Lecture 1: Fundamentals of Spatial Statistics

Description versus Inference

- Description and descriptive statistics
  - Concerned with obtaining summary measures to describe a set of data
- Inference and inferential statistics
  - Concerned with making inferences from samples about populations
  - Concerned with making legitimate inferences about underlying processes from observed patterns

We will be looking at both!

Classic Descriptive Statistics: Univariate

Measures of Central Tendency and Dispersion

- Central Tendency: single summary measure for one variable:
  - mean (average)
  - median (middle value)
  - mode (most frequently occurring)
  - geometric mean (more robust, less sensitive)
- Dispersion: measure of spread or variability
  - Variance
  - Standard deviation (square root of variance)

Formulae for mean

$$\overline{X} = \sum_{i=1}^{N} \frac{X_i}{N}$$

Formulae for variance

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{N} \quad = \frac{\sum_{i=1}^{n} X_i^2 - [(\sum X)^2/N]}{N}$$

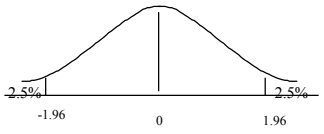
These may be obtained in ArcGIS by:
--opening a table, right clicking on column heading, and selecting *Statistics*
--going to *ArcToolbox>Analysis>Statistics>Summary Statistics*

Classic Descriptive Statistics: Univariate

Frequency distributions

A counting of the frequency with which values occur on a variable

- Most easily understood for a categorical variable (e.g. ethnicity)
- For a continuous variable, frequency can be:
  - calculated by dividing the variable into categories or "bins" (e.g income groups)
  - represented by the proportion of the area under a frequency curve



In ArcGIS, you may obtain frequency counts on a categorical variable via:
--*ArcToolbox>Analysis>Statistics>Frequency*

Classic Descriptive Statistics: Bivariate

Pearson Product Moment Correlation Coefficient (r)

- Measures the degree of association or strength of the relationship between two continuous variables
- Varies on a scale from –1 thru 0 to +1

  -1 implies perfect negative association
  - As values on one variable rise, those on the other fall (price and quantity purchased)

  0 implies no association

  +1 implies perfect positive association
  - As values rise on one they also rise on the other (house price and income of occupants)

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n \quad S_x S_y}$$

Where Sx and Sy are the standard deviations of X and Y, and X and Y are the means.

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{N}} \qquad S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{N}}$$

Classic Descriptive Statistics: Bivariate

Calculation Formulae for Pearson Product Moment Correlation Coefficient (r)

Correlation Coefficient example using "calculation formulae"

$$r = \frac{\dfrac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{X} \bullet \overline{Y}}{S_x S_y},$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n} x^2}{n} - \overline{X}^2}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{n} y^2}{n} - \overline{Y}^2}.$$

As we explore spatial statistics, we will see many analogies to the mean, the variance, and the correlation coefficient, and their various formulae

Inferential Statistics: Are differences real?

- Frequently, we lack data for an entire *population* (all possible occurrences) so most measures (statistics) are estimated based on *sample* data
  - *Statistics* are measures calculated from samples which are *estimates* of *population parameters*
- the question must always be asked if an observed difference (say between two statistics) could have arisen due to chance associated with the sampling process, or reflects a real difference in the underlying population(s)
- Answers to this question involve the concepts of *statistical inference and statistical hypothesis testing*
- Although we do not have time to go into this in detail, it is always important to explore before any firm conclusions are drawn.
- However, never forget: **statistical significance does not always equate to scientific (or substantive) significance**
  - With a big enough sample size (and data sets are often large in GIS), statistical significance is often easily achievable

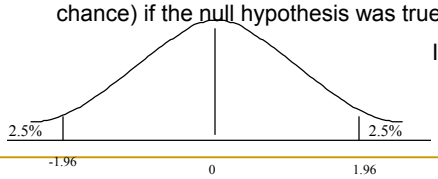
Statistical Hypothesis Testing: Classic Approach

Statistical hypothesis testing usually involves 2 values; don't confuse them!

- A measure(s) or index(s) derived from samples (e.g. the mean center or the Nearest Neighbor Index)
  - We may have two sample measures (e.g. one for males and another for females), or a single sample measure which we compare to "spatial randomness"
- A *test statistic*, derived from the measure or index, whose probability distribution is known when repeated samples are made,
  - this is used to test the *statistical significance* of the measure/index

We proceed from the null hypothesis (H_o) that, in the population, there is "no difference" between the two sample statistics, or from spatial randomness*

- If the test statistic we obtain is very unlikely to have occurred (less than 5% chance) if the null hypothesis was true, the null hypothesis is rejected

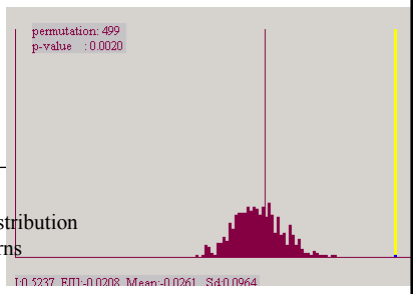


If the test statistic is beyond +/- 1.96 (assuming a Normal distribution), we reject the null hypothesis (of no difference) and assume a statistically significant difference at at least the 0.05 significance level.

Statistical Hypothesis Testing: Simulation Approach

- Because of the complexity inherent in spatial processes, it is sometime difficult to derive a legitimate test statistic whose probability distribution is known
- An alternative approach is to use the computer to simulate multiple random spatial patterns (or samples)--say 100, the spatial statistic (e.g. NNI or LISA) is calculated for each, and then displayed as a frequency distribution.
  - This simulated sampling distribution can then be used to assess the probability of obtaining our observed value for the Index if the pattern had been random.

Our observed value:
--highly unlikely to have occurred if the process was random
--conclude that process is not random



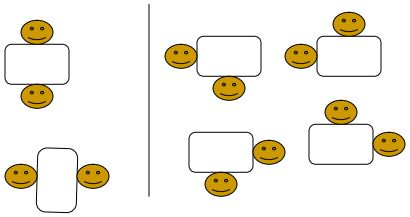
Empirical frequency distribution from 499 random patterns ("samples")

*This approach is used in Anselin's **GeoDA** software*

Is it Spatially Random? *Tougher than it looks to decide!*

- Fact: It is observed that about twice as many people sit catty/corner rather than opposite at tables in a restaurant
  - Conclusion: psychological preference for nearness

- In actuality: an outcome to be expected from a random process: *two ways to sit opposite, but four ways to sit catty/corner*



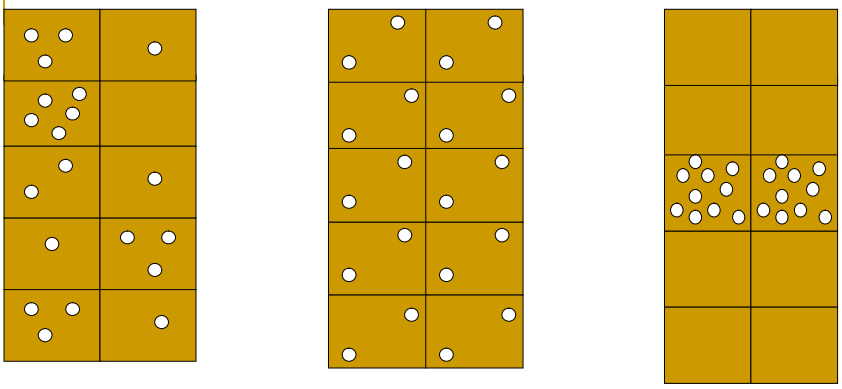
Why Processes differ from Random

Processes differ from random in two fundamental ways

- Variation in the receptiveness of the study area to receive a point
  - Diseases cluster because people cluster (e.g. cancer)
  - Cancer cases cluster ‘cos chemical plants cluster
  - *First order effect*
- Interdependence of the points themselves
  - Diseases cluster ‘cos people catch them from others who have the disease (colds)
  - *Second order effects*

In practice, it is very difficult to disentangle these two effects merely by the analysis of spatial data

What do we mean by spatially random?



