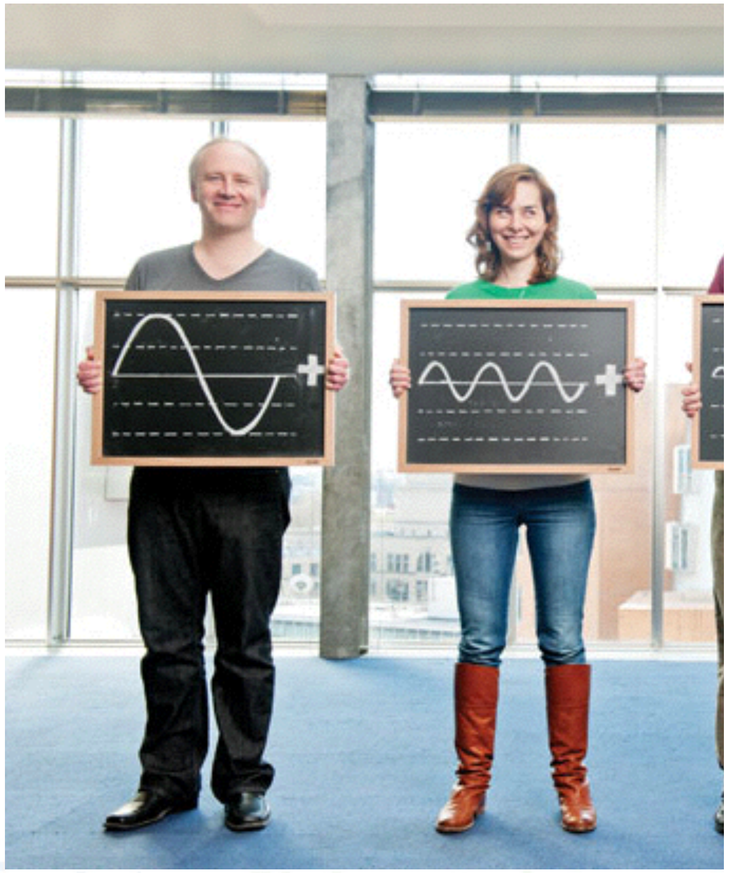
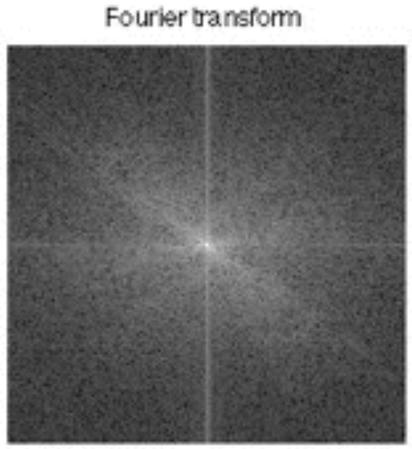
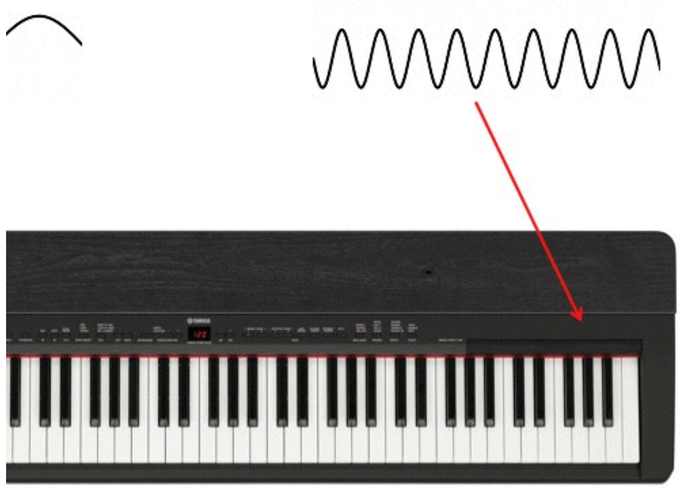
Introduction to Spatial Analysis

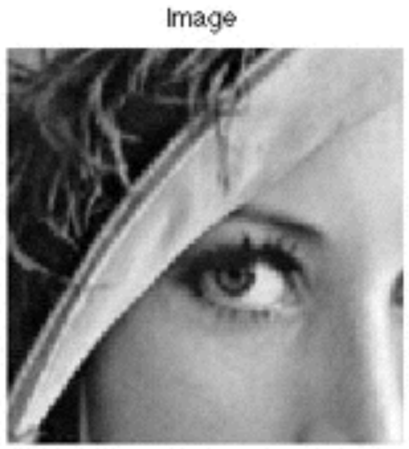
Huy T. Vo
(many slides retrieved and modified from Prof. Briggs
and Dr. Arribas-Bel under CCAS4)

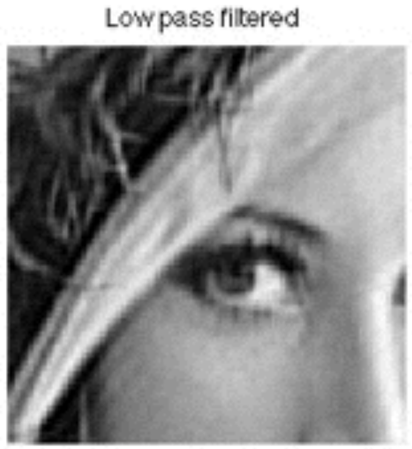
from last time — Fourier Transform











also works in 2D+! (spatial)

[http://www.askamathematician.com]

Today — Spatial Autocorrelation!

