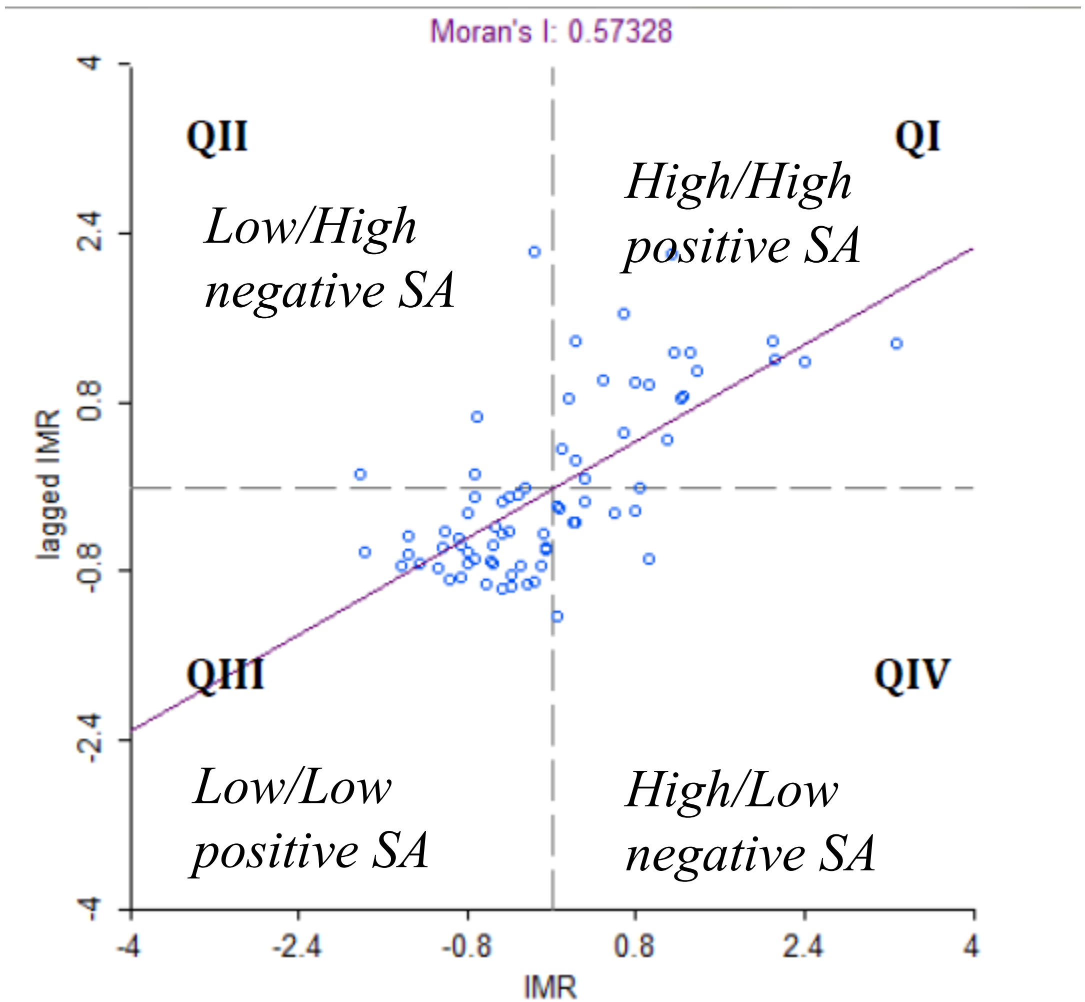
Standard Correlation

shows the association or relationship between two <u>different</u> variables





Bivariate Moran Scatter Plot



Local Measures of Spatial Autocorrelation

Local Indicators of Spatial Association (LISA)