RANDOM

UNIFORM/ DISPERSED

CLUSTERED

- Types of Distributions
  - **Random:** any point is equally likely to occur at any location, and the position of any point is not affected by the position of any other point.
  - **Uniform:** every point is as far from all of its neighbors as possible: “unlikely to be close”
  - **Clustered:** many points are concentrated close together, and there are large areas that contain very few, if any, points: “unlikely to be distant”

Centrographic Statistics

- Basic descriptors for spatial point distributions

  Measures of Centrality **Measures of Dispersion**

  - Mean Center -- Standard Distance
  - Centroid -- Standard Deviational Ellipse
  - Weighted mean center
  - Center of Minimum Distance
- Two dimensional (spatial) equivalents of standard descriptive statistics for a single-variable distribution
- May be applied to polygons by first obtaining the centroid of each polygon
- Best used in a **comparative context** to compare one distribution (say in 1990, or for males) with another (say in 2000, or for females)

Mean Center

- Simply the mean of the X and the Y coordinates for a set of points
- Also called center of gravity or centroid
- Sum of differences between the mean X and all other X is zero (same for Y)
- Minimizes sum of squared distances between itself and all points

$$\bar{X} = \sum_{i=1}^{N} \frac{X_i}{N}$$

$$\bar{Y} = \sum_{i=1}^{N} \frac{Y_i}{N}$$

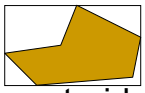
$$\min \sum d_{iC}^{2}$$

Distant points have large effect.

Provides a single point summary measure for the location of distribution.

Centroid

- The equivalent for polygons of the mean center for a point distribution
- The center of gravity or balancing point of a polygon
- if polygon is composed of straight line segments between nodes, centroid again given "average X, average Y" of nodes
- Calculation sometimes approximated as center of bounding box
  - Not good

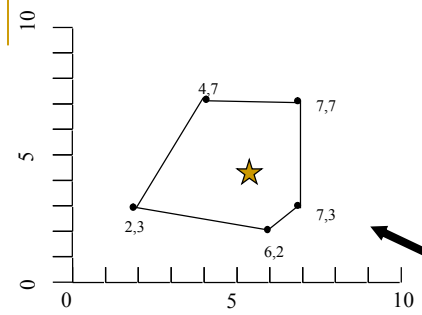


- By calculating the centroids for a set of polygons can apply Centrographic Statistics to polygons

Weighted Mean Center

- Produced by weighting each X and Y coordinate by another variable (W*i*)
- Centroids derived from polygons can be weighted by any characteristic of the polygon

$$\overline{X} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad \overline{Y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

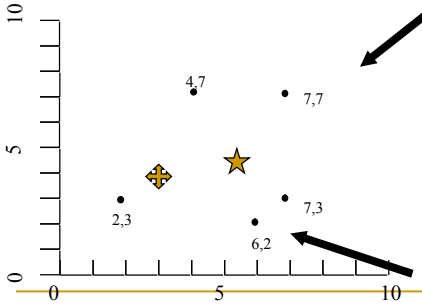


Calculating the centroid of a polygon or the mean center of a set of points.

1	2	3
2	4	7
3	7	7
4	7	3
5	6	2
sum	26	22
Centroid/MC	**5.2**	**4.4**

(same example data as for area of polygon)

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$



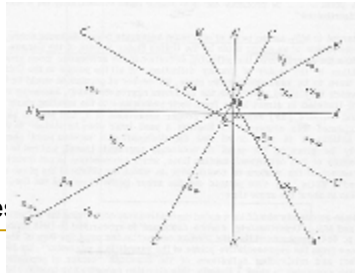
Calculating the weighted mean center. Note how it is pulled toward the high weight point.

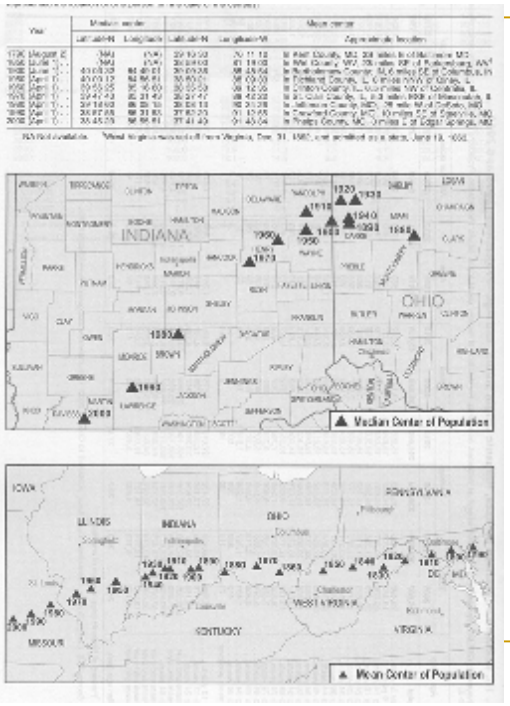
i	X	Y	weight	wX	wY
1	2	3	3,000	6,000	9,000
2	4	7	500	2,000	3,500
3	7	7	400	2,800	2,800
4	7	3	100	700	300
5	6	2	300	1,800	600
sum	26	22	4,300	13,300	16,200
w MC				**3.09**	**3.77**

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum w_i}, \bar{Y} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum w_i}$$

Center of Minimum Distance or *Median Center*

- Also called *point of minimum aggregate travel*
- That point (MD) which minimizes sum of distances between itself and all other points ($_i$) $\min \sum d_{iMD}$
- No direct solution. Can only be derived by approximation
- Not a determinate solution. Multiple points may meet this criteria—see next bullet.
- Same as Median center:
  - Intersection of two orthogonal lines (at right angles to each other), such that each line has half of the points to its left and half to its right
  - Because the orientation of the axis for these lines is arbitrary, multiple points may meet this criteria. Source: Neft, 1966





Median and Mean Centers for US Population

Median Center:
Intersection of a north/south and an east/west line drawn so half of population lives above and half below the e/w line, and half lives to the left and half to the right of the n/s line

Mean Center:
Balancing point of a weightless map, if equal weights placed on it at the residence of every person on census day.

Standard Distance Deviation

- Represents the standard deviation of the distance of each point from the mean center
- Is the two dimensional equivalent of standard deviation for a single variable
- Given by:

Formulae for standard deviation of single variable

$$\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{N}}$$

$$\sqrt{\frac{\sum_{i=1}^{n}(X_i - X_c)^2 + \sum_{i=1}^{n}(Y_i - Y_c)^2}{N}}$$

Or, with weights

$$\sqrt{\frac{\sum_{i=1}^{n} w_i(X_i - X_c)^2 + \sum_{i=1}^{n} w_i(Y_i - Y_c)^2}{\sum_{i=1}^{n} w_i}}$$

which by Pythagoras reduces to:

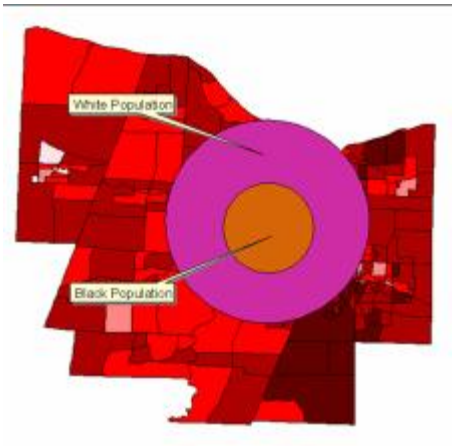
$$\sqrt{\frac{\sum_{i=1}^{n} d_{iC}{}^2}{N}}$$

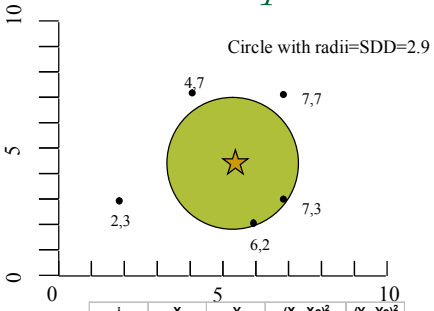
---essentially the average distance of points from the center

Provides a single unit measure of the spread or dispersion of a distribution.

We can also calculate a weighted standard distance analogous to the weighted mean center.

