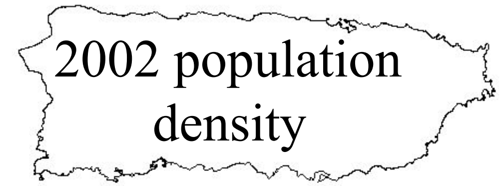
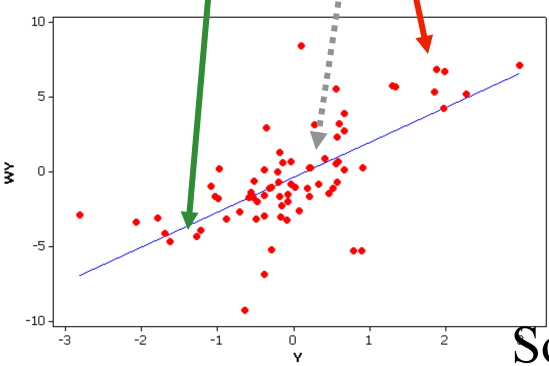
Spatial Autocorrelation of Areal Data

<u>Positive spatial autocorrelation</u>

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



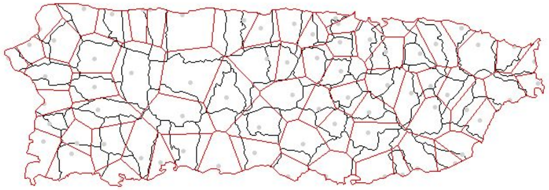


Source: Ron Briggs of UT Dallas

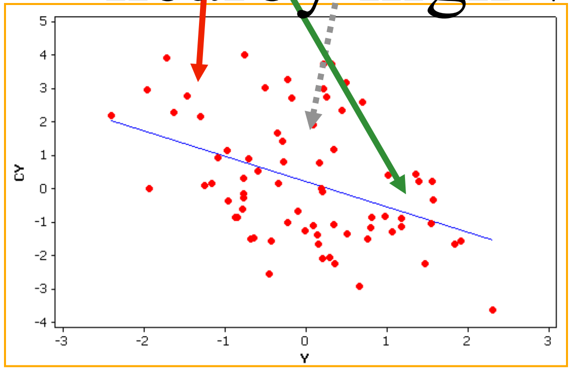
Negative spatial autocorrelation

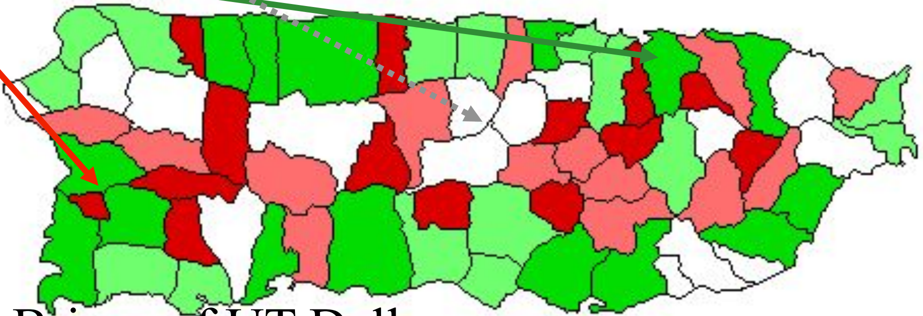
- high values surrounded by nearby low values
- intermediate values surrounded by nearby intermediate values
- low values surrounded by nearby high values

competition for space



Grocery store density





Source: Ron Briggs of UT Dallas

Spatial Weight Matrix

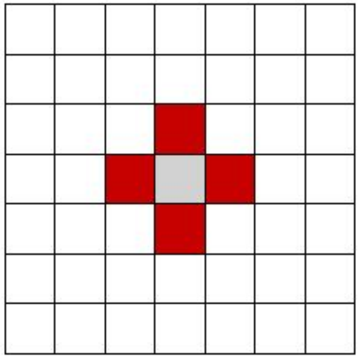
- **Core** concept in statistical analysis of areal data
- Two steps involved:
  - define which relationships between observations are to be given a nonzero weight, i.e., define spatial neighbors
  - assign weights to the neighbors
- Making the neighbors and weights is not easy as it seems to be
  - Which states are near Texas?

Spatial Neighbors

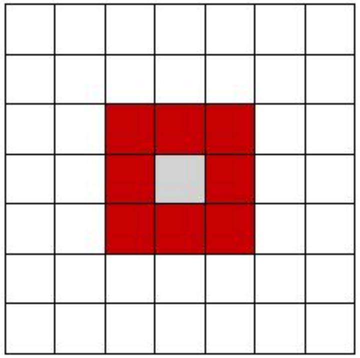
- **Contiguity-based neighbors**
  - Zone *i* and j are neighbors if zone i is contiguity or adjacent to zone *j*
  - But what constitutes contiguity?
- **Distance-based neighbors**
  - Zone i and j are neighbors if the distance between them are less than the threshold distance
  - But what distance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

Contiguity-based Spatial Neighbors

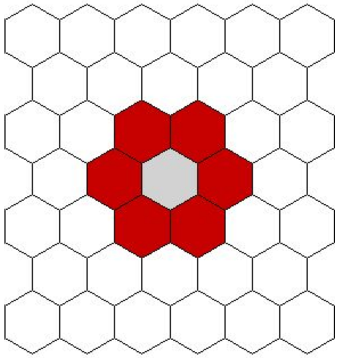
- Sharing a border or boundary
  - Rook: sharing a border
  - Queen: sharing a border <u>or</u> a point



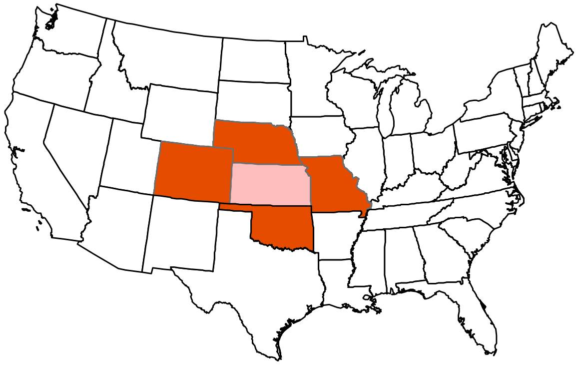
rook



queen



Hexagons



Irregular

Which use?

