COA 616 Geostatistics in Environmental Sciences

Lecture 2 Spatial Autocorrelation

Wei Wu

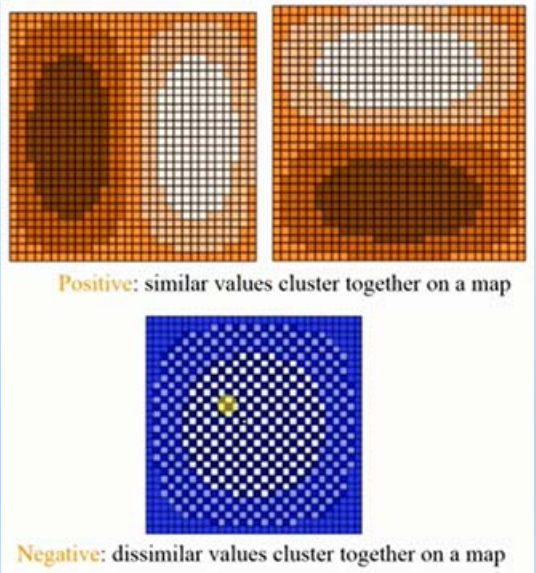
September 18, 2018

Definition

- Tobler's first law of geography

  Everything is related to everything else, but near things are more related than distant things
- Autocorrelation – correlation of a variable with itself
- Spatial autocorrelation – Correlation of a variable with itself across space; Correlation between the same attribute at two locations

  Z(s) is the attribute Z observed in the plane at spatial location s = (x,y), then spatial autocorrelation refers to the correlation between $Z(s_i)$ and $Z(s_j)$ or $Z(x_i, y_i)$ and $Z(x_j, y_j)$.



Why spatial autocorrelation matters

- Spatial autocorrelation is of interest in its own right because it suggests the operation of a spatial process (social economic process, environmental process…).
- Most statistical analyses 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

Why spatial autocorrelation matters cont.

- In ordinary least squares regression (OLS), the regression coefficient will be biased and their precision exaggerated

  - Bias implies regression coefficients may be higher than they really are

  - Exaggerated precision (lower standard error) implies they are more likely to be found "statistically significant".

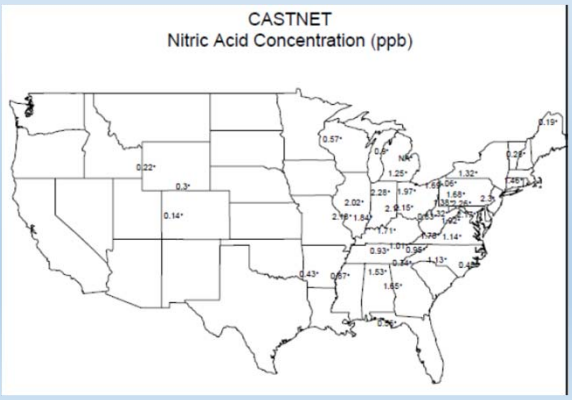
$$\sigma = \frac{s}{\sqrt{n}}$$

$$z = \frac{x - \mu}{\sigma}$$

Types of spatial data

We denote a spatial process in d dimensions as $\{Z(s): s \in D \subset R^d\}$

1) Geostatistical data: The domain D is continuous and fixed set.



Examples:

- Weekly concentrations of Nitric Acid in the US
- Weekly concentrations of ozone in the US
- Annual acidic deposition in the US

Types of spatial data cont.

2) Lattice data (regional data): The domain D is fixed and discrete. They are spatially aggregated over areal regions.

Examples:

Remote sensing data (pixel)

US Census bureau data (census tract)

Number of deaths, crimes reported for counties or zip codes

3) Point patterns: The domain D is random. Point patterns arise when the important variable to be analyzed is the location of events.

Examples:

Locations of rare species

Location at which weeds emerge in a garden

Locations of birds' nests in a suitable habitat – evidence of territoriality?