Standard Distance Deviation *Example*





i	X	Y	$(X - Xc)^2$	$(Y - Yc)^2$
1	2	3	10.2	2.0
2	4	7	1.4	6.8
3	7	7	3.2	6.8
4	7	3	3.2	2.0
5	6	2	0.6	5.8
sum	26	22	18.8	23.2
Centroid	5.2	4.4		
			sum of sums	42
			divide N	8.4
			sq rt	2.90

$$sdd = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_c)^2 + \sum_{i=1}^{n}(Y_i - Y_c)^2}{N}}$$

Standard Deviational Ellipse: *concept*

- Standard distance deviation is a good single measure of the dispersion of the incidents around the mean center, but it does not capture any *directional bias*
  - doesn't capture the shape of the distribution.
- The *standard deviation ellipse* gives dispersion in two dimensions
- Defined by 3 parameters
  - Angle of rotation
  - Dispersion along major axis
  - Dispersion along minor axis

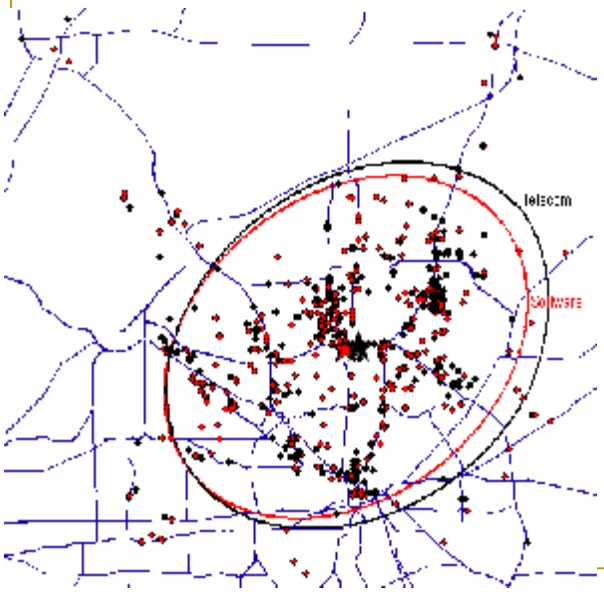
The major axis defines the direction of maximum spread of the distribution

The minor axis is perpendicular to it and defines the minimum spread

Standard Deviational Ellipse: *calculation*

- Formulae for calculation may be found in references cited at end. For example
  - Lee and Wong pp. 203-208
- Basic concept is to:
  - Calculate the mean center
  - Transfer the coordinates of points to the MC
  - Find the axis going through maximum dispersion (thus derive angle of rotation: formula 5.13)
  - Calculate standard deviation of the points along this axis (thus derive the length (radii) of major axis: formula 5.14)
  - Calculate standard deviation of points along the axis perpendicular to major axis (thus derive the length (radii) of minor axis: formula 5.15)

Mean Center & Standard Deviational Ellipse:



example :

There appears to be no major difference between the location of the software and the telecommunications industry in North Texas.

Lecture 2: Point Pattern

Point Pattern Analysis

Analysis of spatial properties of the entire body of points rather than the derivation of single summary measures

Two primary approaches:

- ***Point Density*** approach using **Quadrat Analysis** based on observing the frequency distribution or density of points within a set of grid squares.
  - Variance/mean ratio approach
  - Frequency distribution comparison approach
- ***Point interaction*** approach using **Nearest Neighbor Analysis** based on distances of points one from another

Although the above would suggest that the first approach examines *first order* effects and the second approach examines *second order* effects, in practice the two cannot be separated.

Exhaustive census
--used for secondary (e.g census) data

Random sampling
--useful in field work

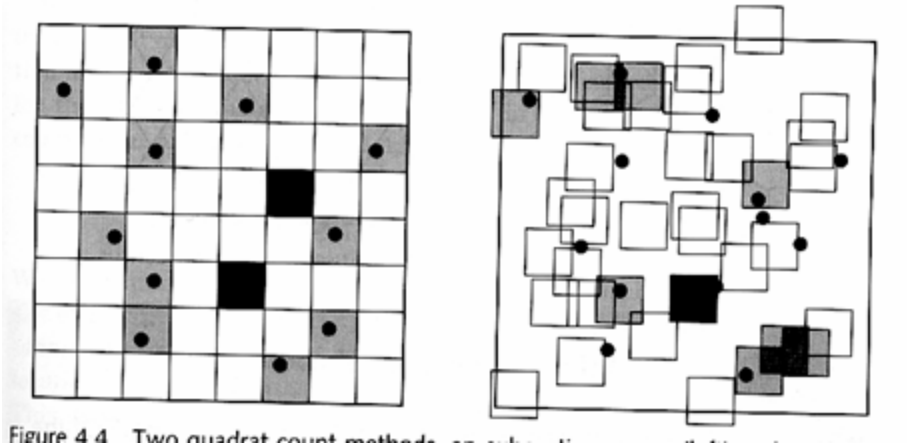
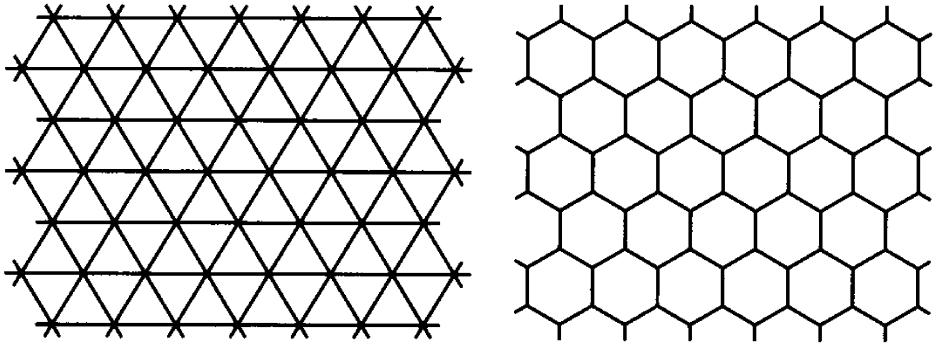


Figure 4.4 Two quadrat count methods

Multiple ways to create quadrats
--and results can differ accordingly!



Frequency counts by Quadrat would be:

	Census Q = 64		Sampling Q = 38	
Number of points in Quadrat	Count	Proportion	Count	Proportion
0	51	0.797	29	0.763
1	11	0.172	8	0.211
2	2	0.031	1	0.026
3	0	0.000	0	0.000
Q = # of quadarts				
P = # of points =		15		

Quadrats don't have to be square
--and their size has a big influence

Quadrat Analysis: *Frequency Distribution Comparison*

- We can compare observed frequencies in the quadrats (Q= number of quadrats) with expected frequencies that would be generated by
  - a random process (modeled by the Poisson frequency distribution)
  - a clustered process (e.g. one cell with P points, Q-1 cells with 0 points)
  - a uniform process (e.g. each cell has P/Q points)
- The standard *Kolmogorov-Smirnov test* for comparing two frequency distributions can then be applied – see next slide
- See Lee and Wong pp. 226-229 for example and further discussion.

$$\sqrt{\frac{2 * A}{P}}$$

Where:
A = area of region
P = # of points

Kolmogorov-Smirnov (K-S) Test

- The test statistic "D" is simply given by:

$$D = \max [\text{ Cum Obser. Freq} - \text{Cum Expect. Freq}]$$

The largest difference (irrespective of sign) between observed cumulative frequency and expected cumulative frequency

- The critical value at the 5% level is given by:

$$D_{(\text{at } 5\%)} = \sqrt{\frac{1.36}{Q}}$$ where Q is the number of quadrats

- Expected frequencies for a random spatial distribution are derived from the Poisson frequency distribution and can be calculated with:

$p(0) = e^{-\lambda} = 1 / (2.71828^{P/Q})$ and $p(x) = p(x - 1) * \lambda / x$

Where x = number of points in a quadrat and p(x) = the probability of x points

P = total number of points Q = number of quadrats

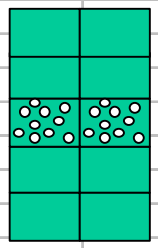
λ = P/Q (the average number of points per quadrat)

See next slide for worked example for cluster case

Calculation of Poisson Frequencies for Kolmogorov-Smirnov test

CLUSTERED pattern as used in lecture

A	B	C	D	E	F	G	H	
		=ColA * Cc	=Col B / q			!Col E - Col G		
Number of Points in quadrat	Observed Quadrat Count	Total Point	Observed Probability	Cumulative Observed Probability	Poisson Probability	Cumulative Poisson Probability	Absolute Difference	
0	8	0	0.8000	0.8000	0.1353	0.1353	0.6647	
1	0	0	0.0000	0.8000	0.2707	0.4060	0.3940	
2	0	0	0.0000	0.8000	0.2707	0.6767	0.1233	
3	0	0	0.0000	0.8000	0.1804	0.8571	0.0571	
4	0	0	0.0000	0.8000	0.0902	0.9473	0.1473	
5	0	0	0.0000	0.8000	0.0361	0.9834	0.1834	
6	0	0	0.0000	0.8000	0.0120	0.9955	0.1955	
7	0	0	0.0000	0.8000	0.0034	0.9989	0.1989	
8	0	0	0.0000	0.8000	0.0009	0.9998	0.1998	
9	0	0	0.0000	0.8000	0.0002	1.0000	0.2000	
10	2	20	0.2000	1.0000	0.0000	1.0000	0.0000	