Example

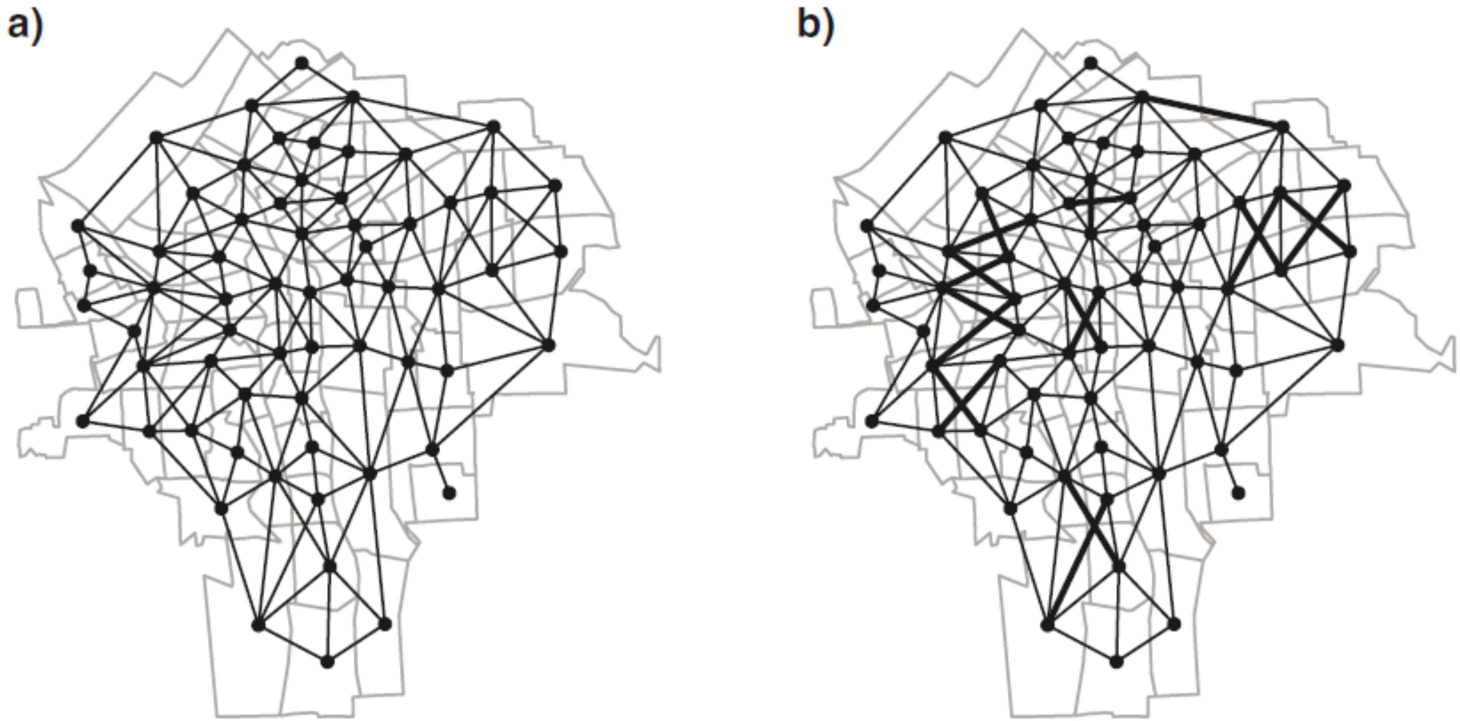
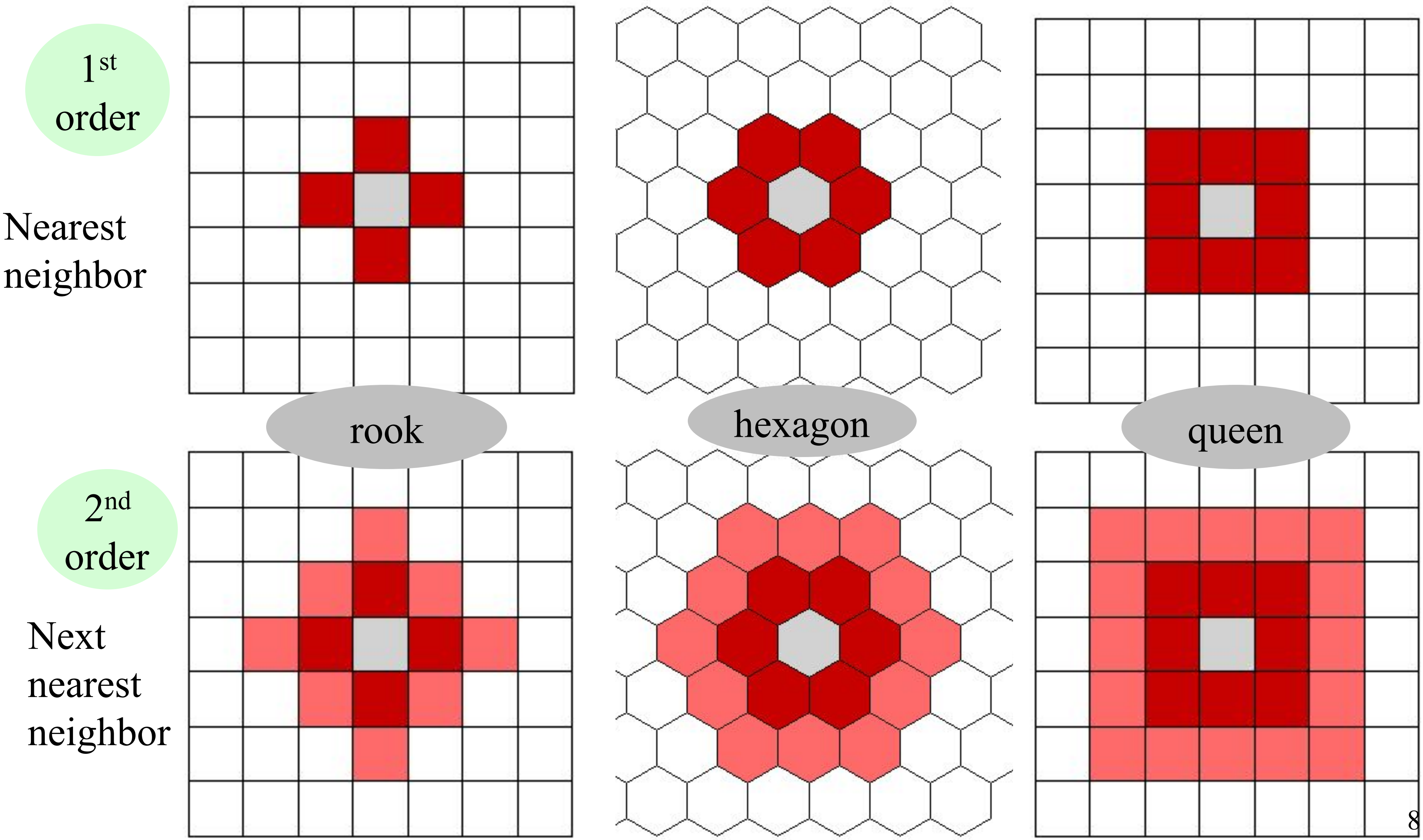


Fig. 9.3. (**a**) Queen-style census tract contiguities, Syracuse; (**b**) Rook-style contiguity differences shown as thicker lines

Source: Bivand and Pebesma and Gomez-Rubio

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

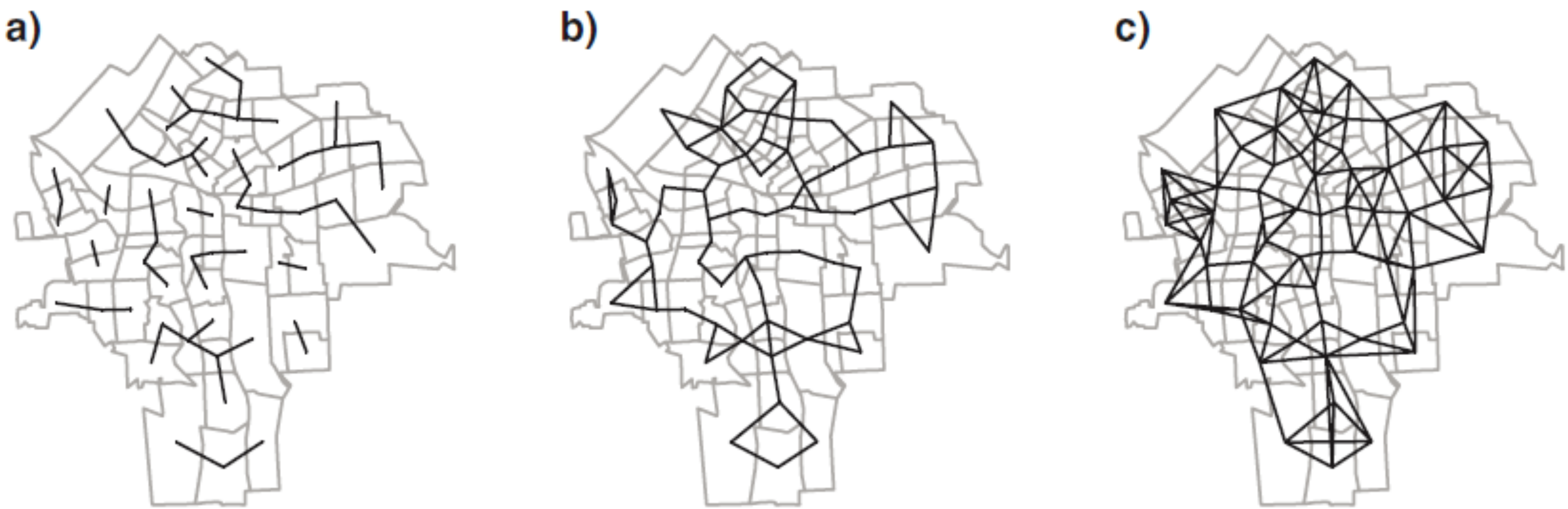


Fig. 9.5. (**a**) $k = 1$ neighbours; (**b**) $k = 2$ neighbours; (**c**) $k = 4$ neighbours

Source: Bivand and Pebesma and Gomez-Rubio

Distance-based Neighbors

- thresh-hold distance (buffer)