Location of disease

Steps in determining spatial autocorrelation
Lattice data

- Choose a neighborhood criterion – Which areas are linked?
- Assign weights to the areas that are linked – Create a spatial weights matrix
- Run statistical test, using weights matrix, to examine spatial autocorrelation

Moran's I

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

Where n is the number of regions (sample size), i and j are locations , w_{ij} is a measure of spatial proximity between locations i and j, X is the only random variable.

Compared to Pearson correlation coefficient:

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

Modifying r by adding w_{ij} (binary indicator – neighbors or not), and replacing y with x, then we get:

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

Interpretation of Moran's I

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

Most powerful test statistics for spatial autocorrelation.

Pairs of regions where both regions exhibit above-average scores (or below average scores) will contribute positive terms to the numerator.

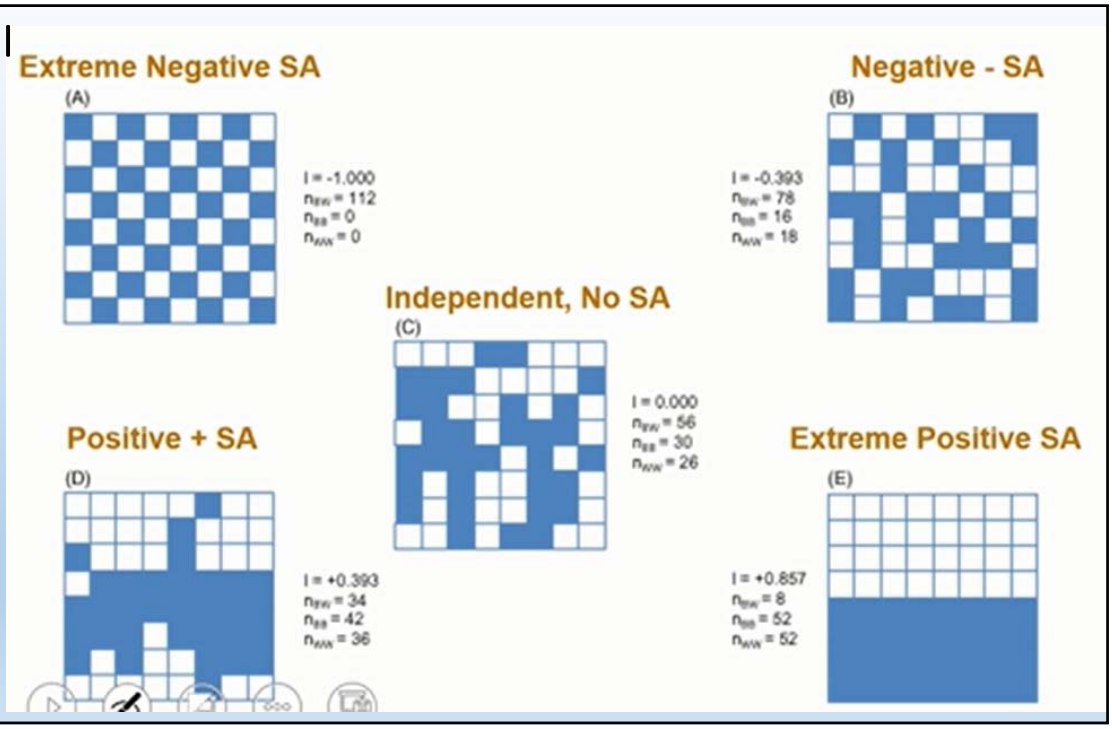
Pairs of regions where one region exhibit above average scores and the other below average scores will contribute negative terms to the numerator.