The Kolmogorov-Smirnov D test statistic is the largest Absolute Difference

	= largest value in Column h						0.6647	
Critical Value at 5% for one sample given by:				1.36/sqrt(Q)			0.4301	Significant
Critical Value at 5% for two sample given by:				1.36*sqrt((Q1+Q2)/Q1*Q2))				
number of quadrats			Q	10	(sum of column B)			
number of points			P	20	(sum of Col C)			
number of points in a quadrat			x					
poisson probability	p(x) =		p(x-1)*(P/Q)/x	(Col E, Row 11 onwards)				
if x=0 then p(x) = p(0)=2.71828^P/Q				(Col E, Row 10)				
Euler's constant		2.7183						

Row 10

The spreadsheet spatstat.xls contains worked examples for the Uniform/ Clustered/ Random data previously used, as well as for Lee and Wong's data

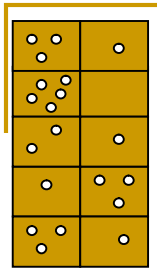
Quadrat Analysis: *Variance/Mean Ratio (VMR)*

- Apply uniform or random grid over area (A) with width of square given by:

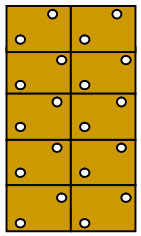
$$\sqrt{\frac{2 * A}{P}}$$

Where:
A = area of region
P = # of points

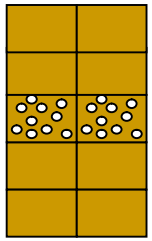
- Treat each cell as an observation and count the number of points within it, to create the variable X
- Calculate variance and mean of X, and create the variance to mean ratio: *variance / mean*
- For an uniform distribution, the variance is zero.
  - Therefore, we expect a variance-mean ratio close to 0
- For a random distribution, the variance and mean are the same.
  - Therefore, we expect a variance-mean ratio around 1
- For a clustered distribution, the variance is relatively large
  - Therefore, we expect a variance-mean ratio above 1



RANDOM



UNIFORM/ DISPERSED



CLUSTERED

3	1
5	0
2	1
1	3
3	1

Quadrat #	x Number of Points Per Quadrat	x^2
1	3	9
2	1	1
3	5	25
4	0	0
5	2	4
6	1	1
7	1	1
8	3	9
9	3	9
10	1	1
	20	60

Variance	2.222
Mean	2.000
Var/Mean	1.111

random

2	2
2	2
2	2
2	2
2	2

Quadrat #	x Number of Points Per Quadrat	x^2
1	2	4
2	2	4
3	2	4
4	2	4
5	2	4
6	2	4
7	2	4
8	2	4
9	2	4
10	2	4
	20	40

Variance	0.000
Mean	2.000
Var/Mean	0.000

uniform

0	0
0	0
10	10
0	0
0	0

Quadrat #	x Number of Points Per Quadrat	x^2
1	0	0
2	0	0
3	0	0
4	0	0
5	10	100
6	10	100
7	0	0
8	0	0
9	0	0
10	0	0
	20	200

Variance	17.778
Mean	2.000
Var/Mean	8.889

Clustered

Formulae for variance

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{N-1} \quad = \frac{\sum_{i=1}^{n} X_i^2 - [(\sum X)^2/N]}{N-1}$$

Note:

N = number of Quadrats = 10

Ratio = Variance/mean

Significance Test for VMR

- A significance test can be conducted based upon the chi-square frequency
- The test statistic is given by: (sum of squared differences)/Mean

$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\overline{X}} \quad = \frac{\sum_{i=1}^{n} X_i^2 - [(\sum X)^2/N]}{\overline{X}}$$

- The test will ascertain if a pattern is significantly more clustered than would be expected by chance (but does not test for a uniformity)
- The values of the test statistics in our cases would be:

$$\underline{\text{random}} \qquad \frac{60-(20^2)/10}{2} = 10$$

$$\underline{\text{uniform}} \qquad \frac{40-(20^2)/10}{2} = 0$$

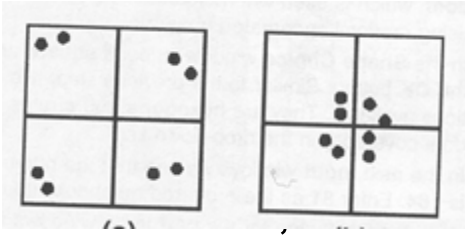
$$\underline{\text{clustered}} \qquad \frac{200-(20^2)/10}{2} = 80$$

- For degrees of freedom: N - 1 = 10 - 1 = 9, the value of chi-square at the 1% level is 21.666.
- Thus, there is only a 1% chance of obtaining a value of 21.666 or greater if the points had been allocated randomly. Since our test statistic for the clustered pattern is 80, we conclude that there is (considerably) less than a 1% chance that the clustered pattern could have resulted from a random process
- T-test in p.p.234

Weakness of Quadrat Analysis

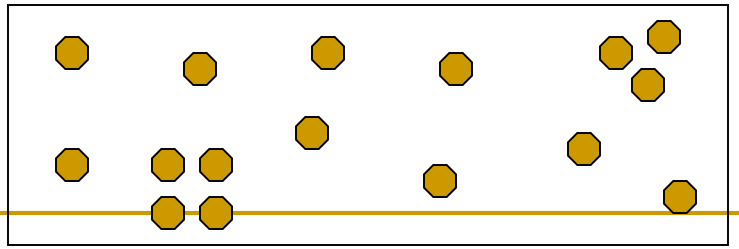
- Results may depend on quadrat size and orientation (Modifiable areal unit problem)
  - test different sizes (or orientations) to determine the effects of each test on the results
- Is a measure of dispersion, and not really pattern, because it is based primarily on the density of points, and not their arrangement in relation to one another

For example, quadrat analysis cannot distinguish between these two, obviously different, patterns



- Results in a single measure for the entire variations within the region are not recognized (could have clustering locally in some areas, but not overall)

For example, overall pattern here is dispersed, but there are some local clusters



Nearest-Neighbor Index (NNI)

- uses distances between points as its basis.
- Compares the mean of the distance observed between each point and its nearest neighbor with the expected mean distance that would occur if the distribution were random:

NNI=Observed Aver. Dist / Expected Aver. Dist

For random pattern, NNI = 1

For clustered pattern, NNI = 0

For dispersed pattern, NNI = 2.149

- We can calculate a **Z** statistic to test if observed pattern is significantly different from random:
- $Z = \frac{\text{Av. Dist Obs - Av. Dist. Exp.}}{\text{Standard Error}}$

if Z is below –1.96 or above +1.96, we are 95% confident that the distribution is not randomly distributed. (If the observed pattern was random, there are less than 5 chances in 100 we would have observed a z value this large.)

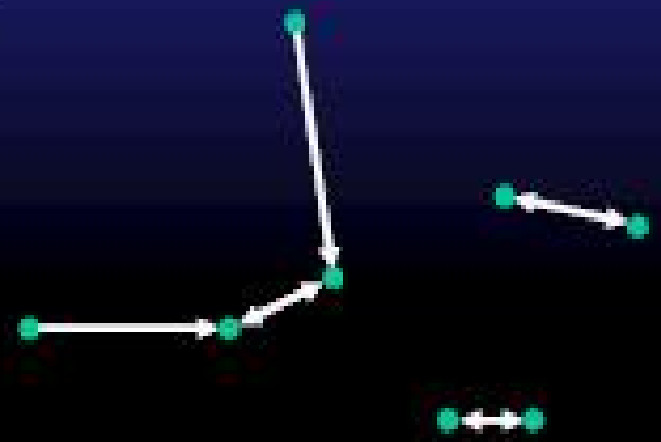
(in the example that follows, the fact that the NNI for uniform is 1.96 is coincidence!)

d_i: Distance to the nearest point from point i

n: Number of points

W: (Average) nearest neighbor distance

$$W = \frac{1}{n}\sum_{i=1}^{n} d_i$$



Nearest Neighbor Formulae

Index

NN Index: The Nearest Neighbor Index (Uncorrected)

$$NNI = \frac{\bar{d}}{E(\bar{d})}$$

Where:

Avg. Dist.: Average Nearest Neighbor Distance

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

Exp. Avg.: Expected Average Nearest Neighbor Distance (Uncorrected)

$$E(\bar{d}) = 0.5\sqrt{\frac{A}{n}}$$

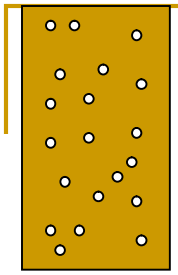
Significance test

SD: Standard deviation (Standard error)

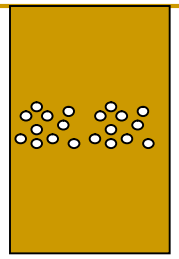
$$SD = \sqrt{\left(\frac{1}{4\tan^{-1}1} - \frac{1}{4}\right)\frac{A}{n^2}} = \frac{0.26136}{\sqrt{n^2 / A}}$$