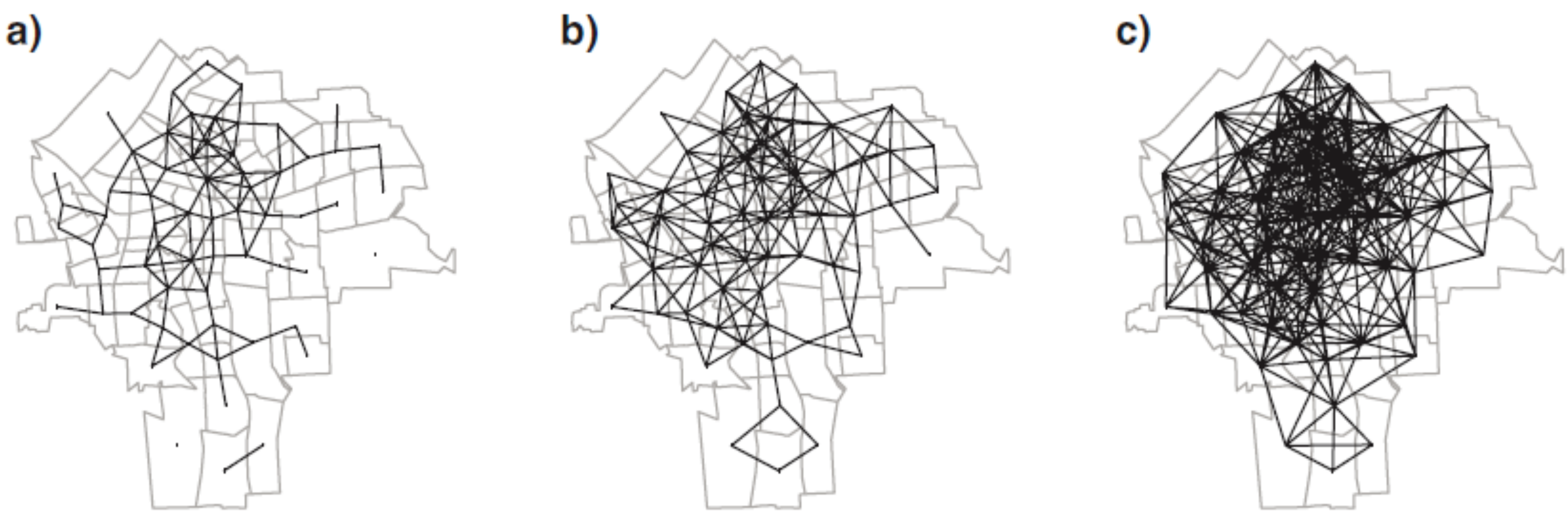


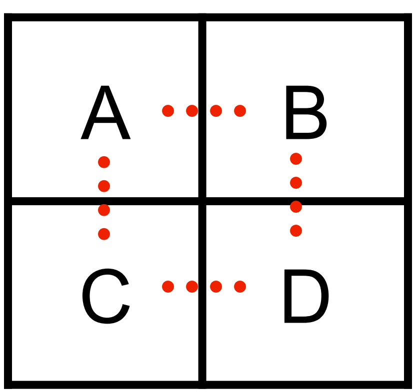
Fig. 9.6. (**a**) Neighbours within 1,158 m; (**b**) neighbours within 1,545 m; (**c**) neighbours within 2,317 m

Source: Bivand and Pebesma and Gomez-Rubio

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

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

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

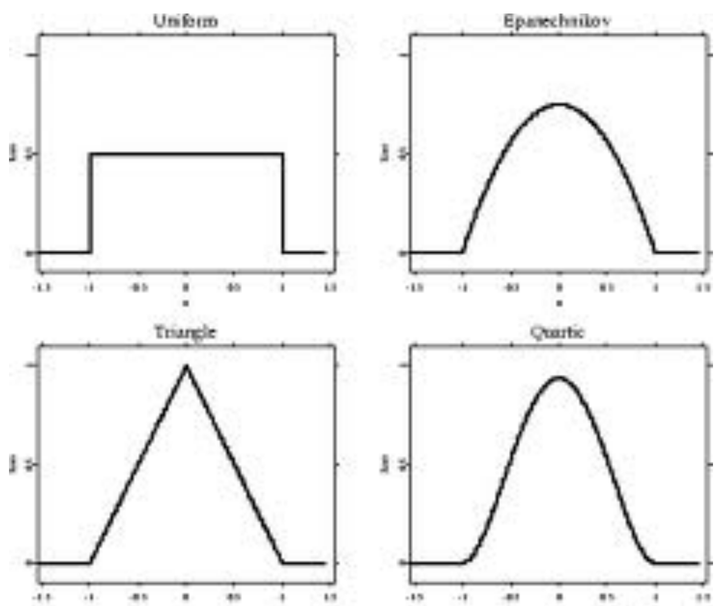
	A	B	C	D	E	F	Row Sum
A	0	1	0	1	0	0	2
B	1	0	1	0	1	0	***3***
C	0	1	0	0	0	1	2
D	1	0	0	0	1	0	2
E	0	1	0	1	0	1	***3***
F	0	0	1	0	1	0	2

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* $w_{ij} = e^{-d^2}$)



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
  - Join Count
  - Moran’s I, Getis-Ord’s G
- Local Measures
  - Local Moran’s I , Getis-Ord’s G

Formula for Moran's I

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

- 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

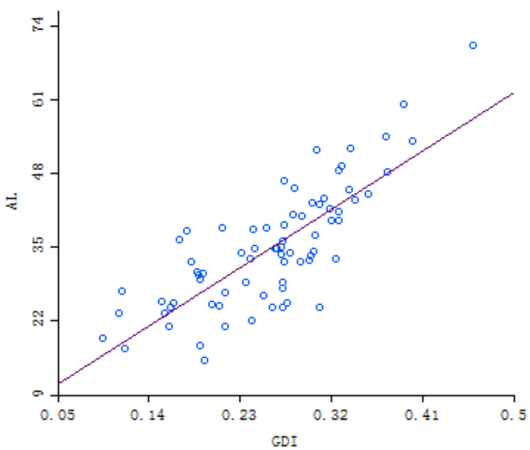
- Expectation of Moran’s I under no spatial autocorrelation

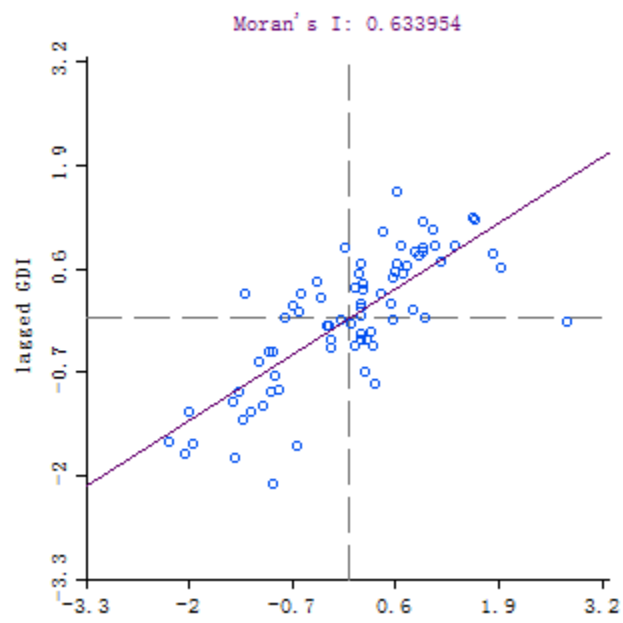
$$E(\mathrm{I}) = -1/(\mathrm{N}-1)$$

- Variance of Moran’s is complex and exact equation is given at textbook d&G&L
- [-1, 1]