Values near +1 indicate a strong positive spatial autocorrelation (high values located close to high values, and low values located close to low values)

Values near -1 indicate a strong negative spatial autocorrelation (high values located close to low values)

Values near 0 indicate no spatial autocorrelation (no clustering, no uniformity, random pattern)

Moran's I



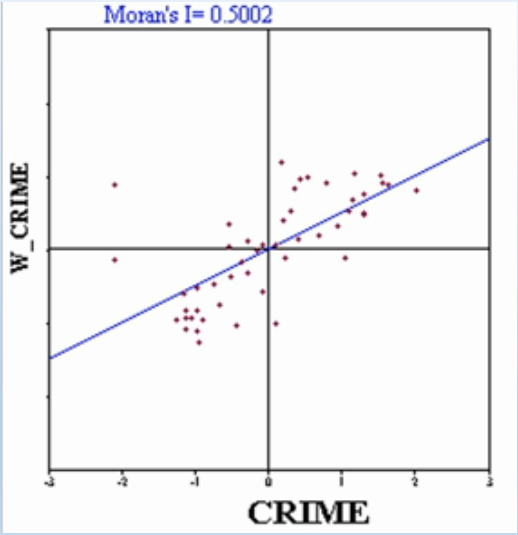
Moran's I scatterplot

Analogous to correlation

Horizontal axis – Standardized X
Vertical axis – Neighborhood average of standardized X

Moran's I is imply the correlation between standardized X and the neighborhood average of standardized X.

Assess the direction (slope) and strength of spatial autocorrelation (how far away these point clouds are from the best fitted line)



Moran's I – Hypothesis testing

Null hypothesis:

- Spatial randomness
- Values observed at one location doe not depend on values observed at neighboring locations
- Observed spatial pattern of values is equally likely as any other spatial pattern
- The location of values may be altered without affecting the information content of the data

Null hypothesis:

- Moran's I is normally distributed

$$Z_{test} = \frac{I - E(I)}{\sqrt{Var(I)}}$$

- E(I) – Expected I when X is totally random.
- The test statistics can easily be compared to critical Z-scores so that significance can be determined.
- The variance term in this equation is very complicated.

Moran's I – Hypothesis testing

The expected value of Moran's I under hypothesis of no spatial autocorrelation is

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

Its variance equals

$$Var(I) = \frac{NS_4 - S_3 S_5}{(N-1)(N-2)(N-3)(\sum_i \sum_j w_{ij})^2}$$

where

$$S_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2$$

$$S_2 = \frac{\sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2}{1}$$

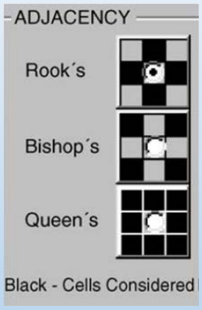
$$S_3 = \frac{N^{-1} \sum_i (x_i - \bar{x})^4}{(N^{-1} \sum_i (x_i - \bar{x})^2)^2}$$

$$S_4 = \frac{(N^2 - 3N + 3)S_1 - NS_2 + 3(\sum_i \sum_j w_{ij})^2}{1}$$

$$S_5 = S_1 - 2NS_1 + \frac{6(\sum_i \sum_j w_{ij})^2}{1}$$

The W matrix

W = $\{w_{ij}\}$ is the matrix that defines the level of spatial connectivity between locations on the map



Neighborhoods can be defined based on

- Continuity (common boundary, 1st order, 2nd order)
- Distance

W matrix can be measured using

- Binary connectivity based on contiguity
  w_{ij} = 1 if regions i and j are contiguous (neighbors), otherwise w_{ij} = 0
- Defined as a function of the distance between i and j
  inverse distance: $w_{ij} = d_{ij}^{-\beta}$
  negative exponential: $w_{ij} = \exp(-\beta d_{ij})$

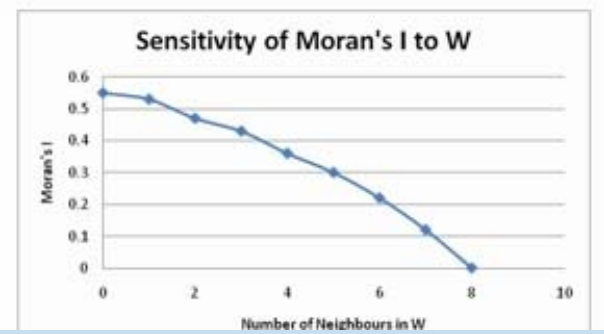
Row-normalized – the total influence of all neighbors is equal to 1: $w_{ij}^{*} = w_{ij} / \sum_{j=1}^{n} w_{ij}$

The W matrix cont.