Z: Standard z-value (Uncorrected)

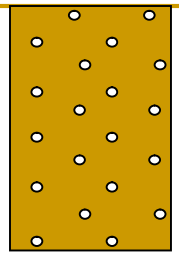
$$Z = \frac{\bar{d} - E(\bar{d})}{SD}$$



RANDOM



CLUSTERED



UNIFORM

Point	Nearest Neighbor	Distance
1	2	1
2	3	0.1
3	2	0.1
4	5	1
5	4	1
6	5	2
7	6	2.7
8	10	1
9	10	1
10	9	1
		10.9

Mean distance	1.09
Area of Region	50
Density	0.2
Expected Mean	1.118034
NNI	0.974926

$Z = -0.1515$

Point	Nearest Neighbor	Distance
1	2	0.1
2	3	0.1
3	2	0.1
4	5	0.1
5	4	0.1
6	5	0.1
7	6	0.1
8	9	0.1
9	10	0.1
10	9	0.1
		1

Mean distance	0.1
Area of Region	50
Density	0.2
Expected Mean	1.118034
NNI	0.089443

$Z = 5.508$

Point	Nearest Neighbor	Distance
1	3	2.2
2	4	2.2
3	4	2.2
4	5	2.2
5	7	2.2
6	7	2.2
7	8	2.2
8	9	2.2
9	10	2.2
10	9	2.2
		22

Mean distance	2.2
Area of Region	50
Density	0.2
Expected Mean	1.118034
NNI	1.96774

$Z = 5.855$

Source: Lembro

Higher-ordered NNI

$$r_{\exp} = 0.75\sqrt{A/n}$$

$$SE_r = 0.272\sqrt{A/n^2}$$

2nd ordered

$$r_{\exp}(k) = \gamma_1(k)\sqrt{A/n} = \frac{k(2k)!}{(2^k k!)^2}\sqrt{A/n}$$

$$SE_r(k) = \gamma_2(k)\sqrt{A/n^2}$$

K-ordered

$$r_{\exp} = 0.5\sqrt{A/n} + (0.0514 + \frac{0.041}{\sqrt{n}})\frac{B}{n}$$

$$SE_r^2 = 0.0683\frac{A}{n^2} + 0.037B\sqrt{\frac{A}{n^5}}$$

Edge adjustment

Evaluating the Nearest Neighbor Index

- Advantages
  - NNI takes into account distance
  - No quadrat size problem to be concerned with
- However, NNI not as good as might appear
  - Index highly dependent on the boundary for the area
    - its size and its shape (perimeter)
  - Fundamentally based on only the mean distance
  - Doesn't incorporate local variations (could have clustering locally in some areas, but not overall)
  - Based on point location only and doesn't incorporate magnitude of phenomena at that point
- An "adjustment for edge effects" available but does not solve all the problems
- Some alternatives to the NNI are the **G** and **F** functions, based on the entire frequency distribution of nearest neighbor distances, and the **K** function based on all interpoint distances.
  - See O and U pp. 89-95 for more detail.
  - Note: the G Function and the General/Local G statistic (to be discussed later) are related but not identical to each other

Lecture 3: Point Pattern cont.

K-function

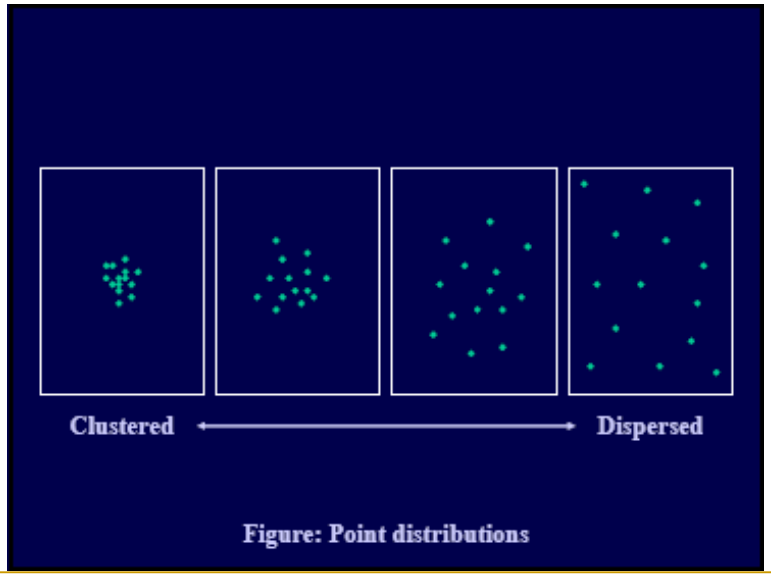


Figure: Point distributions

Definition of *K*-function

K-function indicates the average number of points within a certain distance *h* from points. It is, therefore, a function of the distance *h*.

K-function counts the number of points located in the circular region of radius *h* from every point in *S*, and divides it by the number and density of points for standardization.

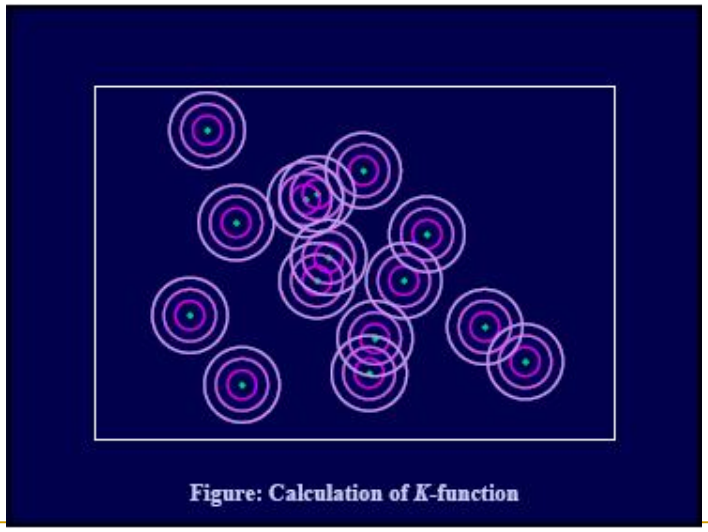


Figure: Calculation of *K*-function

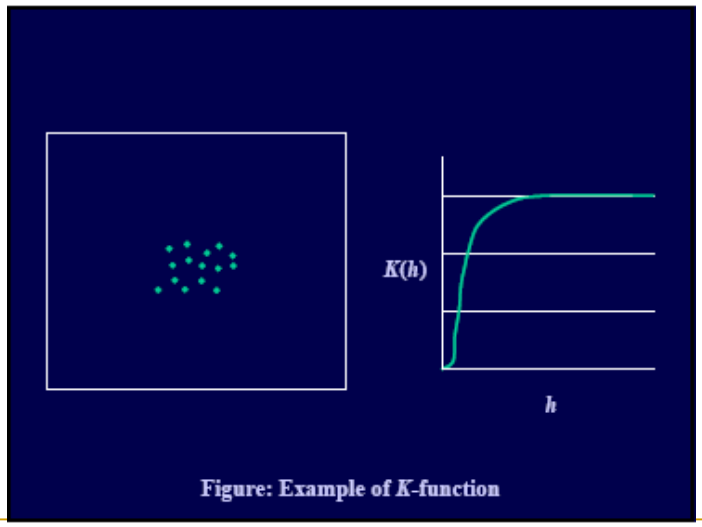


Figure: Example of *K*-function

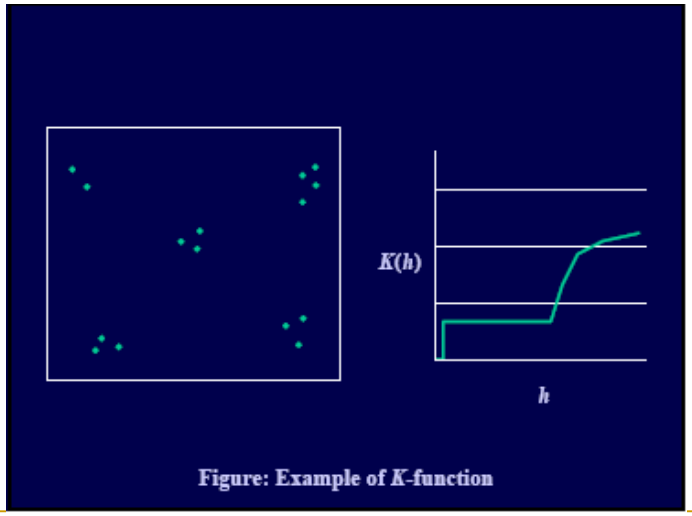


Figure: Example of *K*-function

Formal definition of *K*-function

n: Number of points
$\mathbf{x}_i$: Locational vector of point i
λ: Point density ($=n/A$, A: the area of the region)
h: Distance parameter
σ_{ij}: Binary function defined by

$$\sigma_{ij}(h) = \begin{cases} 1 & \text{if } |\mathbf{x}_i - \mathbf{x}_j| \le h \\ 0 & \text{otherwise} \end{cases}$$

K-function is mathematically defined by

$$K(h) = \frac{\sum_i \sum_{j \neq i} \sigma_{ij}(h)}{n\lambda}$$

The numerator is the total number of points within a certain distance h from points.