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

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

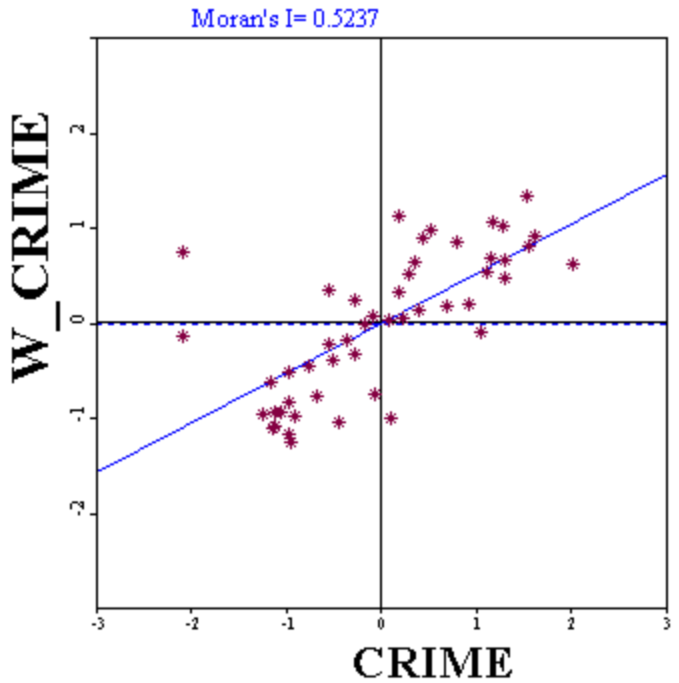
E(I) is the expected value if random

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

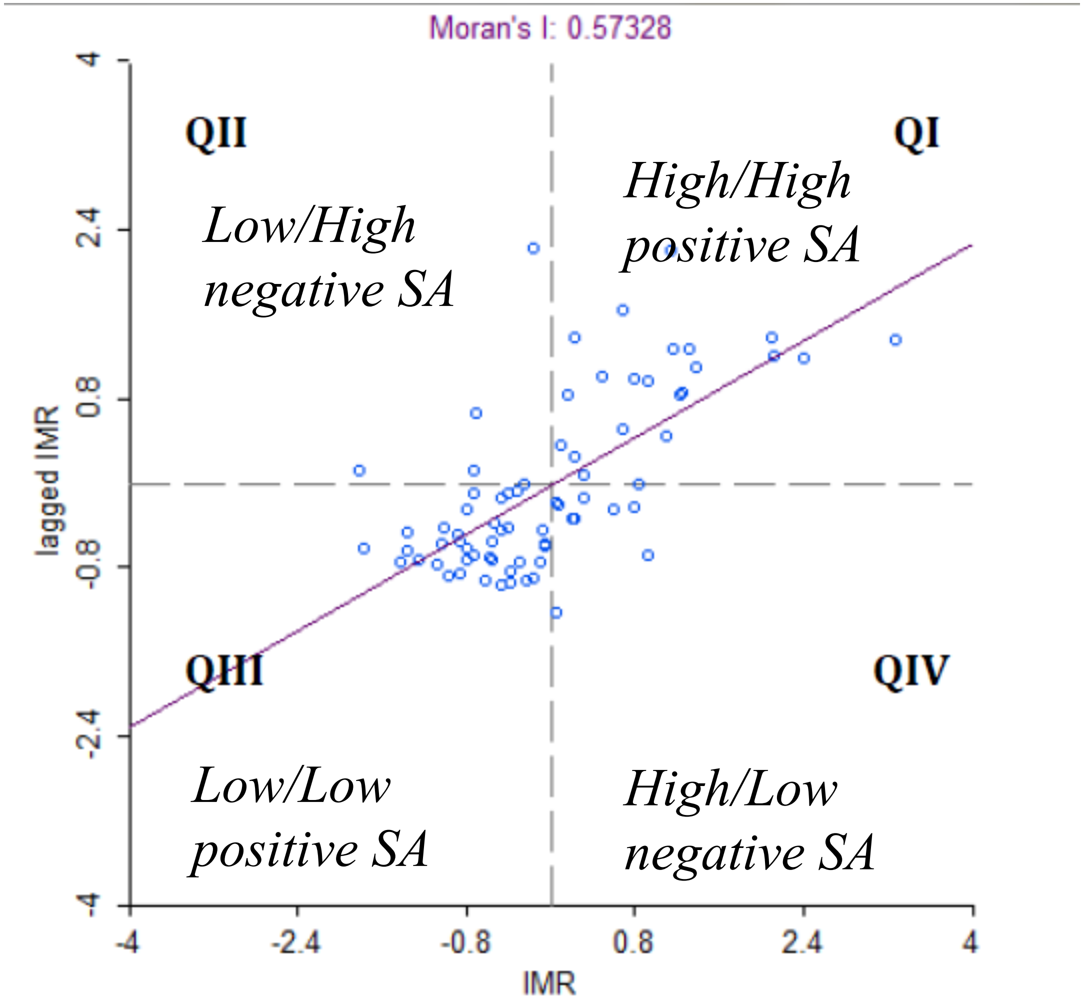
Moran Scatter Plots

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

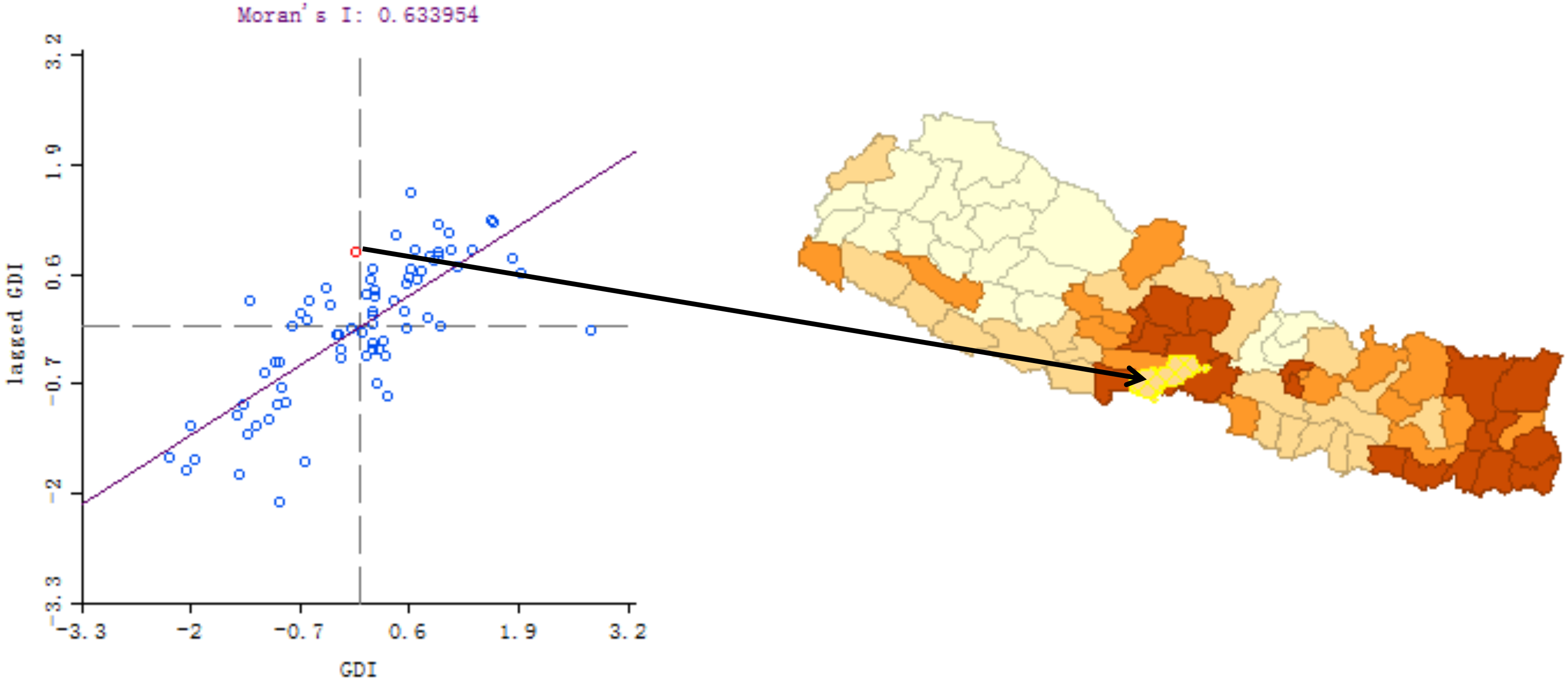


The <u>slope</u> of this *regression line* is Moran's I

Moran Scatter Plots



Moran Scatterplot: Example

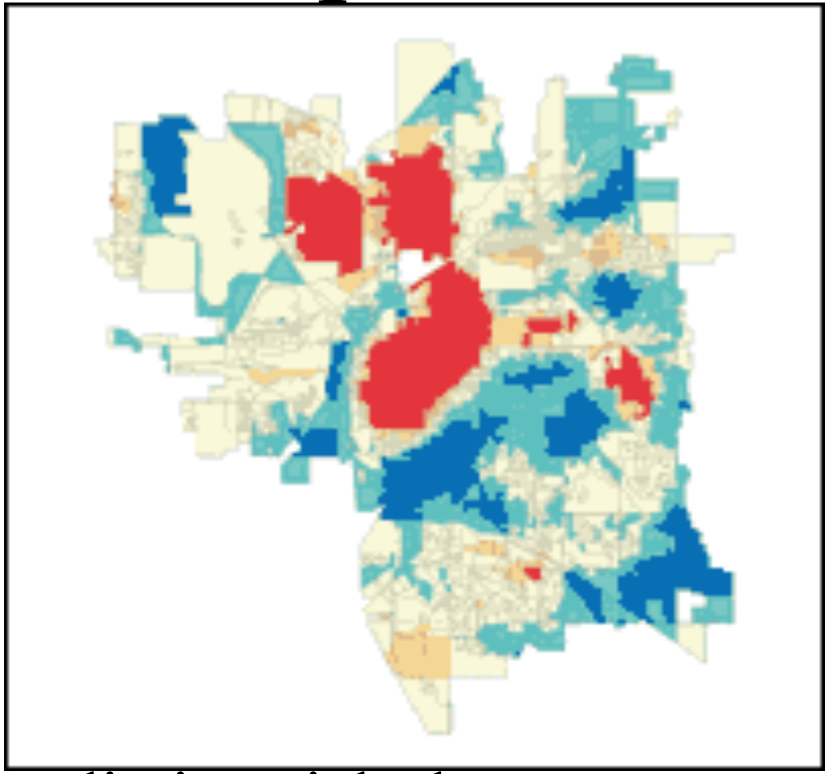


Moran's I for rate-based data

- Moran's I is often calculated for rates, such as crime rates (e.g. number of crimes per 1,000 population) or infant mortality rates (e.g. number of deaths per 1,000 births)
- An adjustment should be made, especially if the denominator in the rate (population or number of births) varies greatly (as it usually does)
- Adjustment is know as the *EB adjust*ment:
  - see Assuncao-Reis *Empirical Bayes Standardization* Statistics in Medicine, 1999
- *GeoDA* software includes an option for this adjustment

Hot Spots and Cold Spots

- What is a *hot spot*?
  - A place where high values cluster together
- What is a *cold spot*?
  - A place where low values cluster together



- Moran's I and Geary's C cannot distinguish them
  - They only indicate clustering
  - Cannot tell if these are hot spots, cold spots, or both

Getis-Ord General/Global G-Statistic

- The G statistic distinguishes between hot spots and cold spots. It identifies *spatial concentrations*.
  - G is relatively large if high values cluster together
  - G is relatively low if low values cluster together