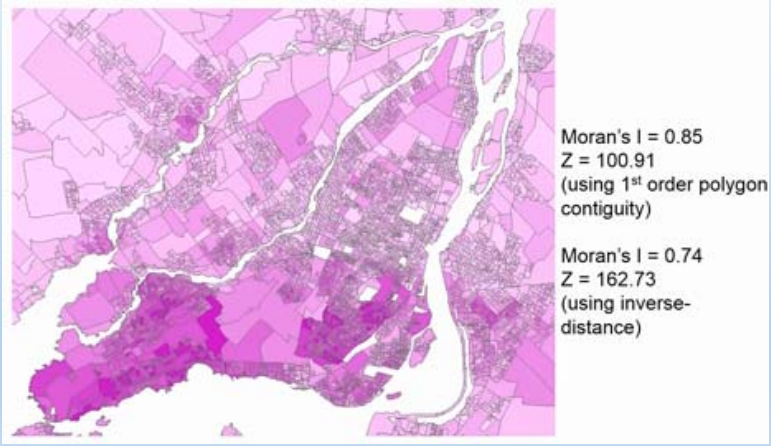
There are not many good theories to define W matrix

The value of Moran's I is very sensitive to the choice of W matrix, so sensitivity analysis is performed and Moran's I is calculated for various W matrices.

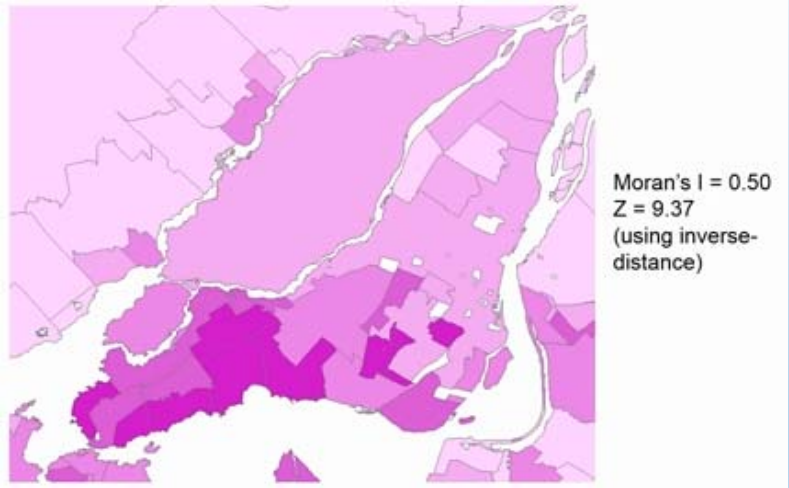
Increase of spatial size of neighborhood, spatial autocorrelation tends to decrease. So spatial autocorrelation is scale-dependent!