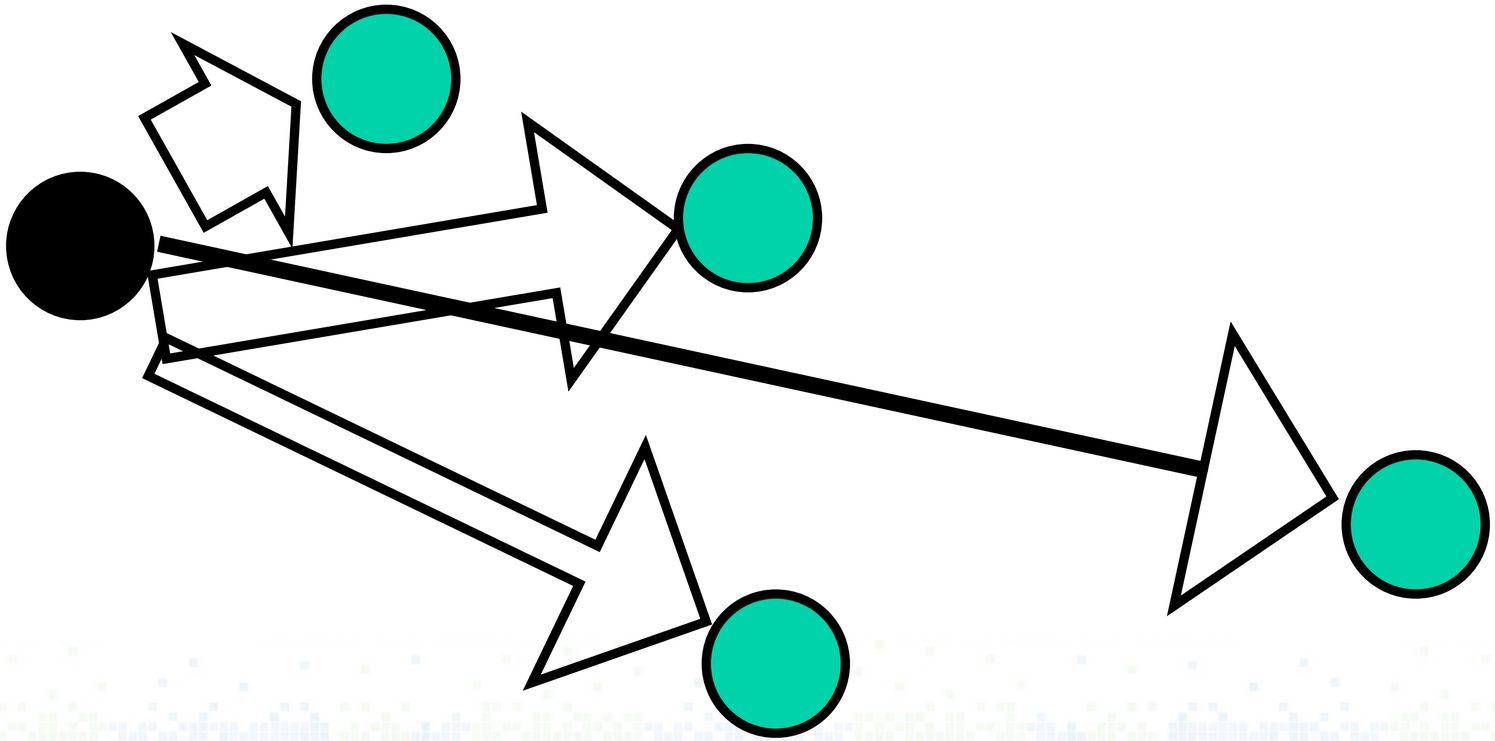
- Some claims "*The Single Most Important Concept in Geography and GIS!*"

Spatial Autocorrelation

- The confirmation of Tobler's first law of geography
  - Everything is related to everything else, but near things are more related than distant things.
- Using similarity
  - The degree to which characteristics at one location are similar (or dissimilar) to those nearby.
- Using probability
  - Measure of the extent to which the occurrence of an event in one geographic area makes more probable, or less probable, the occurrence of a similar event in a neighboring geographic area.
- Using correlation
  - Correlation of a variable with itself through space.
  - The correlation between an observation's value on a variable and the value of near-by observations on the same variable

Spatial Autocorrelation: Tobler's Law

- *Everything is related to everything else, but near things are more related than distant things.*

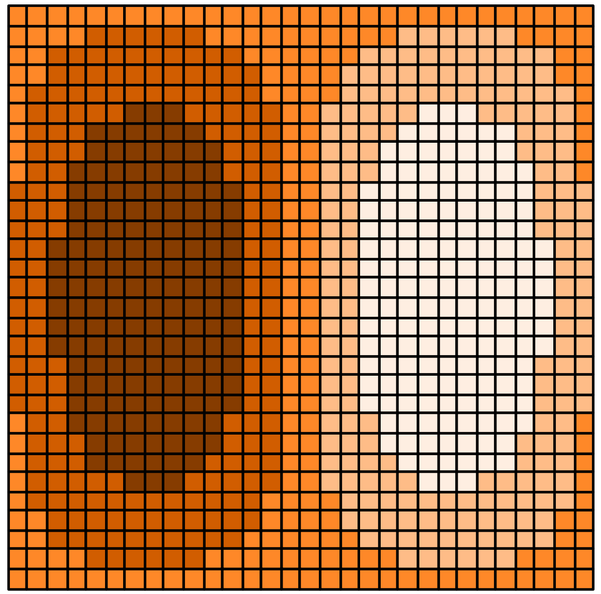


[*Tobler W., (1970) "A computer movie simulating urban growth in the Detroit region". Economic Geography, 46(2): 234-240]

Spatial Autocorrelation — interpretation

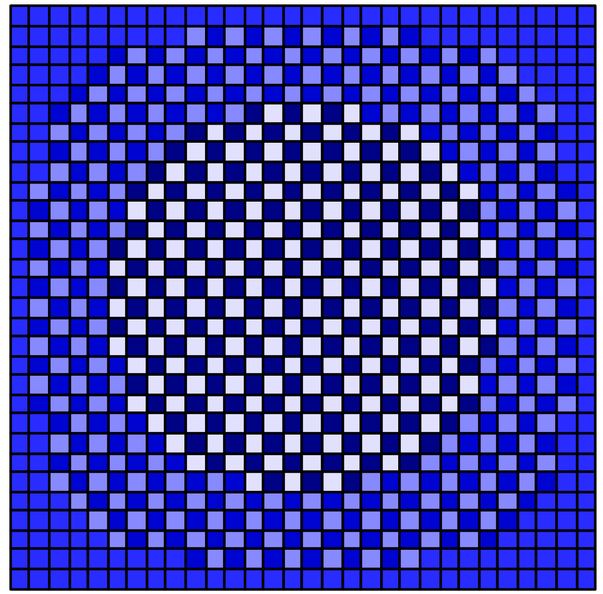
Spatial:
On a map
Auto:
Self
Correlation:
Degree of relative similarity

Positive
Spatial Autocorrelation



Positive: similar values cluster together on a map

Negative
Spatial Autocorrelation



Negative: dissimilar values cluster together on a map

Source: Dr Dan Griffith, with modification

Spatial Autocorrelation: similarity

- The degree to which characteristics at one location are similar to (or different from) those nearby.

Similar to = positive spatial autocorrelation

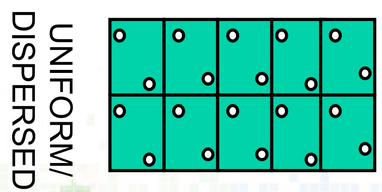
Different from (dissimilar) = negative spatial autocorrelation

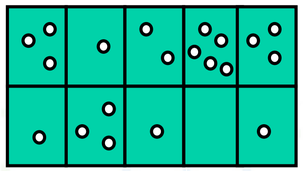
Positive spatial autocorrelation much more common than negative

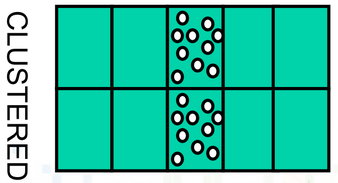
Spatial Autocorrelation: probability

- Measure of the extent to which the occurrence of an event in one geographic unit (polygon) makes more probable, or less probable, the occurrence of a similar event in a neighboring unit.

high negative spatial autocorrelation	no spatial autocorrelation*	high positive spatial autocorrelation
Dispersed Pattern	Random Pattern	Clustered Pattern