- The General G statistic is interpreted relative to its *expected value*
  - The value for which there is no spatial association
  - G > (larger than) *expected value* ➔ potential "hot spots"
  - G < (smaller than) *expected value* ➔ potential "cold spots"
- A Z *test statistic* is used to test if the difference is statistically significant
- Calculation of G based on a *neighborhood distance* within which cluster is expected to occur

Getis, A. and Ord, J.K. (1992) *The analysis of spatial association by use of distance statistics* Geographical Analysis, *24(3) 189-206*

Comments on General G

- General G will not show negative spatial autocorrelation
- Should only be calculated for ratio scale data
  - data with a "natural" zero such as crime rates, birth rates
- Although it was defined using a contiguity (0,1) weights matrix, any type of spatial weights matrix can be used
  - ArcGIS gives multiple options
- There are two global versions: G and G*
  - G does not include the value of X_i itself, only "neighborhood" values
  - G* includes X_i as well as "neighborhood" values
- Significance test on General G and G* follows the similar procedure as used for Moran's I

Local Measures of Spatial Autocorrelation

Local Indicators of Spatial Association (LISA)

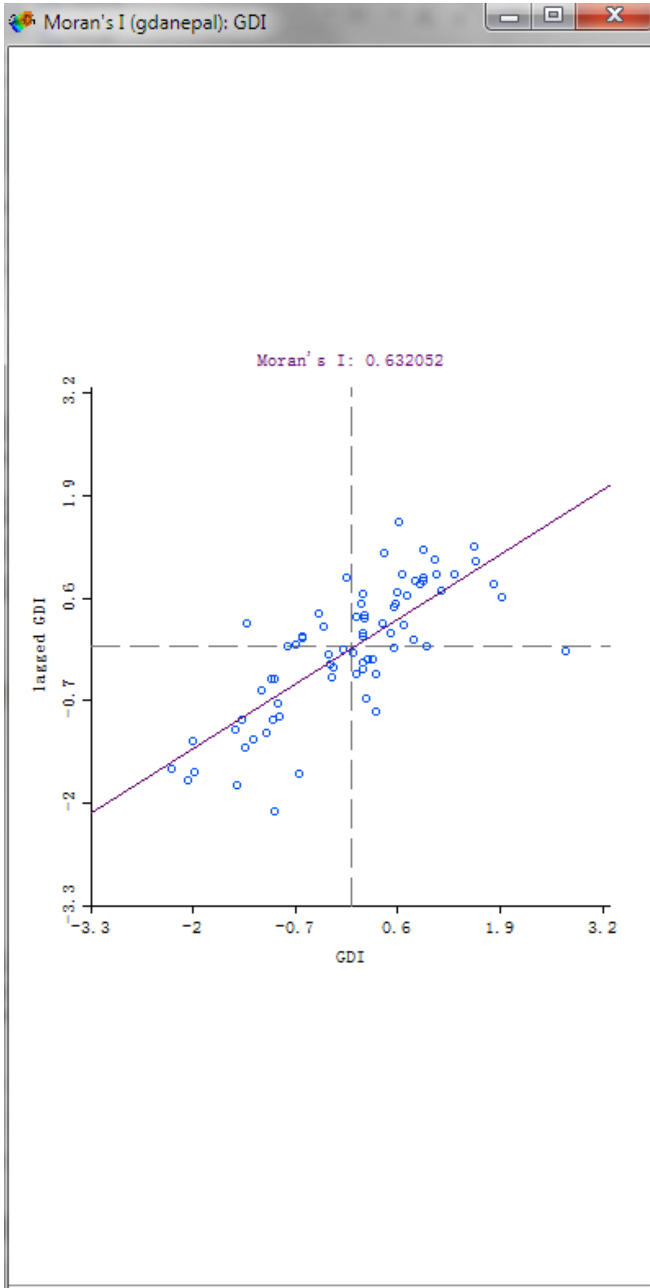
- Local versions of *Moran's I, Geary's C, and the Getis-Ord G statistic*
- Moran's I is most commonly used, and the local version is often called Anselin's LISA, or just LISA

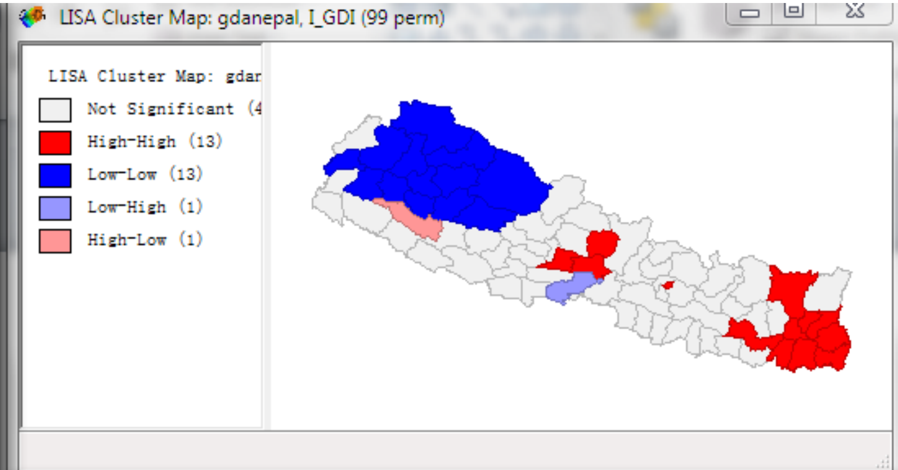
See:
Luc Anselin 1995 *Local Indicators of Spatial Association-LISA* Geographical Analysis 27: 93-115

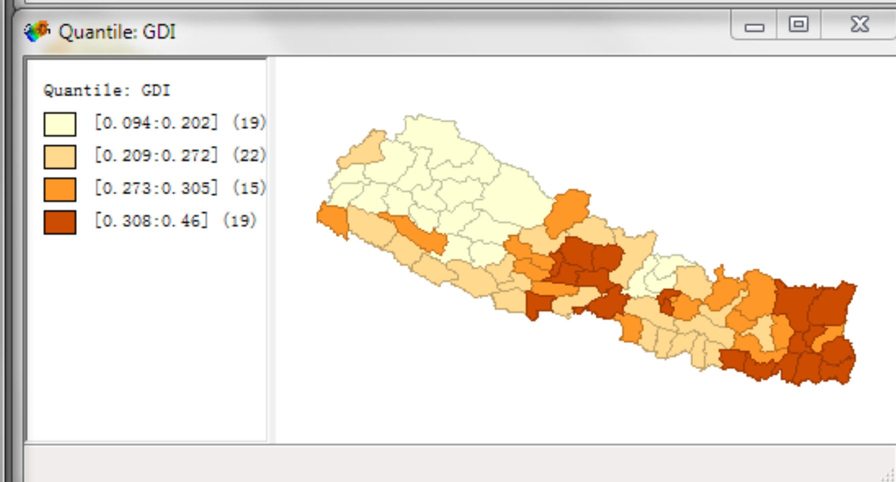
Local Indicators of Spatial Association (LISA)