Moran's I is sensitive to choice of W matrix

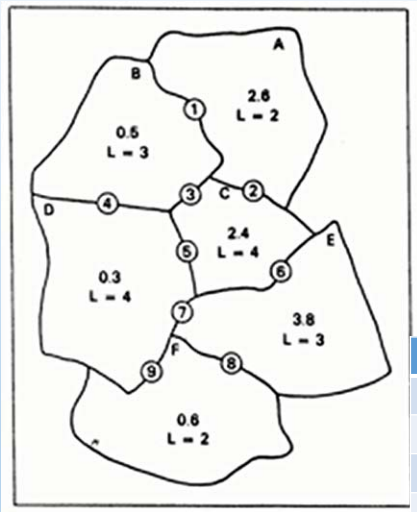


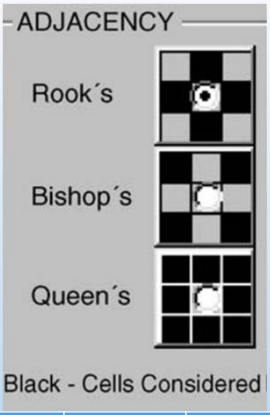
Moran's I is more sensitive to spatial scale (aggregation of data)



Calculation of Moran's I

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = 18$$





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

Calculating Moran's I

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

	X	d=$(X - \bar{X})$	$(X - \bar{X})^2$
A	2.6	0.9	0.81
B	0.5	-1.2	1.44
C	2.4	0.7	0.49
D	0.3	-1.4	1.96
E	3.8	2.1	4.41
F	0.6	-1.1	1.21

$\bar{X}$=1.7 $\Sigma_i(X_i - \bar{X})^2$=10.32

Calculating Moran's I

$$I = \left(\frac{6}{18}\right)\frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{10.32}$$

i,j	1	2	3	4	5	6
1	0	d_1d_2	d_1d_3	0	0	0
2	d_2d_1	0	d_2d_3	d_2d_4	0	0
3	d_3d_1	d_3d_2	0	d_3d_4	d_3d_5	0
4	0	d_4d_2	d_4d_3	0	d_4d_5	d_4d_6
5	0	0	d_5d_3	d_5d_4	0	d_5d_6
6	0	0	0	d_6d_4	d_6d_5	0

Where: $d_i = (X_i - \bar{X})$

Calculating Moran's I

$$I = \left(\frac{6}{18}\right)\frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{10.32}$$

i,j	1	2	3	4	5	6
1	0	-1.08	0.63	0	0	0
2	-1.08	0	-0.84	1.68	0	0
3	0.63	-0.84	0	-0.98	1.47	0
4	0	1.68	-0.98	0	-2.94	1.54
5	0	0	1.47	-2.94	0	-2.31
6	0	0	0	1.54	-2.31	0

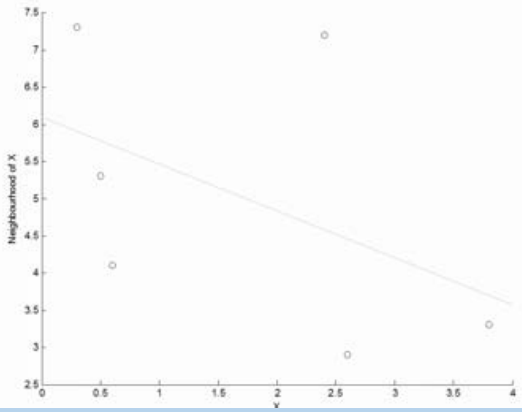
→ $\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X}) = -5.66$

Calculating Moran's I

$$I = \left(\frac{6}{18}\right)\frac{-5.66}{10.32}$$

$$I = -\left(\frac{1}{3}\right)0.55$$

$$I = -0.183$$



→ Some weak negative spatial autocorrelation

$$z = \frac{I - E(I)}{\sqrt{Var(I)}} \qquad E(I) = -\frac{1}{n-1} = -0.2$$

Local statistics

Used to detect the locations of clusters of high and low values.