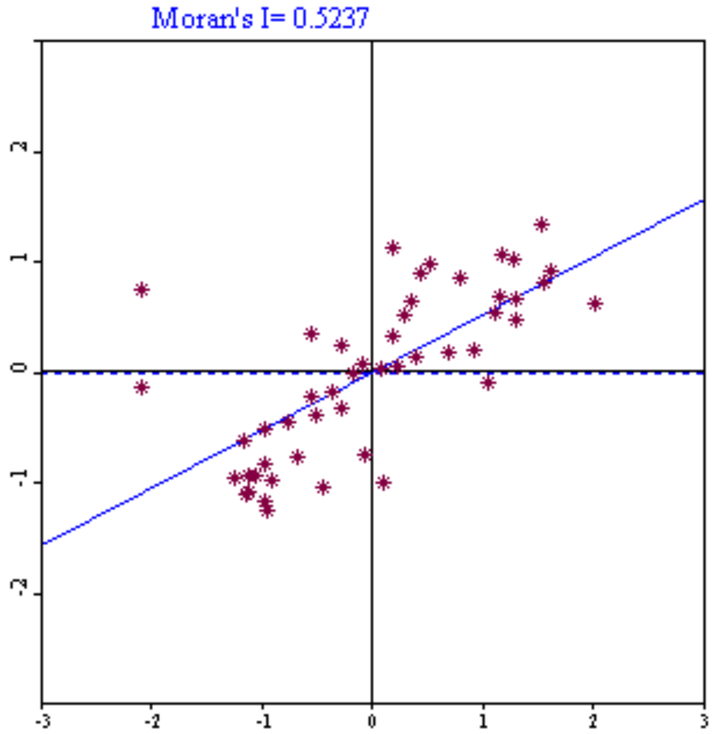




Spatial Autocorrelation: correlation

- The correlation between an observation's value on a variable and the value of nearby observations on the **same** variable.

Crime rate in **nearby** area



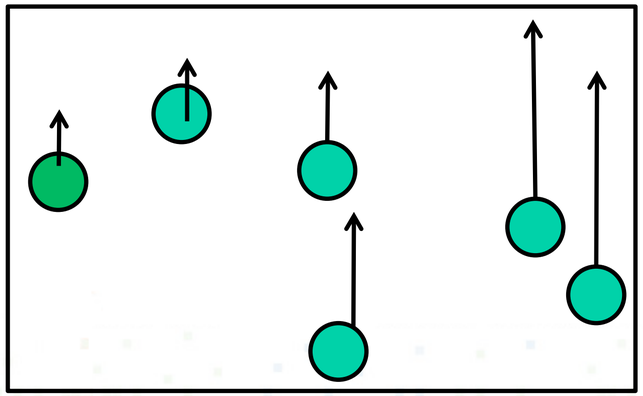
Crime rate in an area

Why is SA important?

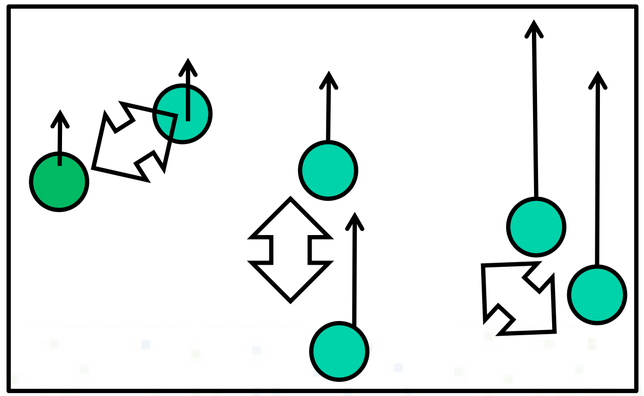
- Because it implies the existence of a spatial process
  - Why are near-by areas similar to each other?
  - Why do high income people live “next door” to each other?
  - These are GEOGRAPHICAL questions.
    - They are about location
- It invalidates most traditional statistical inference tests
  - If SA exists, then the results of standard statistical inference tests may be wrong
  - We need to use spatial statistical inference tests

Why are standard statistical tests wrong?

- Statistical tests are based on the assumption that the values of observations in each sample are independent of one another
- Spatial Autocorrelation violates this
  - samples taken from nearby areas are related to each other and are not independent

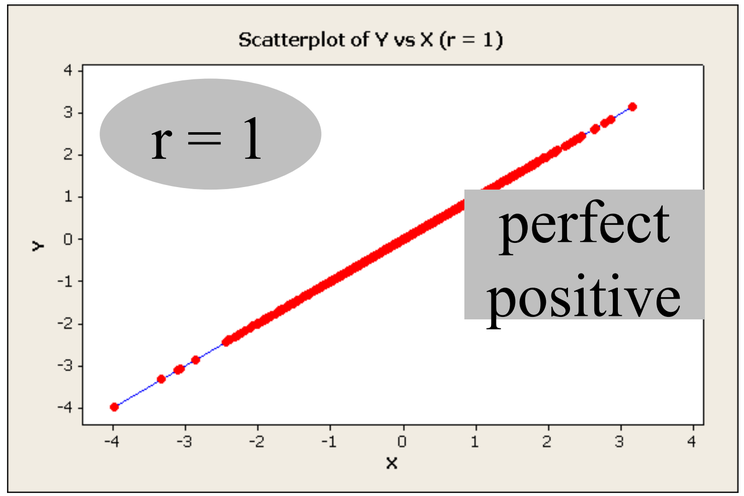


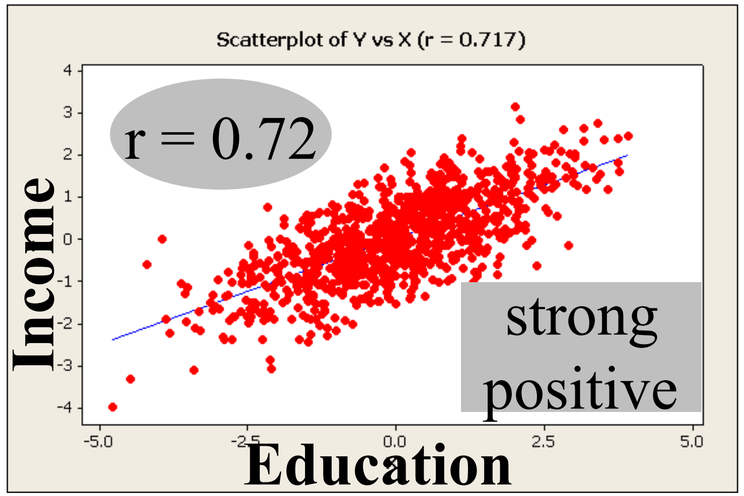
Values near each other are similar in magnitude…