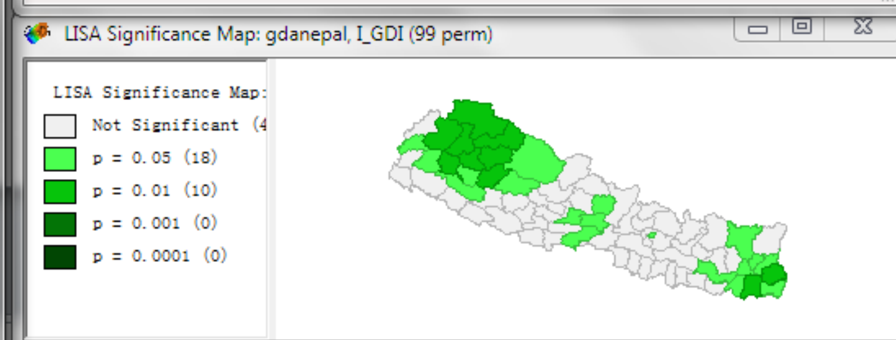
- The statistic is calculated for **each** areal unit in the data
- For each polygon, the index is calculated based on neighboring polygons with which it shares a border
- A measure is available for each polygon, these can be mapped to indicate how spatial autocorrelation varies over the study region
- 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
3.2
1.9
0.6
-0.7
-2
-3.3
-3.3
-2
-0.7
0.6
1.9
3.2


LISA Cluster Map: gdanepal, I_GDI (99 perm)
LISA Cluster Map: gdar
Not Significant (4
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 (4
p = 0.05 (18)
p = 0.01 (10)
p = 0.001 (0)
p = 0.0001 (0)

Calculating Anselin's LISA

- The local Moran statistic for areal unit *i* is:

$$I_i = z_i \sum_j w_{ij} z_j$$

where z_i is the original variable x_i in "standardized form" $z_i = \frac{x_i - \bar{x}}{SD_x}$

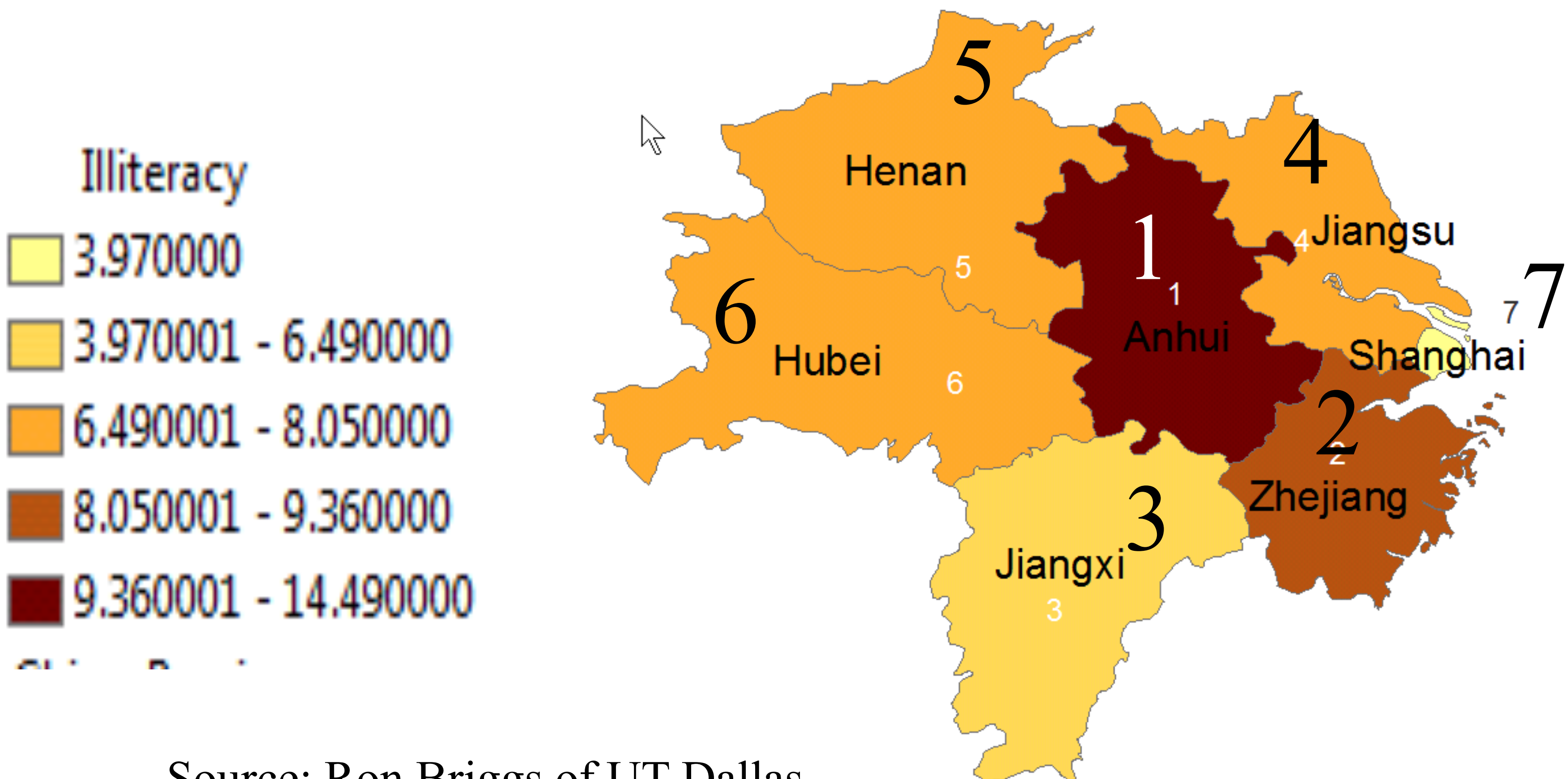
or it can be in "deviation form" $x_i - \bar{x}$

and w_{ij} is the spatial weight

The summation $\sum_j$ is across each <u>row</u> *i* of the spatial weights matrix.

An example follows

Contiguity Matrix		1	2	3	4	5	6	7			
	Code	Anhui	Zhejiang	Jiangxi	Jiangsu	Henan	Hubei	Shanghai	Sum	Neighbors	**Illiteracy**
Anhui	1	0	1	1	1	1	1	0	5	6 5 4 3 2	**14.49**
Zhejiang	2	1	0	1	1	0	0	1	4	7 4 3 1	**9.36**
Jiangxi	3	1	1	0	0	0	1	0	3	6 2 1	**6.49**
Jiangsu	4	1	1	0	0	0	0	1	3	7 2 1	**8.05**
Henan	5	1	0	0	0	0	1	0	2	6 1	**7.36**
Hubei	6	1	0	1	0	1	0	0	3	1 3 5	**7.69**
Shanghai	7	0	1	0	1	0	0	0	2	2 4	**3.97**



Source: Ron Briggs of UT Dallas

Contiguity Matrix and Row Standardized Spatial Weights Matrix

Contiguity Matrix

		1	2	3	4	5	6	7	
	Code	Anhui	Zhejiang	Jiangxi	Jiangsu	Henan	Hubei	Shanghai	Sum
Anhui	1	0	1	1	1	1	1	0	5
Zhejiang	2	1	0	1	1	0	0	1	4
Jiangxi	3	1	1	0	0	0	1	0	3
Jiangsu	4	1	1	0	0	0	0	1	3
Henan	5	1	0	0	0	0	1	0	2
Hubei	6	1	0	1	0	1	0	0	3
Shanghai	7	0	1	0	1	0	0	0	2

1/3

Row Standardized Spatial Weights Matrix

	Code	Anhui	Zhejiang	Jiangxi	Jiangsu	Henan	Hubei	Shanghai	Sum
Anhui	1	0.00	0.20	0.20	0.20	0.20	0.20	0.00	1
Zhejiang	2	0.25	0.00	0.25	0.25	0.00	0.00	0.25	1
Jiangxi	3	0.33	0.33	0.00	0.00	0.00	0.33	0.00	1
Jiangsu	4	0.33	0.33	0.00	0.00	0.00	0.00	0.33	1
Henan	5	0.50	0.00	0.00	0.00	0.00	0.50	0.00	1
Hubei	6	0.33	0.00	0.33	0.00	0.33	0.00	0.00	1
Shanghai	7	0.00	0.50	0.00	0.50	0.00	0.00	0.00	1

Source: Ron Briggs of UT Dallas

Calculating standardized (z) scores

$$z_i = \frac{x_i - \bar{x}}{SD_x}$$

Deviations from Mean and z scores.