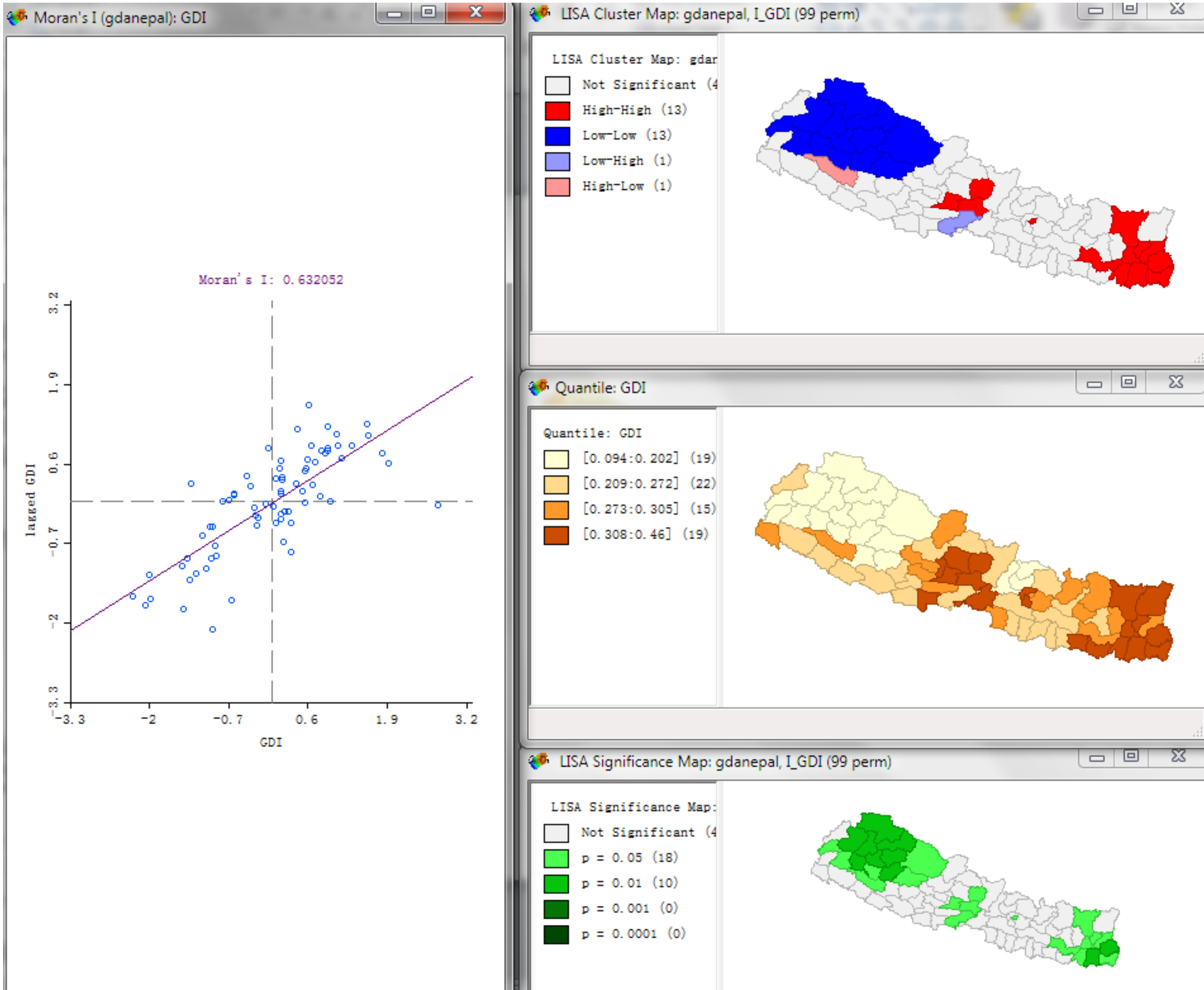
- Local versions of *Moran's I*
- Moran's I is most commonly used, and the local version is often called Anselin's LISA, or just LISA

See:
Luc Anselin 1995 *Local Indicators of Spatial Association-LISA* Geographical Analysis 27: 93-115

Local Indicators of Spatial Association (LISA)

- The statistic is calculated for **each** areal unit in the data
- For each polygon, the index is calculated based on neighboring polygons with which it shares a border
- A measure is available for each polygon, these can be mapped to indicate how spatial autocorrelation varies over the study region
- Each index has an associated test statistic, we can also map which of the polygons has a statistically significant relationship with its neighbors, and show type of relationship

Example:

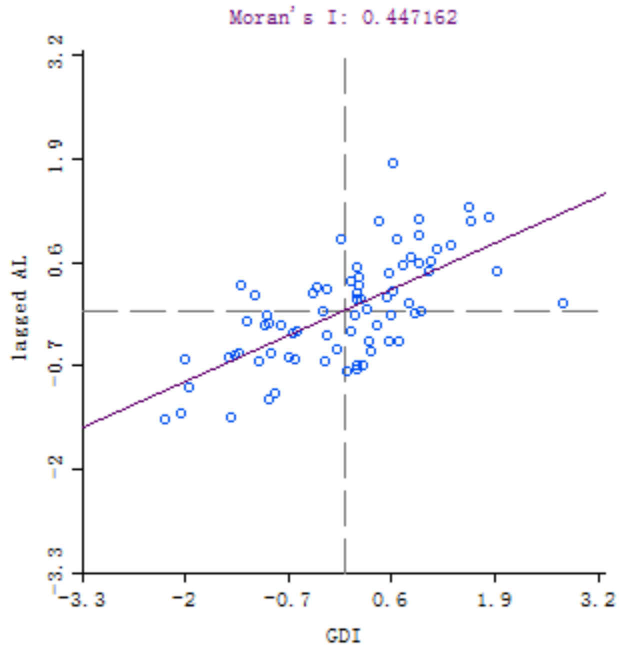


Moran's I (gdanepal): GDI
Moran's I: 0.632052
lagged GDI
GDI
-3.3
-2
-0.7
0.6
1.9
3.2
LISA Cluster Map: gdanepal, I_GDI (99 perm)
LISA Cluster Map:
Not Significant
High-High (13)
Low-Low (13)
Low-High (1)
High-Low (1)
Quantile: GDI
Quantile: GDI
[0.094:0.202] (19)
[0.209:0.272] (22)
[0.273:0.305] (15)
[0.308:0.46] (19)
LISA Significance Map: gdanepal, I_GDI (99 perm)
LISA Significance Map:
Not Significant
p = 0.05 (18)
p = 0.01 (10)
p = 0.001 (0)
p = 0.0001 (0)

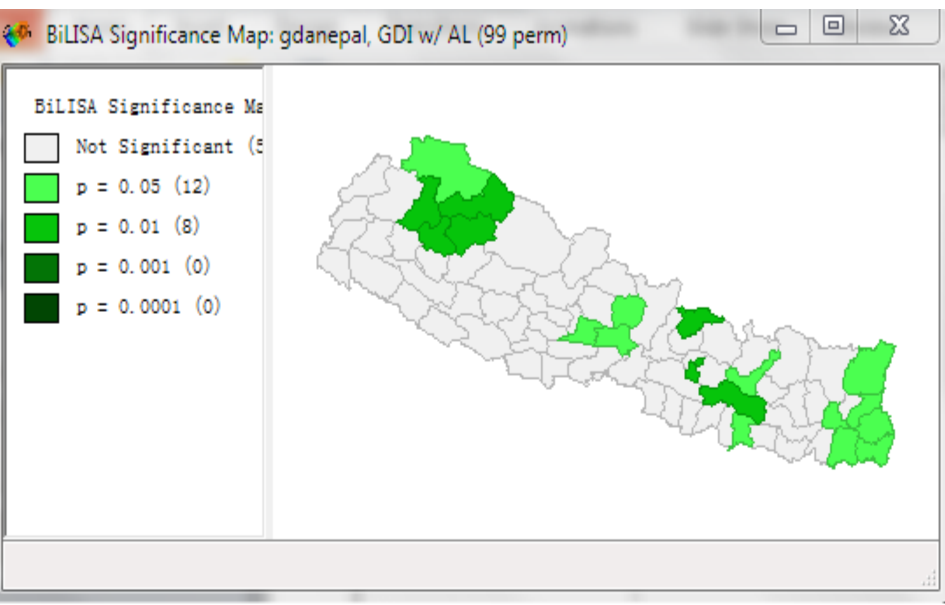
Bivariate LISA

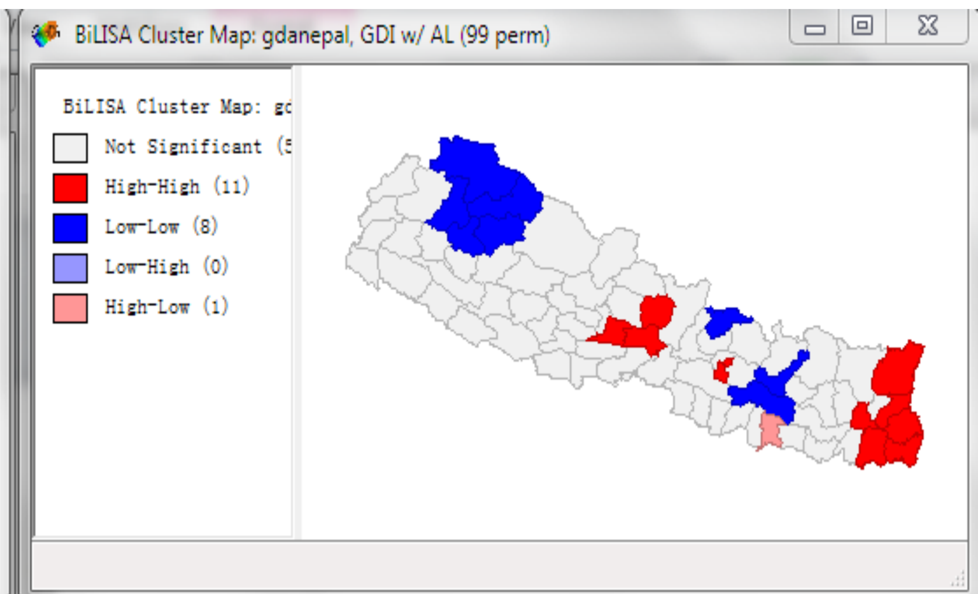
- Moran's I is the correlation between X and Lag-X--the same variable but in nearby areas
  - Univariate Moran's I
- Bivariate Moran's I is a correlation between X and a different variable in nearby areas.