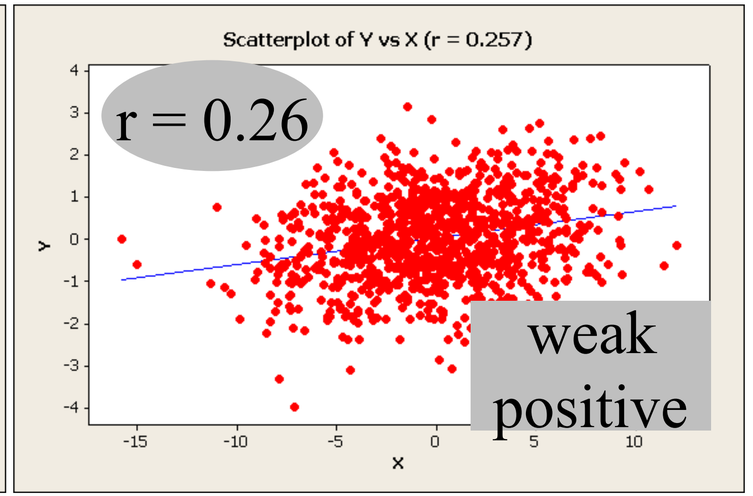


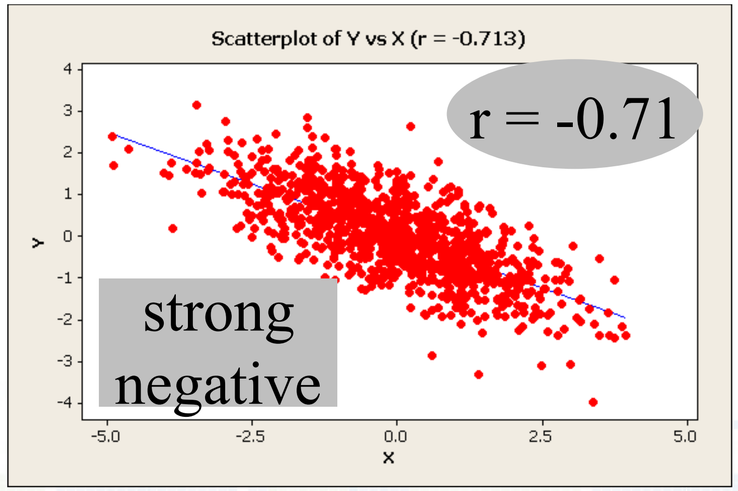
…implies a relationship between nearby observations

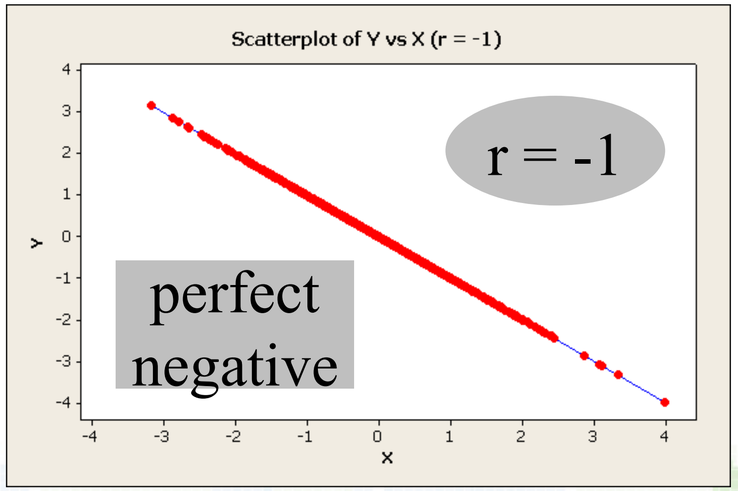
Example of spatial relationship









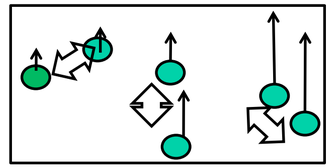


Example of spatial relationship

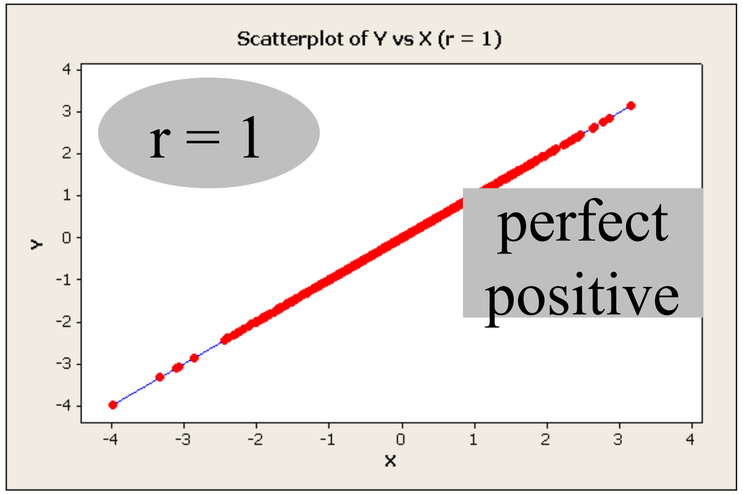
- If Spatial Autocorrelation exists:
  - Correlation coefficients appear to be bigger than they really are, and
  - They are more likely to be found “statistically significant”

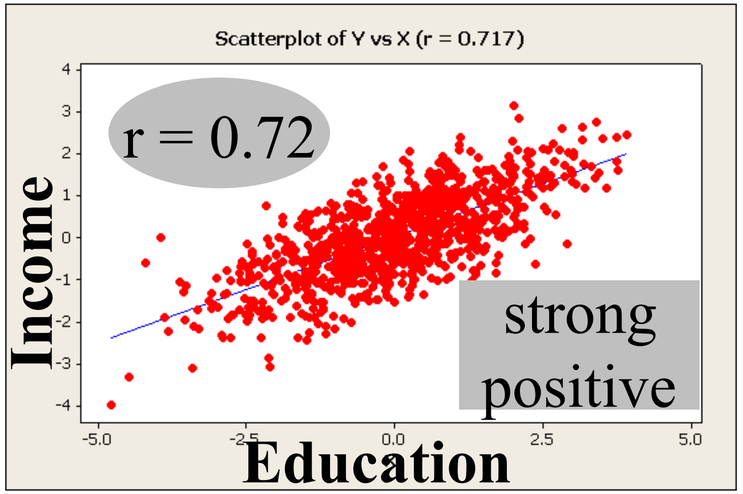
- Consequently:
  - we tend to incorrectly conclude a relationship exists when it does not
  - we believe that the relationship is stronger than it really is