	X	X-Xmean	X-Mean2	z
Anhui	14.49	6.29	39.55	2.101
Zhejiang	9.36	1.16	1.34	0.387
Jiangxi	6.49	(1.71)	2.93	(0.572)
Jiangsu	8.05	(0.15)	0.02	(0.051)
Henan	7.36	(0.84)	0.71	(0.281)
Hubei	7.69	(0.51)	0.26	(0.171)
Shanghai	3.97	(4.23)	17.90	(1.414)

Mean and Standard Deviation

Sum	57.41		0.00		62.71
Mean	57.41	/	7	=	8.20
Variance	62.71	/	7	=	8.96
SD	√ 8.96			=	2.99

Source: Ron Briggs of UT Dallas

Calculating LISA

Row Standardized Spatial Weights Matrix

	Code	Anhui	Zhejiang	Jiangxi	Jiangsu	Henan	Hubei	Shanghai
Anhui	1	0.00	0.20	0.20	0.20	0.20	0.20	0.00
Zhejiang	2	0.25	0.00	0.25	0.25	0.00	0.00	0.25
Jiangxi	3	0.33	0.33	0.00	0.00	0.00	0.33	0.00
Jiangsu	4	0.33	0.33	0.00	0.00	0.00	0.00	0.33
Henan	5	0.50	0.00	0.00	0.00	0.00	0.50	0.00
Hubei	6	0.33	0.00	0.33	0.00	0.33	0.00	0.00
Shanghai	7	0.00	0.50	0.00	0.50	0.00	0.00	0.00

w_{ij}

$$I_i = z_i \sum_j w_{ij} z_j$$

z_j

Z-Scores for row Province and its potential neighbors

	Zi	Anhui	Zhejiang	Jiangxi	Jiangsu	Henan	Hubei	Shanghai
Anhui	2.101	***2.101***	0.387	(0.572)	(0.051)	(0.281)	(0.171)	(1.414)
Zhejiang	0.387	2.101	***0.387***	(0.572)	(0.051)	(0.281)	(0.171)	(1.414)
Jiangxi	(0.572)	2.101	0.387	***(0.572)***	(0.051)	(0.281)	(0.171)	(1.414)
Jiangsu	(0.051)	2.101	0.387	(0.572)	***(0.051)***	(0.281)	(0.171)	(1.414)
Henan	(0.281)	2.101	0.387	(0.572)	(0.051)	*(0.281)*	(0.171)	(1.414)
Hubei	(0.171)	2.101	0.387	(0.572)	(0.051)	(0.281)	***(0.171)***	(1.414)
Shanghai	(1.414)	2.101	0.387	(0.572)	(0.051)	(0.281)	(0.171)	***(1.414)***

$w_{ij} z_j$

Spatial Weight Matrix multiplied by Z-Score Matrix (cell by cell multiplication)

	Zi	**Anhui**	**Zhejiang**	**Jiangxi**	**Jiangsu**	**Henan**	**Hubei**	**Shanghai**	**SumWijZj**	**LISA**	**Lisa from GeoDA**
									0.000		
Anhui	**2.101**	-	0.077	(0.114)	(0.010)	(0.056)	(0.034)	-	**(0.137)**	**-0.289**	-0.248
Zhejiang	**0.387**	0.525	-	(0.143)	(0.013)	-	-	(0.353)	**0.016**	**0.006**	0.005
Jiangxi	**(0.572)**	0.700	0.129	-	-	-	(0.057)	-	**0.772**	**-0.442**	-0.379
Jiangsu	**(0.051)**	0.700	0.129	-	-	-	-	(0.471)	**0.358**	**-0.018**	-0.016
Henan	**(0.281)**	1.050	-	-	-	-	(0.085)	-	**0.965**	**-0.271**	-0.233
Hubei	**(0.171)**	0.700	-	(0.191)	-	(0.094)	-	-	**0.416**	**-0.071**	-0.061
Shanghai	**(1.414)**	-	0.194	-	(0.025)	-	-	-	**0.168**	**-0.238**	-0.204

Source: Ron Briggs of UT Dallas

Local Getis-Ord G and G* Statistics

- **Local Getis-Ord G**
  - It is the proportion of all x values in the study area accounted for by the neighbors of location *I*
  - G* will include the self value

$$G_i(d) = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}$$

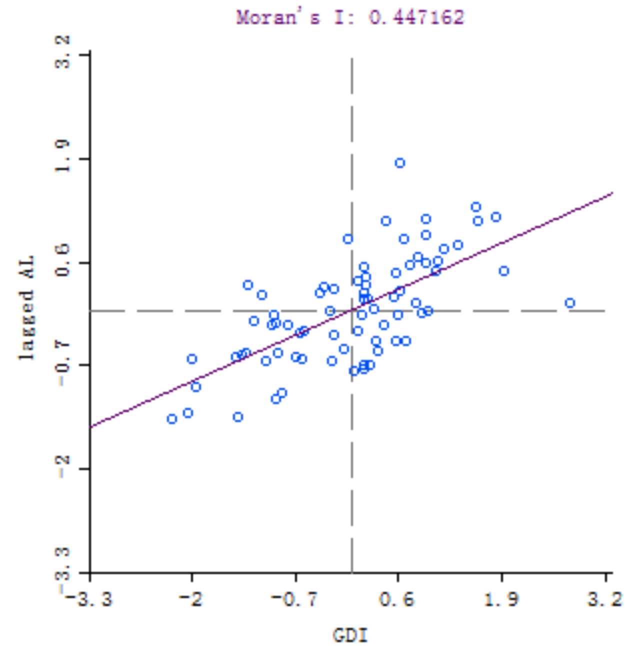
G will be <u>high</u> where <u>high</u> values cluster
G will be <u>low</u> where <u>low</u> values cluster
Interpreted relative to expected value
if randomly distributed.