K-function of points under CSR

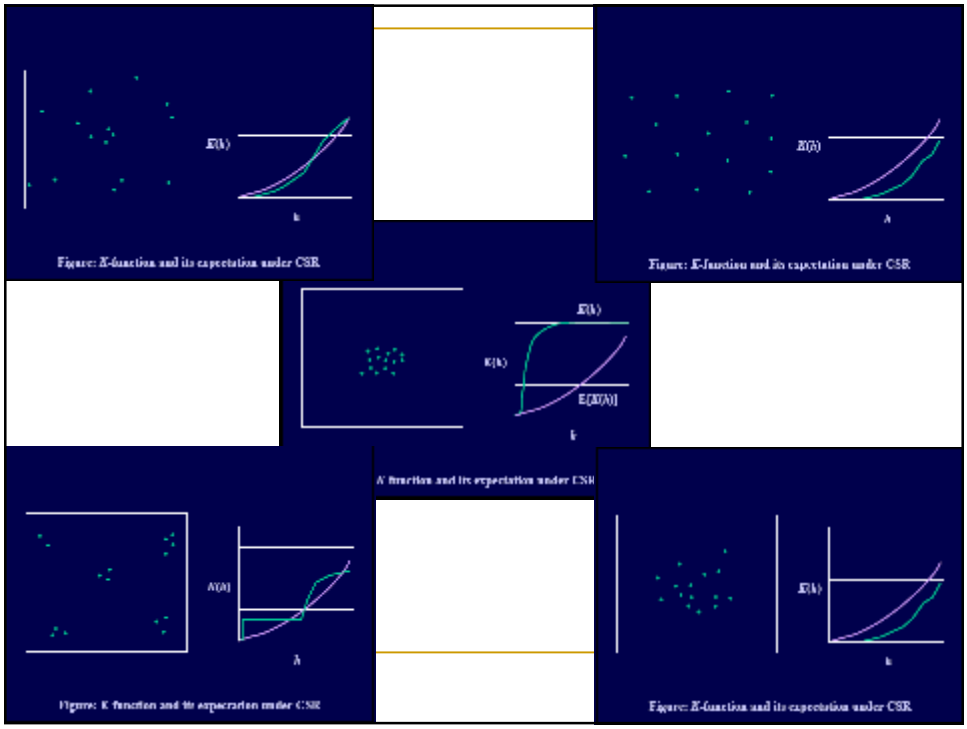
To evaluate the degree of spatial clustering of points, we again consider the distribution of points under CSR.

The expectation of *K*-function of points under CSR is

$$\mathrm{E}\left[K(h)\right] = \pi h^2$$

Comparing $K(h)$ and its expectation, we can classify point distributions into one of three categories:

$K(h) > \pi h^2$: Points are clustered
$K(h) = \pi h^2$: Points are randomly distributed
$K(h) < \pi h^2$: Points are dispersed



Standardization of *K*-function

To compare *K*-function values among different *h* values, we standardize the *K*-function

$$L(h) = \sqrt{\frac{K(h)}{\pi}} - h$$

The standardized *K*-function is called *L*-function.

Point distributions are then classified as below.

$L(h) > 0$:	Points are clustered
$L(h) \approx 0$:	Points are randomly distributed
$L(h) < 0$:	Points are dispersed

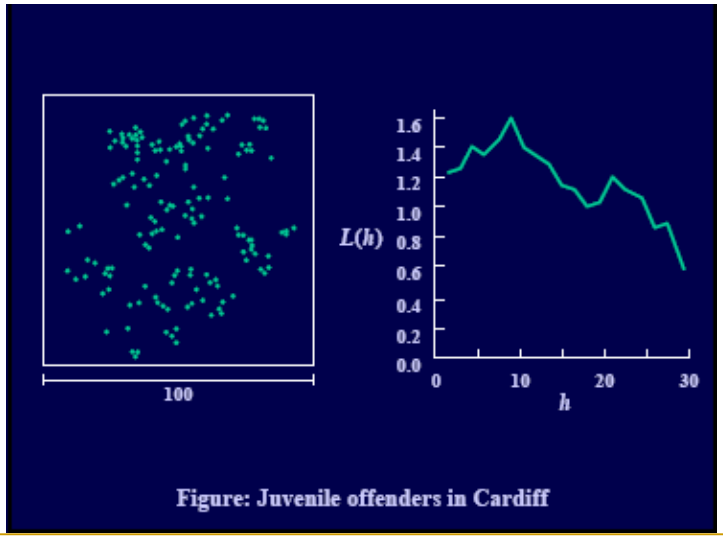


Figure: Juvenile offenders in Cardiff

Lecture 4: Spatial Autocorrelation (point)

Moran *I*

Geary *C*

Getis and Ord's *G*

Spatial Autocorrelation

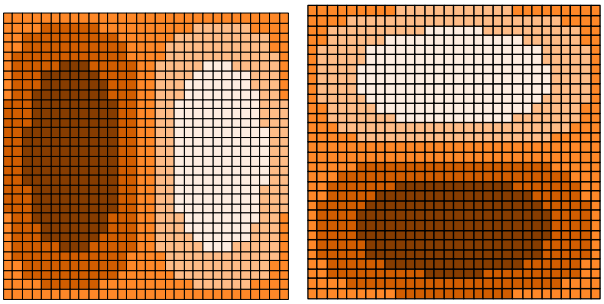
- The instantiation of Tobler's first law of geography
  - Everything is related to everything else, but near things are more related than distant things.
- Correlation of a variable with itself through space.
- The correlation between an observation's value on a variable and the value of close-by observations on the same variable
- The degree to which characteristics at one location are similar (or dissimilar) to those nearby.
- Measure of the extent to which the occurrence of an event in an areal unit constrains, or makes more probable, the occurrence of a similar event in a neighboring areal unit.
- Several measures available:
  - Join Count Statistic (polygon)
  - Moran's I (point and polygon)
  - Geary's C ratio (point and polygon)
  - General (Getis-Ord) G (point and polygon)
  - Anselin's Local Index of Spatial Autocorrelation (LISA) (point and polygon)

These measures may be "global" or "local"

Representation of Spatial Effects: WY, Autocorrelation

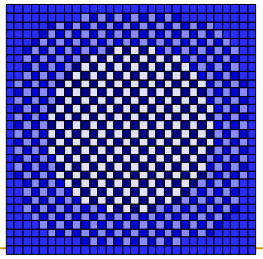
- **W**
  - The Spatial Weights Matrix
  - The Spatial Association of All Sites to All Other Sites
  - d, d^2, 1/0, 1/d
- **Y**
  - The Attribute Association Matrix
  - The Association of the Attributes at Each Site to the Attributes at All Other Sites
  - +,-,/,x

Spatial Autocorrelation



Positive: similar values cluster together on a map

Auto: self
Correlation: degree of relative correspondence



Source: Dr Dan Griffith, with modification

Negative: dissimilar values cluster together on a map

Why Spatial Autocorrelation Matters

- Spatial autocorrelation is of interest in its own right because it suggests the operation of a spatial process
- Additionally, most statistical analyses are based on the assumption that the values of observations in each sample are independent of one another
  - Positive spatial autocorrelation violates this, because samples taken from nearby areas are related to each other and are not independent

Why Spatial Autocorrelation Matters

- In ordinary least squares regression (OLS), for example, the correlation coefficients will be biased and their precision exaggerated
  - Bias implies correlation coefficients may be higher than they really are
    - They are biased because the areas with higher concentrations of events will have a greater impact on the model estimate
  - Exaggerated precision (lower standard error) implies they are more likely to be found "statistically significant"
    - they will overestimate precision because, since events tend to be concentrated, there are actually a fewer number of independent observations than is being assumed.

The Spatial Weights Matrix

W is the formal expression of the spatial association between objects

(it is the pair-wise geometry of objects being studied).