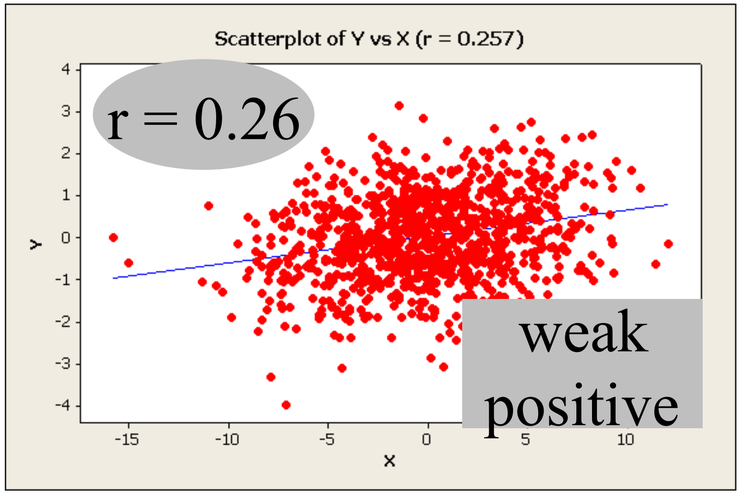
Example of spatial relationship

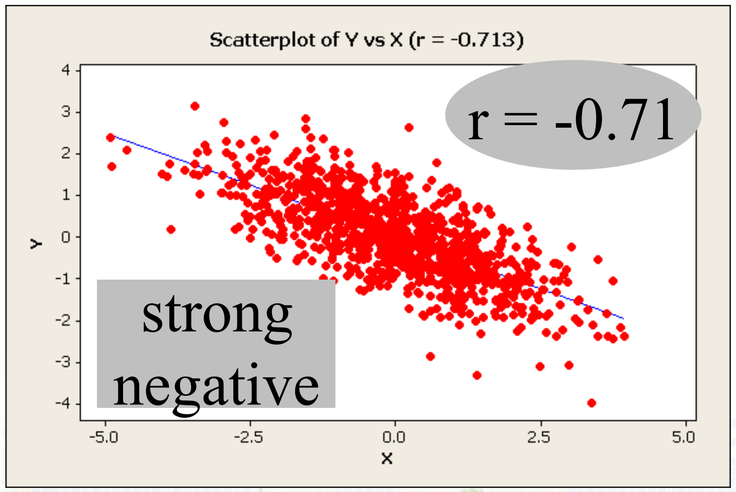


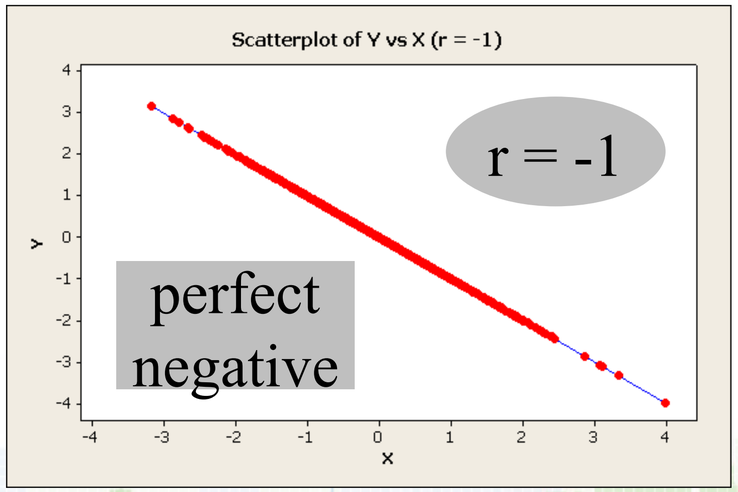
income and education are similar in nearby areas











SA reduces variability, thus, smaller std error

Measuring Spatial Autocorrelation

- First we need to measure the nearness or proximity of a sample:
  - Which points or polygons are “ near” or “next to” other points or polygons?
  - What are the neighborhoods surrounding the West Village?
    - And how each of them influences the neighborhood?

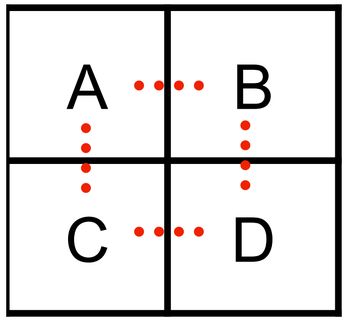
Spatial Weights Matrix — W

- Wij of the spatial weights matrix W measures the relative location of the zone i and j. (e.g. 1 is very near, 0 is very far).
- Different methods of calculating Wij can result in different values for autocorrelation and different conclusions from statistical significance tests!
  - Contiguity-based — binary: 1 for adjacent zones, and 0 otherwise
    - How do we define adjacency?
  - Distance-based — continuous: use distances between zones
    - How do we define points of measurement?

Contiguity-based Weights

- Sharing a border (rook) or boundary (queen)
  - “close” but no common border
  - length of border

4 areal units

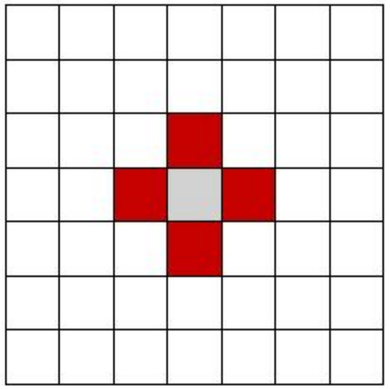


Common border

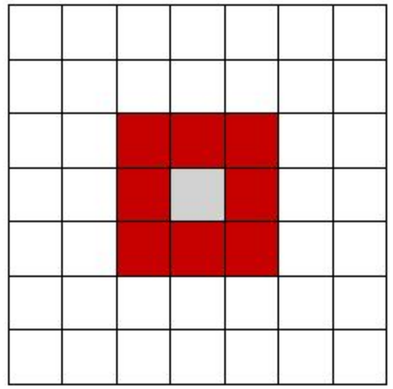
associated geographic connectivity/ weights matrix

4x4 matrix

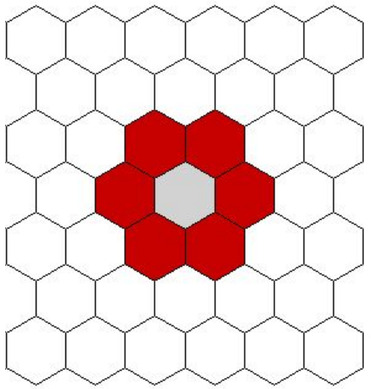
$$\mathbf{W} = \begin{array}{c|cccc} & A & B & C & D \\ A & 0 & 1 & 1 & 0 \\ B & 1 & 0 & 0 & 1 \\ C & 1 & 0 & 0 & 1 \\ D & 0 & 1 & 1 & 0 \end{array}$$