$$E(G_i(d)) = \frac{\sum_j w_{ij}(d)}{n-1}$$

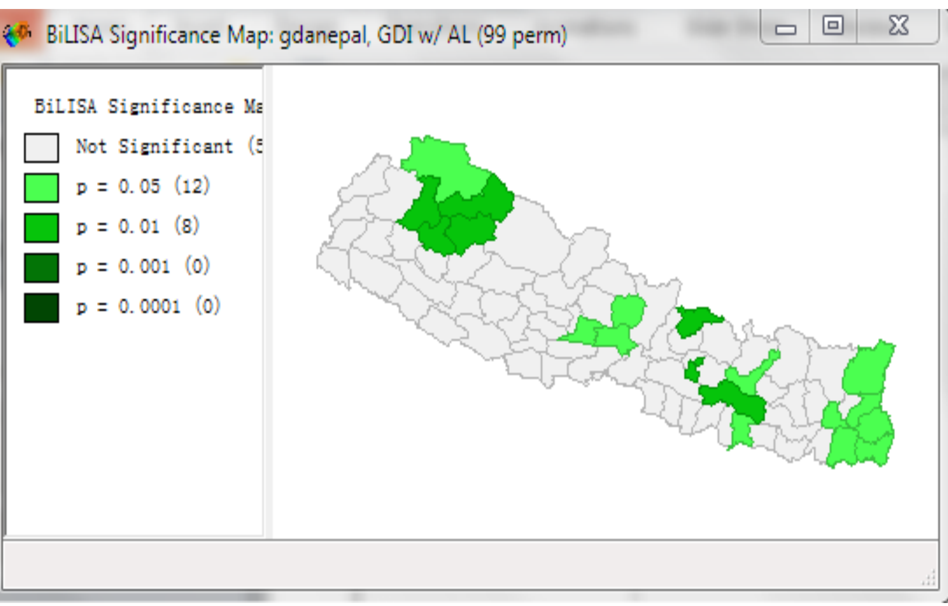
Bivariate LISA

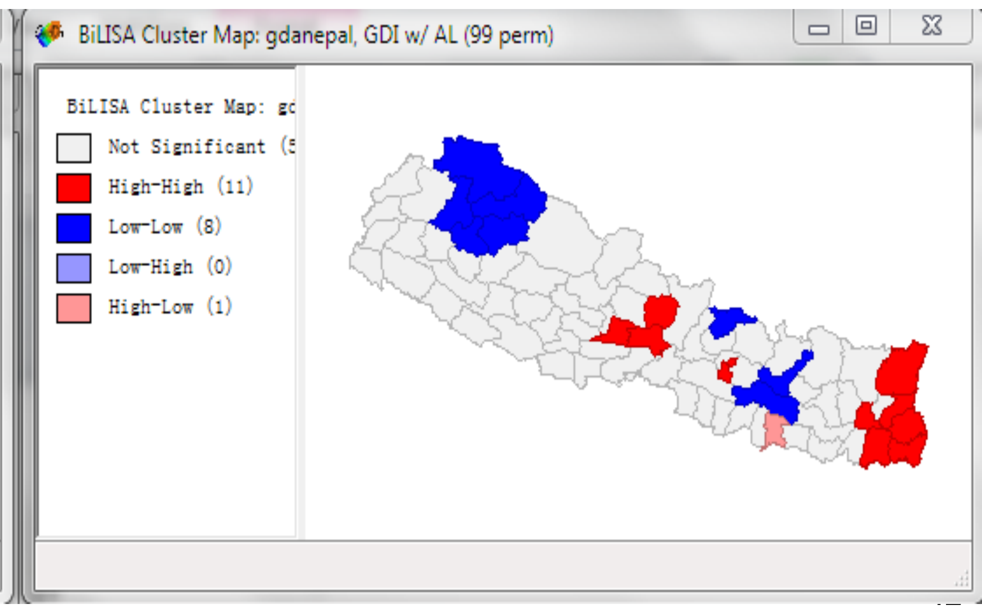
- Moran's I is the correlation between X and Lag-X--the same variable but in nearby areas
  - Univariate Moran's I
- Bivariate Moran's I is a correlation between X and a different variable in nearby areas.

Moran Scatter Plot for GDI vs AL



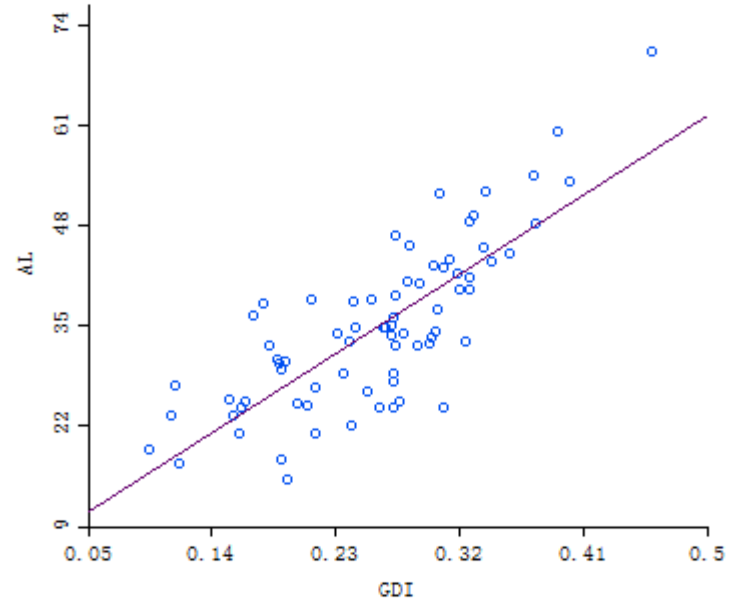
Moran Significance Map for GDI vs. AL

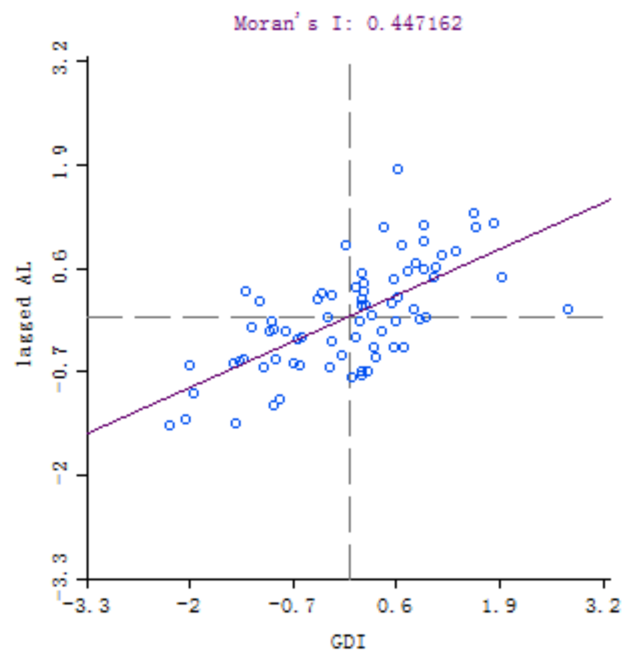




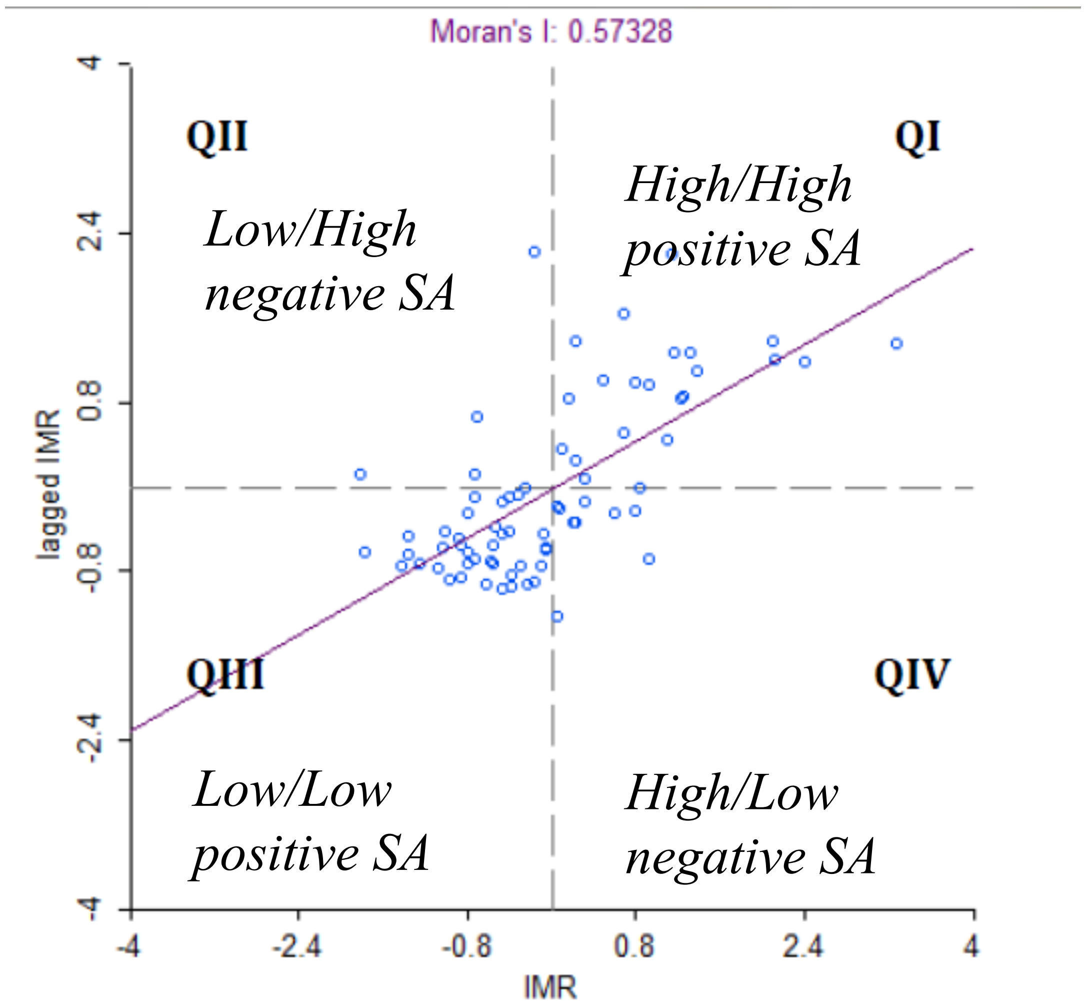
Bivariate LISA
and the Correlation Coefficie

- 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.





Bivariate Moran Scatter Plot



Summary

- Spatial autocorrelation of areal data
- Spatial weight matrix
- Measures of spatial autocorrelation
- Global Measure
  – Moran's I/General G and G*
- Local
  - LISA: Moran's I/General G and G*
  - Bivariate LISA

  – Significance test

- End of this topic