Why use local statistics
- Allow to test if clustering occurs near a specific target location (species near upwelling zone)
- Allow to discover new locations of significant local clustering
- Significant local clusters may exist even though global tests may not find a significant level of global autocorrelation
- Global tests may find significant autocorrelation even due to several extremely clustered locations

Local Moran's I

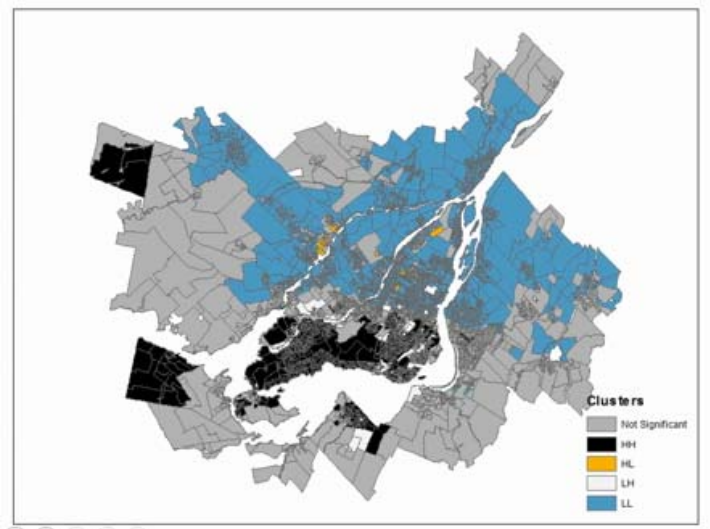
$$I_i = \frac{n(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \sum_j w_{ij}(x_j - \bar{x})$$

The sum of all the local moran's I's sum up to a value proportionately equal to the global Moran's I.

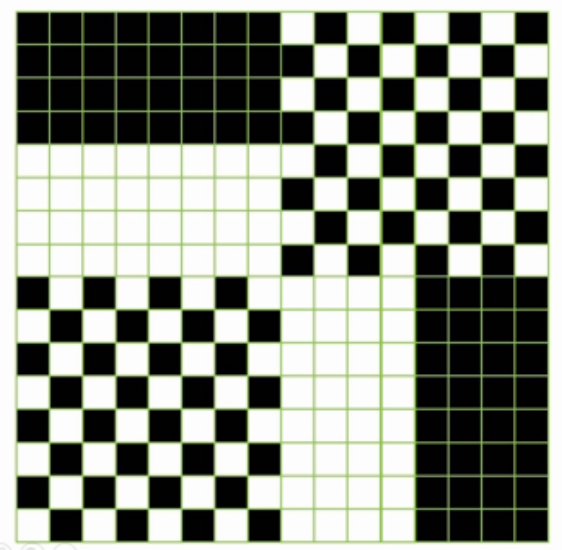
The local Moran's I is normally distributed and locations of significant positive and negative autocorrelation can be mapped out.

Zones can be categorized into 5 types:
1. Insignificant: Not a significant local cluster
2. High-High: A high location that clusters with other high locations (+)
3. Low-Low: A low location that clusters with other low locations (+)
4. Low-High: A low location that clusters with high locations (-)
5. High-Low: A high location that clusters with low locations (-)

Local Moran's I



Global Moran's I misinterpret spatial pattern



Geary ratio (Geary's C)

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

Geary's C typically ranges from 0 to 3 – cannot be negative
An uncorrelated process has an expected C = 1
Values less than 1 indicate positive spatial autocorrelation
Values greater than 1 indicate negative spatial autocorrelation

Geary ratio (Geary's C)

GR relates more directly to the semivariogram plot of geostatistics than does MC

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

Highlighting negative relation between Geary's C and Moran's I – inversely related.
GR captures all of the locational information summarized by Moran's I.
Each squared deviation is weighted by the number of neighbors it has. For irregular partitionings, number of neighbors could differ much, resulting in an overemphasis of such deviation by the GR spatial autocorrelation index for area units with relatively large number of neighbors.
Similar values clustered: C -> 0 and I -> 1
Dissimilar values clustered: C -> 3 and I -> -1

Examples

1. North Carolina

SIDR79 is the death rate per 1000 (1979-84) from sudden infant death syndrome

2. Boston house price

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 ft2
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOX	Nitrogen oxide concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk-0.63)2 where Bkis the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's