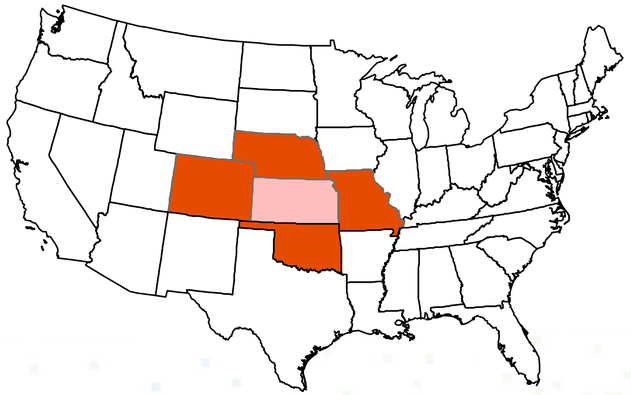
rook



queen



Hexagons

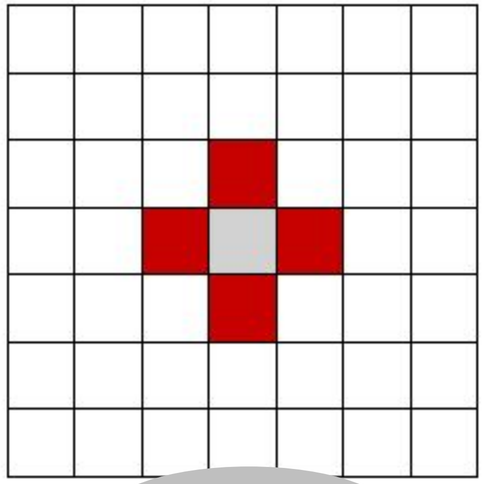


Irregular

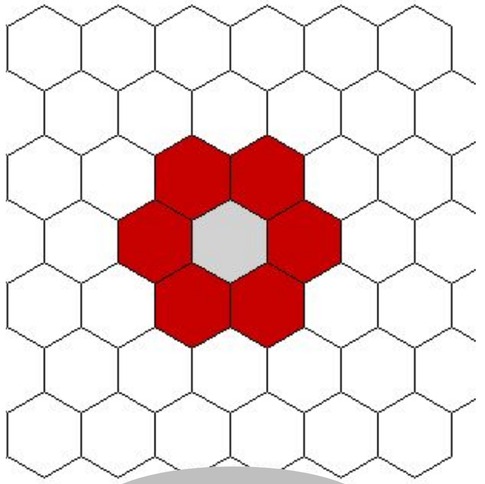
Second-order Contiguity

1st order

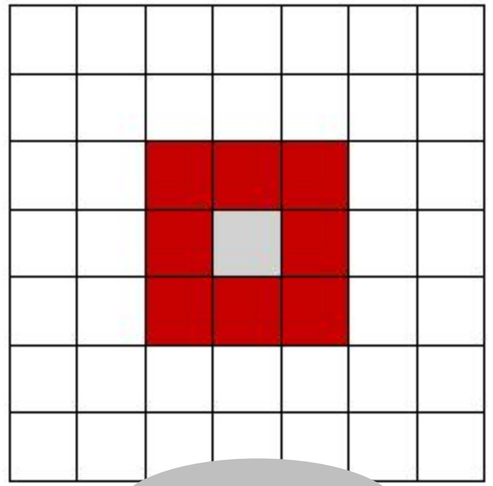
Nearest neighbor



rook



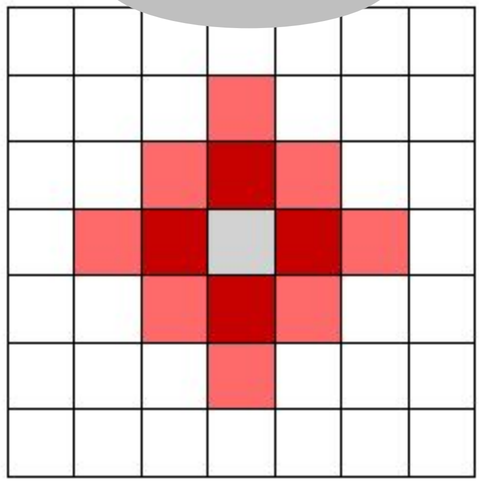
hexagon

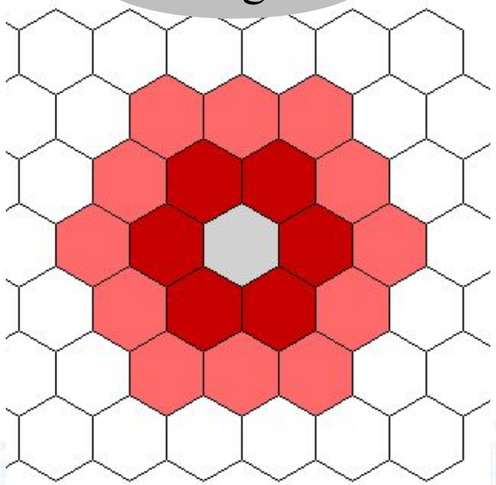


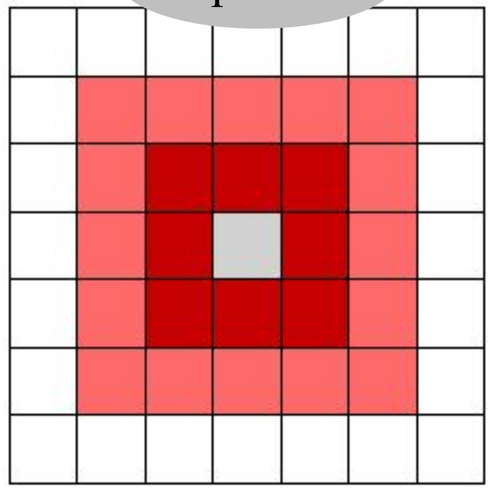
queen

2nd order

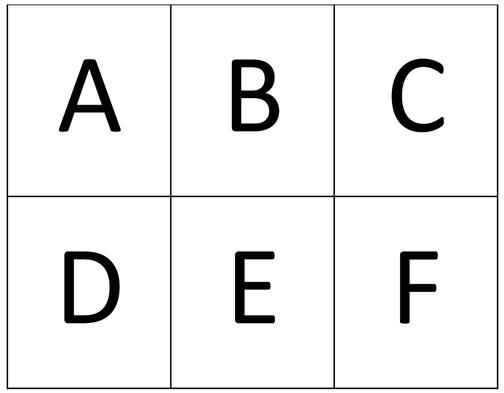
Next nearest neighbor







Row-standardized Contiguity Matrix



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

Distance-based Weights

We want nearness (not distance), weights are the inverse of distances

- 2-D Cartesian distance (for projected data):

$$d_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$$

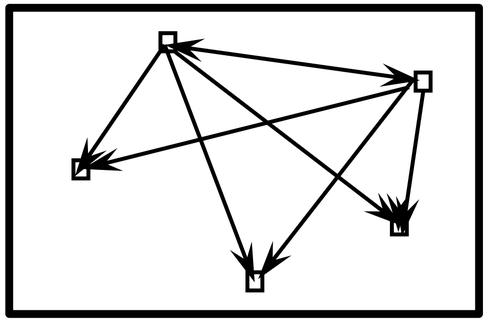
- 3-D Spherical distance via spherical coordinates:

$$d = \text{acos}(\sin \phi 1 \cdot \sin \phi 2 + \cos \phi 1 \cdot \cos \phi 2 \cdot \cos \Delta \lambda) \cdot R$$

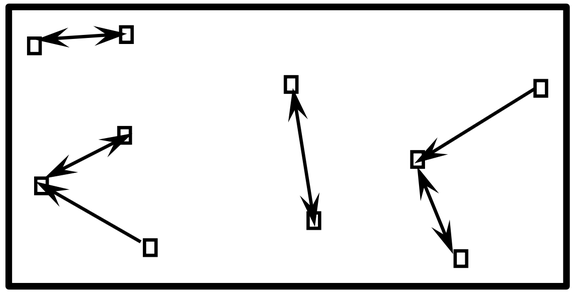
- Or any other methods: city blocks, straight line, etc.

How to assign the weights?

- Include all points/polygon?
  - Matrix could be too big.
  - Zones that are far away are likely not to be related.



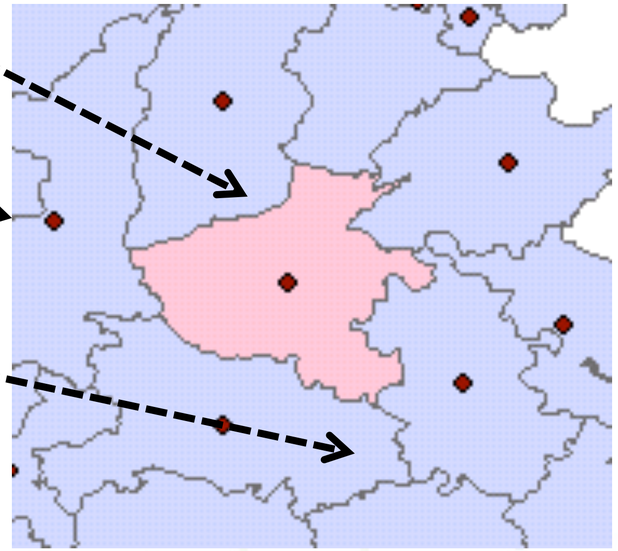
- Include only “n” nearest neighbors
  - Selecting n and the order of contiguity is non-trivial.



- Including zones within a proximity

How to measure polygonal distances?