Measuring Relative Spatial Location

- How do we measure the relative location or distance apart of the points or polygons? Seems obvious but its not!
- Calculation of *Wij, the spatial weights matrix,* indexing the relative location of all points *i* and *j,* is the big issue for all spatial autocorrelation measures
- Different methods of calculation potentially result in different values for the measures of autocorrelation and different conclusions from statistical significance tests on these measures

- Weights based on Contiguity
  - If zone *j* is adjacent to zone *i*, the interaction receives a weight of 1, otherwise it receives a weight of 0 and is essentially excluded
  - But what constitutes contiguity? Not as easy as it seems!
- Weights based on Distance
  - Uses a measure of the actual distance between points or between polygon centroids
  - But what measure, and distance to what points -- All? Some?
- Often, GIS is used to calculate the spatial weights matrix, which is then inserted into other software for the statistical calculations

Weights Based on Contiguity

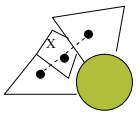
For Regular Polygons

rook case  or queen case 

For Irregular polygons

- All polygons that share a common border
- All polygons that share a common border or have a centroid within the circle defined by the average distance to (or the "convex hull" for) centroids of polygons that share a common border

For points



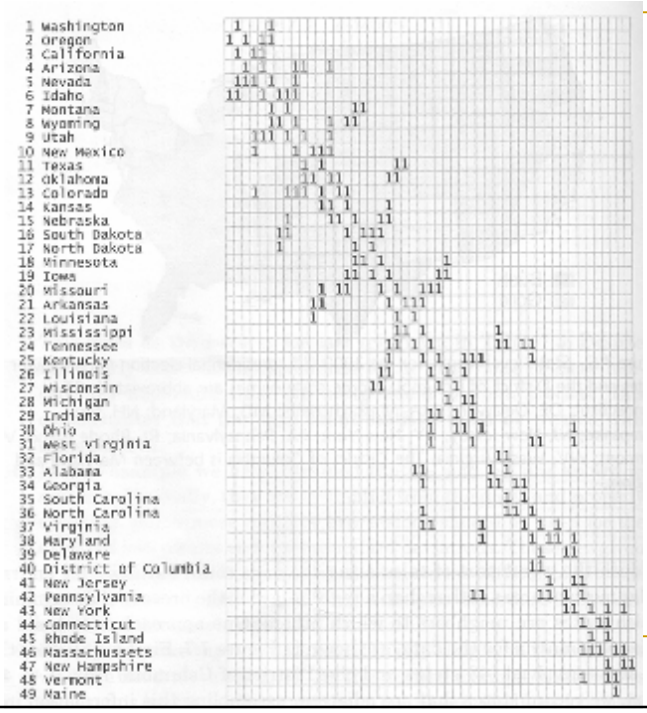
- The closest point (nearest neighbor)

--select the contiguity criteria

--construct n x n weights matrix with 1 if contiguous, 0 otherwise

Weights based on Lagged Contiguity

- We can also use adjacency matrices which are based on lagged adjacency
  - Base contiguity measures on "next nearest" neighbor, not on immediate neighbor
- In fact, can define a range of contiguity matrices:
  - 1st nearest, 2nd nearest, 3rd nearest, etc.



Queens Case <u>Full</u> Contiguity Matrix for US States

- 0s omitted for clarity
- Column headings (same as rows) omitted for clarity
- Principal diagonal has 0s (blanks)
- Can be very large, thus ~~inefficient to use~~.

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		

Queens Case Sparse Contiguity Matrix for US States

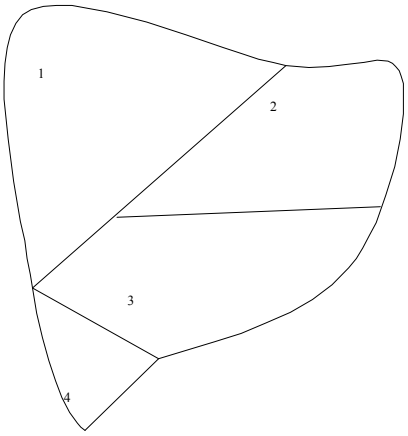
- *Ncount* is the number of neighbors for each state
- Max is 8 (Missouri and Tennessee)
- Sum of Ncount is 218
- Number of common borders (joins)

$\sum$ ncount / 2 = 109

- *N1, N2…* FIPS codes for neighbors

Weights Based on Distance

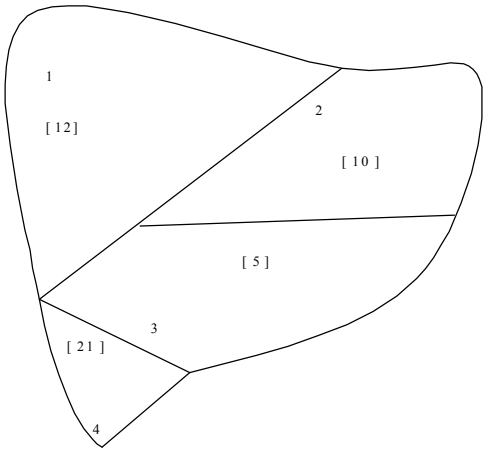
- Most common choice is the inverse (reciprocal) of the distance between locations i and j ($w_{ij} = 1/d_{ij}$)
  - Linear distance?
  - Distance through a network?
- Other functional forms may be equally valid, such as inverse of squared distance ($w_{ij} = 1/d_{ij}^2$), or negative exponential (e^{-d} *or* e^{-d^2})
- Can use length of shared boundary: w_{ij} = *length (ij)/length(i)*
- Inclusion of distance to all points may make it impossible to solve necessary equations, or may not make theoretical sense (effects may only be 'local')
  - Include distance to only the "nth" nearest neighbors
  - Include distances to locations only within a buffer distance
- For polygons, distances usually measured centroid to centroid, but
  - could be measured from perimeter of one to centroid of other
  - For irregular polygons, could be measured between the two closest boundary points (an adjustment is then necessary for contiguous polygons since distance for these would be zero)



0 1 1 0
1 0 1 0
1 1 0 1
0 0 1 0

Typical W

- Spatially contiguous neighbors (rook, queen: one/zero)
- Inverse distances raised to a power: ($1/d$, $1/d^2$, $1/d^5$)
- Geostatistics functions (spherical, gaussian, exponential)
- Lengths of shared borders (perimeters)
- All centroids within distance *d*
- n^{th} nearest neighbor distance
- Links (number of)



0	2	7	- 9
-2	0	5	-11
-7	-5	0	-16
9	11	16	0

A Note on Sampling Assumptions(1/2)

- Another factor which influences results from these tests is the assumption made regarding the type of sampling involved:
  - Free (or normality) sampling assumes that the probability of a polygon having a particular value is not affected by the number or arrangement of the polygons
    - Analogous to sampling with replacement
  - Non-free (or randomization) sampling assumes that the probability of a polygon having a particular value is affected by the number or arrangement of the polygons (or points), usually because there is only a fixed number of polygons (e.g. if n = 20, once I have sampling 19, the 20th is determined)
    - Analogous to sampling without replacement

A Note on Sampling Assumptions(2/2)

- The formulae used to calculate the various statistics (particularly the standard deviation/standard error) differ depending on which assumption is made
  - Generally, the formulae are substantially more complex for randomization sampling—unfortunately, it is also the more common situation!
  - Usually, assuming normality sampling requires knowledge about larger trends from outside the region or access to additional information within the region in order to estimate parameters.

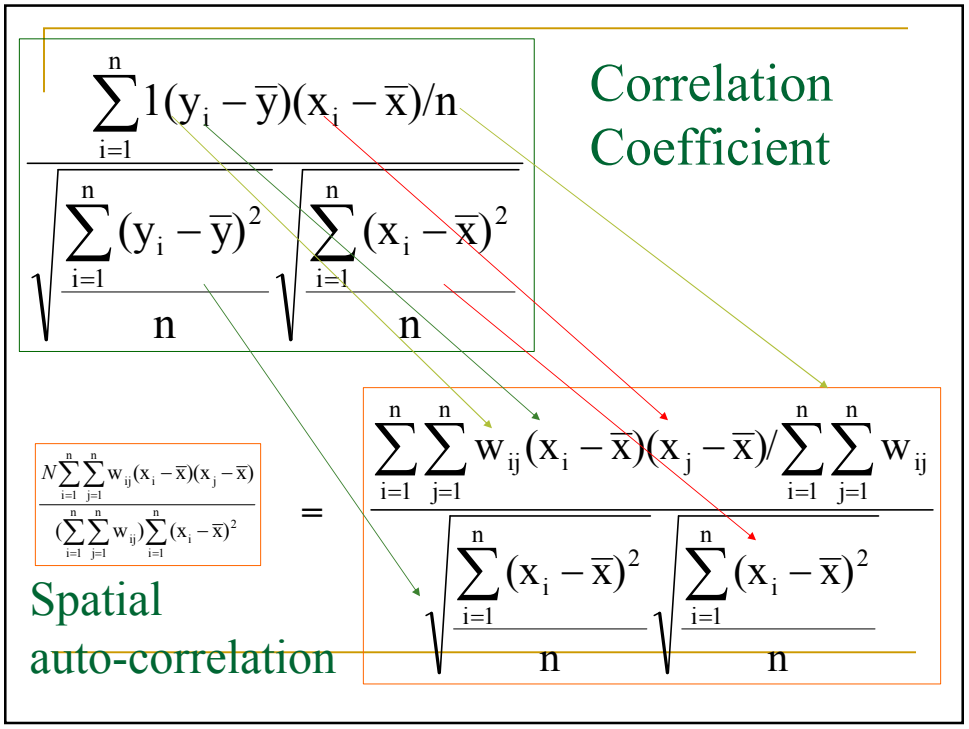
Spatial Autocorrelation Coefficient (SAC)