Moran Scatter Plot for GDI vs AL



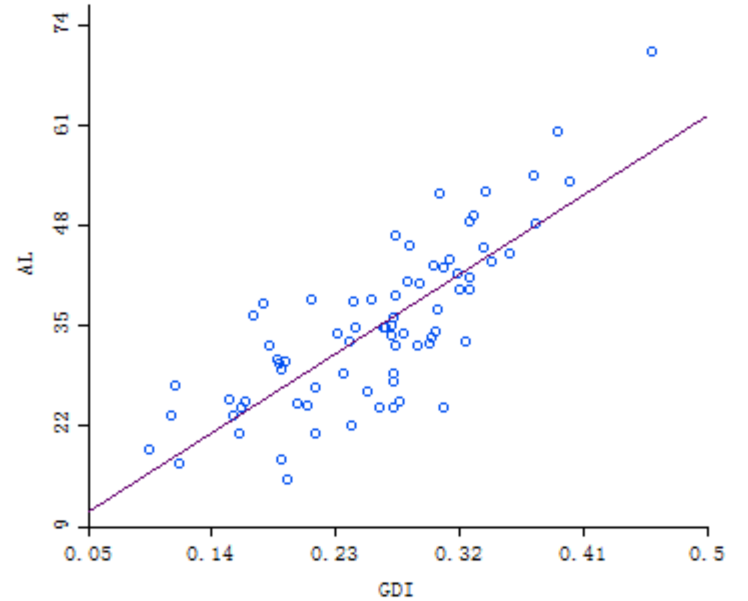
Moran Significance Map for GDI vs. AL

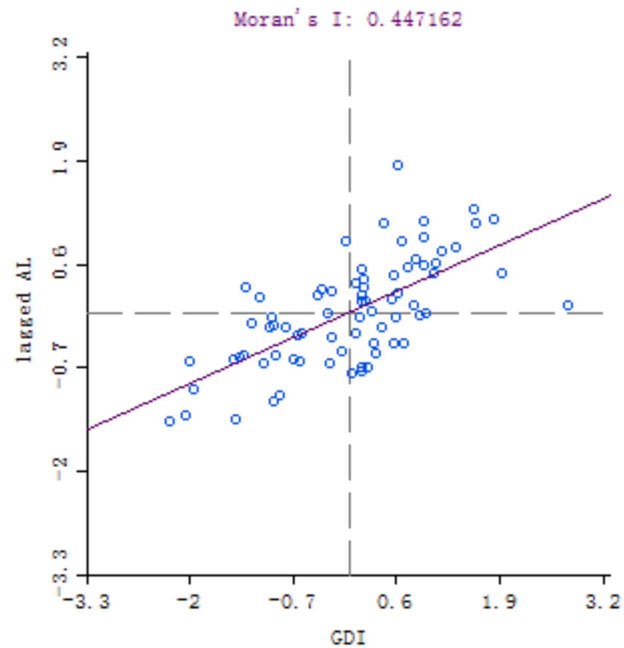




Bivariate LISA
and the Correlation Coefficie

- Correlation Coefficient is the relationship between two <u>different</u> variables in the <u>same</u> area
- Bivariate LISA is a correlation between two <u>different</u> variables in an area and in <u>nearby</u> areas.





Consequences of Ignoring Spatial Autocorrelation

- correlation coefficients and coefficients of determination appear <u>bigger</u> than they really are
  - You think the relationship is stronger than it really is
  - the variables in nearby areas affect each other
- Standard errors appear <u>smaller</u> than they really are
  - *exaggerated precision*
  - You think your predictions are better than they really are since standard errors measure *predictive accuracy*
  - More likely to conclude relationship is *statistically significant*.

Diagnostic of Spatial Dependence

- **For correlation**
  - calculate Moran's I for each variable and test its statistical significance
  - If Moran's I is significant, you may have a problem!
- **For regression**
  - calculate the residuals

    map the residuals: do you see any spatial patterns?
  - Calculate Moran's I for the residuals: is it statistically significant?

Summary

- Spatial autocorrelation of areal data
- Spatial weight matrix
- Measures of spatial autocorrelation
- Global Measure
  - Moran's I
- Consequences of ignoring spatial autocorrelation
- Significance test

- End of this topic