- distances usually measured centroid to centroid, but
- could be measured from boundary of one polygon to centroid of others
- could be measured between the two closest boundary points
- adjustment required for contiguous polygons since distance for these would be zero



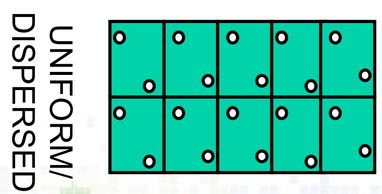
Measuring Spatial Autocorrelation

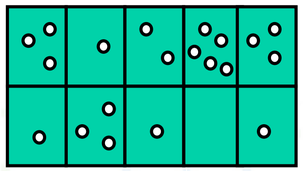
- First we need to measure the nearness or proximity of a sample:
  - Which points or polygons are “ near” or “next to” other points or polygons?
  - What are the neighborhoods surrounding the West Village?
    - And how each of them influences the neighborhood?
- How do we compute and interpret Spatial Autocorrelation?
  - The most common method: ***Moran’s I*** plot and correlation

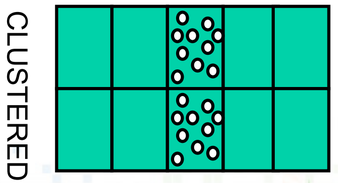
Moran's I

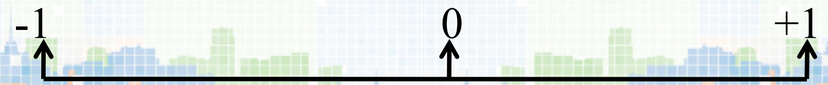
- Use for continuous variables on points and polygons
  - Correlate between a single variable and its "spatial lag"

high negative spatial autocorrelation	no spatial autocorrelation*	high positive spatial autocorrelation
Dispersed Pattern	Random Pattern	Clustered Pattern









Spatial Lag

- A weighted product of neighbor values of a variable X, specify how likely a dependent variable influenced by neighbors

$$X_{sl} = \boldsymbol{W}X$$

$$X_{sl}(i) = \sum_j \boldsymbol{W}_{ij}X_j$$

- Measure that captures the behavior of a variable in the neighborhood of a given observation i
- If $\boldsymbol{W}$ is standardized, the spatial lag is the mean of the variable in the neighborhood

Moran's I

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

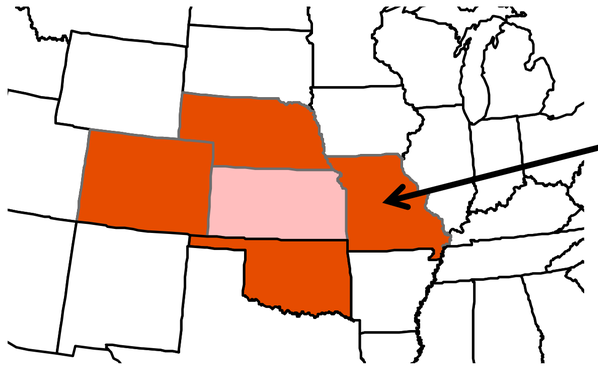
where:

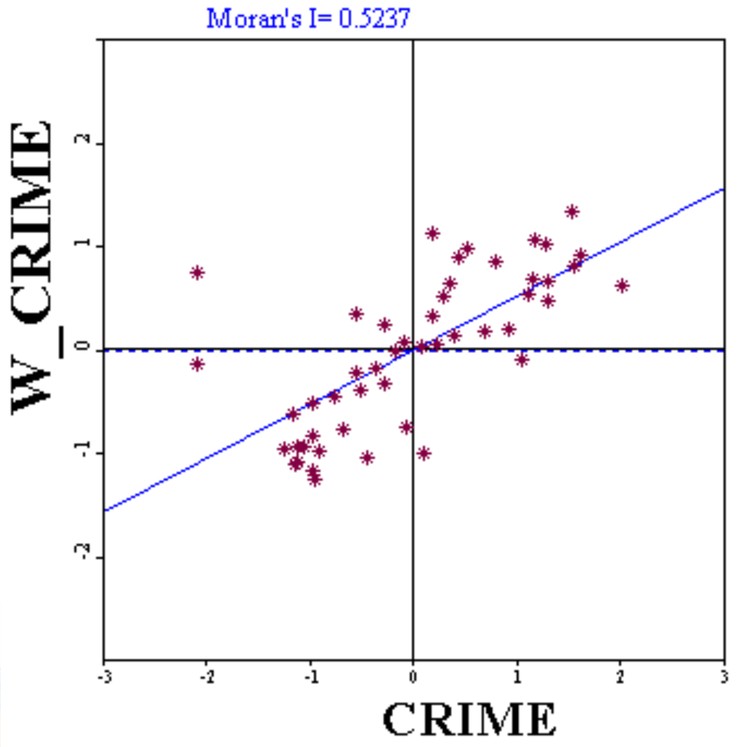
N	is the number of observations (points or polygons)
$\bar{x}$	is the mean of the variable
Xi	is the variable value at a particular location
Xj	is the variable value at another location
Wij	is a weight indexing location of i relative to j

A single value similar to correlation coefficient

Moran Scatterplot

- Scatter plot of a variable X againts its spatial lag
  - Usually, variables are standardized (X-mean(X))/std(X)
    - Observations are well categorized into 4 quadrants
- The slope of the regression line is Moran's I



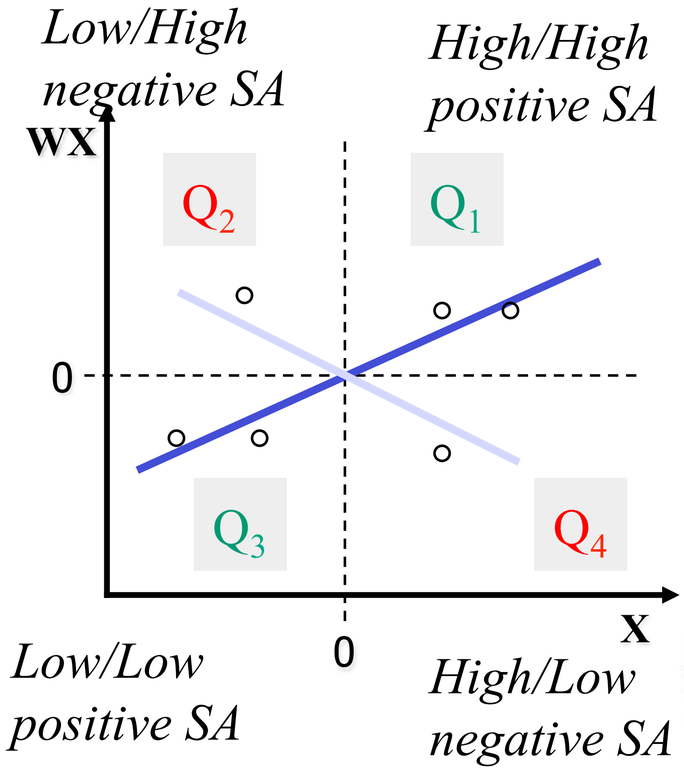


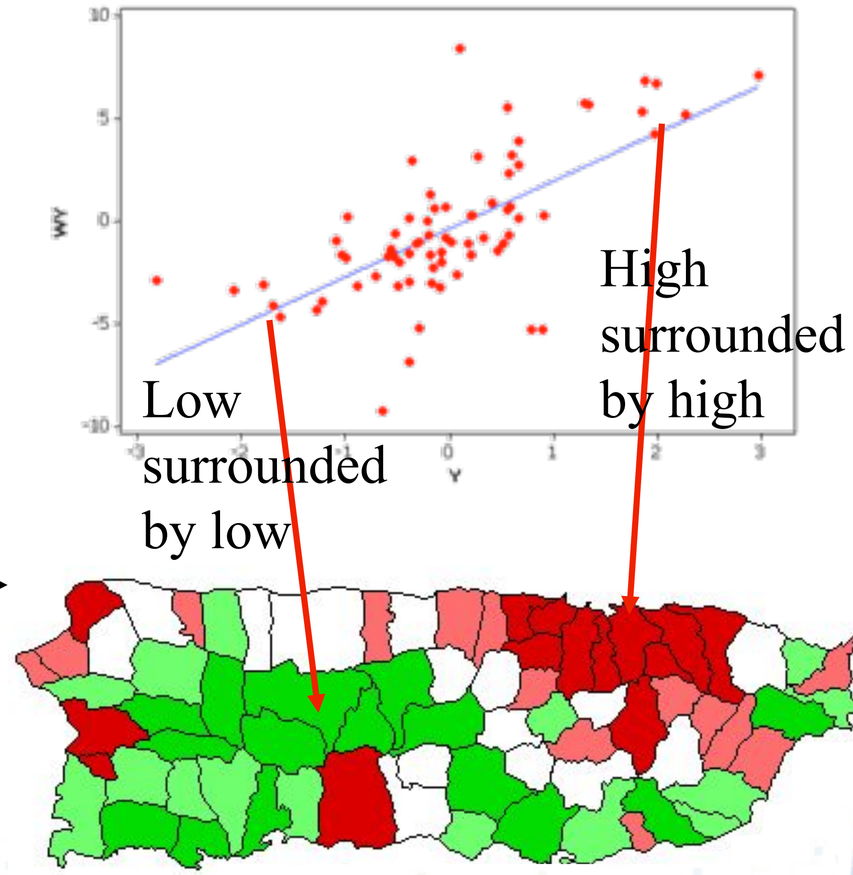
Moran Scatterplot

Locations of positive spatial association
Q1, Q3: "I'm similar to my neighbors"

Locations of negative spatial association
Q2, Q4: "I'm different from my neighbors"

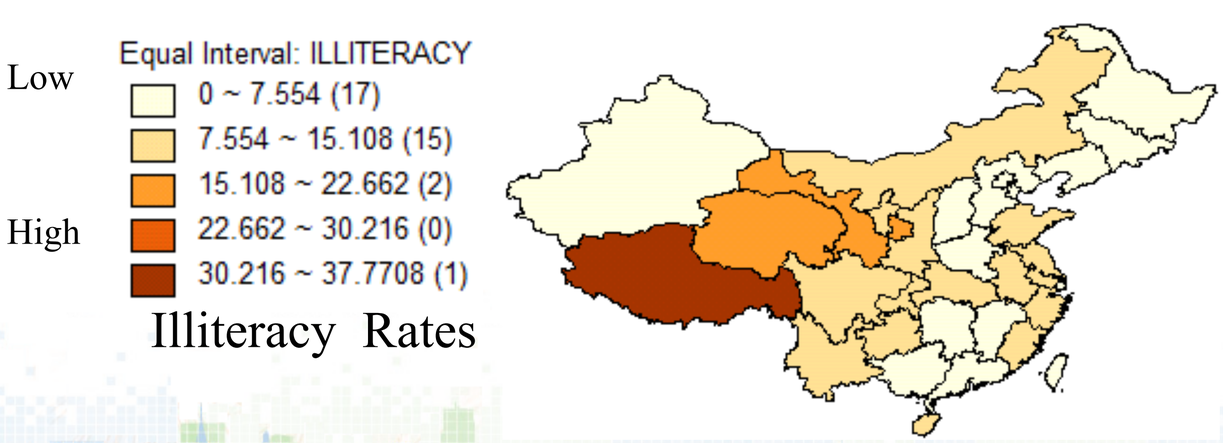
- High-high (Q1): high values are clustered together (hot spots)
- Low-Low (Q3): low values are clustered together (cold spots)

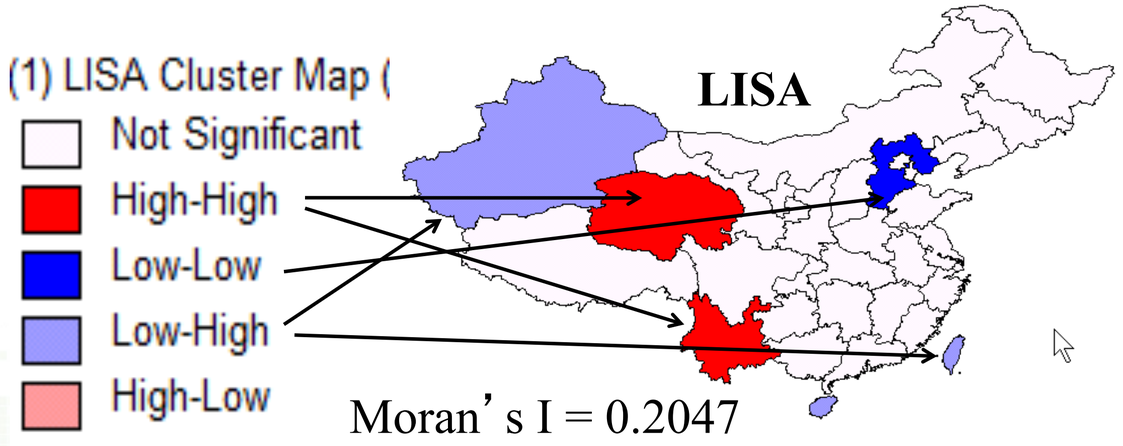




LISA — Local Indicators of Spatial Association

- Moran's I — a single value for the entire data set — global indicator
- LISA — a value is calculated for each observation unit
  - Different patterns may occur in different parts of the region
  - A local version of Moran'I: computing a value for each location





Why is SA important?

- Because it implies the existence of a spatial process
  - Why are near-by areas similar to each other?
  - Why do high income people live “next door” to each other?
  - These are GEOGRAPHICAL questions.
    - They are about location
- It invalidates most traditional statistical inference tests
  - If SA exists, **the results of standard statistical inference tests may be wrong**
  - We need to use spatial statistical inference tests

How to detect the problems?

- For **correlation** — calculate Moran's I for each variable, if I is significant, we may have a problem
- For **regression** — calculate the residuals (differences of actual and predicted values) and:
  - "Ad-hoc"ly plot them on a map: if there are spatial patterns, we may have a problem
  - … or compute Moran's I for the residuals: if it is statistically significant, we may have a problem

What if Spatial Autocorrelation exists?

- Accept that the calculated correlation coefficients may be larger than their true value, and may not be statistically significant
- Include omitted variables that are relevant
- Use a spatial regression model
  - Spatial Autoregressive Models: include the spatial lag/error terms
  - Spatial Filtering
  - Geographically Weighted Regression: multiply by the weights matrix

Wrap-up

- Spatial data can cause problems with standard correlation and regression
- The problems are caused by *spatial autocorrelation*
- We need to use Spatial Regression Models
- Geographers and GIS specialists are experts on spatial data

NYU

CUSP
CENTER FOR URBAN
SCIENCE+PROGRESS

CUNY
The City College
of New York

Thank you!