$$SAC \approx \frac{\sum_i \sum_j s_{ij} w_{ij}}{\sum_i \sum_j w_{ij}}$$

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 cases
  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*
- Applied to a continuous variable for polygons or points
- Similar to correlation coefficient: varies between –1.0 and + 1.0
  - 0 indicates no spatial autocorrelation *[approximate: technically it's –1/(n-1)]*
  - When autocorrelation is high, the I coefficient is close to 1 or -1
  - Negative/positive values indicate negative/positive autocorrelation
- Can also use Moran as index for dispersion/random/cluster patterns
  - Indices close to zero *[technically, close to -1/(n-1)],* indicate random pattern
  - Indices above -1/(n-1) (toward +1) indicate a tendency toward clustering
  - Indices below -1/(n-1) (toward -1) indicate a tendency toward dispersion/uniform
- Differences from correlation coefficient are:
  - Involves one variable only, not two variables
  - Incorporates weights (*wij*) which index relative location
  - Think of it as "the correlation between neighboring values on a variable"
  - More precisely, the correlation between variable, X, and the "spatial lag" of X formed by averaging all the values of X for the neighboring polygons



Adjustment for Short or Zero Distances

- If an inverse distance measure is used, and distances are very short, then w_{ij} becomes very large and distorts I.
- An adjustment for short distances can be used, usually scaling the distance to one mile.
- The units in the adjustment formula are the number of data measurement units in a mile
- In the example, the data is assumed to be in feet.
- With this adjustment, the weights will never exceed 1
- If a contiguity matrix is used (1or 0 only), this adjustment is unnecessary

$$W_{i,j} = \frac{unit}{unit + d_{ij}}$$

$$W_{i,j} = \frac{mile}{mile + d_{ij}}$$

$$W_{i,j} = \frac{5280}{5280 + d_{ij}}$$

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 (mean)
S is the standard error

E(I) = -1/(n-1)

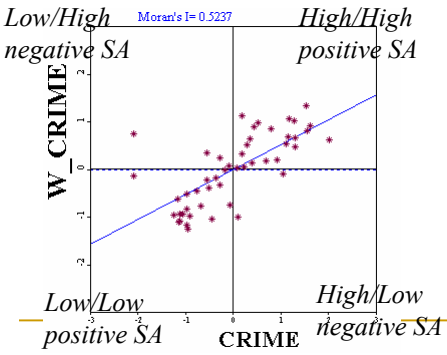
- However, there are two different formulations for the standard error calculation
  - The randomization or nonfree sampling method
  - The normality or free sampling method

  The actual formulae for calculation are in Lee and Wong p. 266
- Consequently, two slightly different values for Z are obtained. In either case, based on the normal frequency distribution, a value 'beyond' +/- 1.96 indicates a statistically significant result at the 95% confidence level ($p <= 0.05$)

Moran Scatter Plots

Moran's I can be interpreted as the correlation between variable, X, and the "spatial lag" of X formed by averaging all the values of X for the neighboring polygons

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



The slope of the regression line is Moran's I
Each quadrant corresponds to one of the four different types of spatial association (SA)

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 death rates (e.g. SIDS rate: number of sudden infant death syndrome deaths per 1,000 births)
- An adjustment should be made in these cases especially if the denominator in the rate (population or number of births) varies greatly (as it usually does)
- Adjustment is know as the EB adjustment:
  - http://en.wikipedia.org/wiki/Empirical_Bayes_method
- Anselin's *GeoDA* software includes an option for this adjustment both for Moran's I and for LISA

Geary's C (Contiguity) Ratio

- Calculation is similar to Moran's I,
  - For Moran, the cross-product is based on the deviations from the mean for the two location values
  - For Geary, the cross-product uses the actual values themselves at each location

$$C = \frac{[(N-1)[\sum_i \sum_j W_{ij}(X_i - X_j)^2]}{2(\sum_i \sum_j W_{ij}(X_i - \bar{X})^2}$$

- However, interpretation of these values is very different, essentially the opposite!

  Geary's C varies on a scale from 0 to 2
  - C of approximately 1 indicates no autocorrelation/random
  - C of 0 indicates perfect positive autocorrelation/clustered
  - C of 2 indicates perfect negative autocorrelation/dispersed
- Can convert to a -/+1 scale by: calculating C* = 1 - C
- Moran's I is usually preferred over Geary's C

Statistical Significance Tests for Geary's C

- Similar to Moran
- Again, based on the normal frequency distribution with

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

Where: C is the calculated value for Moran's I from the sample
E(C) is the expected value (mean)
S is the standard error

however, E(C) = 1

- Again, there are two different formulations for the standard error calculation
  - The randomization or nonfree sampling method
  - The normality or free sampling method

  The actual formulae for calculation are in Lee and Wong p. 265
- Consequently, two slightly different values for Z are obtained. In either case, based on the normal frequency distribution, a value 'beyond' +/- 1.96 indicates a statistically significant result at the 95% confidence level (p <= 0.05)

General G-Statistic

- Moran's I and Geary's C will indicate clustering or positive spatial autocorrelation if high values (e.g. neighborhoods with high crime rates) cluster together (often called *hot spots*) **and/or** if low values cluster together (*cold spots*) , but they cannot distinguish between these situations
- The General 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* (value for which there is no spatial association)
  - Larger than expected value ➔ potential "hot spot"
  - Smaller than expected value ➔ potential "cold spot"
- A Z *test statistic* is used to test if the difference is sufficient to be statistically significant
- Calculation of G must begin by identifying a *neighborhood distance* within which cluster is expected to occur
- Global indicator or "LISA," statistic.

Calculating General G

- Actual Value for G is given by:

$$G(d) = \frac{\sum\sum w_{ij}(d)x_i x_j}{\sum\sum x_i x_j}$$

Where:
d is neighborhood distance
W_{ij} weights matrix has only 1 or 0
1 if j is within d distance of i
0 if its beyond that distance

- Expected value (if no concentration) for G is given by:

$$E(G) = \frac{W}{n(n-1)} \qquad \text{where} \qquad W = \sum_i \sum_j w_{ij}(d),$$

- For the General G, the terms in the numerator (top) are calculated "within a distance bound (d)," and are then expressed relative to totals for the entire region under study.
  - As with all of these measures, if adjacent **x** terms are both large with the same sign (indicating positive spatial association), the numerator (top) will be large
  - If they are both large with different signs (indicating negative spatial association), the numerator (top) will again be large, but negative

Testing General G

- The test statistic for G is normally distributed and is given by:

$$Z = \frac{G - E(G)}{S_{error(G)}} \qquad \text{with} \qquad E(G) = \frac{W}{n(n-1)}$$

However, the calculation of the standard error is complex. See Lee and Wong pp 381-382 for formulae.

- As an example: Lee and Wong find the following values:

  G(d) = 0.5557 E(G) = .5238.

  Since G(d) is greater than E(G) this indicates potential "hot spots" (clusters of high values)

  However, the test statistic Z= 0.3463

  Since this does not lie "beyond +/-1.96, our standard marker for a 0.05 significance level, we conclude that the difference between G(d) and E(G) could have occurred by chance." There is no compelling evidence for a hot spot.

Attribute Relationships

Y

- **Types of Relationships**
  - Additive association (clustering): $(Y_i + Y_j)$
  - Multiplicative association (product): (Y_iY_j)
  - Covariation (correlation): $(Y_i - Ybar)(Y_j - Ybar)$
  - Differences (homogeneity/heterogeneity): $(Y_i - Y_j)$
  - Inverse (relativity): (Y_i/Y_j)
- **All Relationships Subject to Mathematical Manipulation (power, logs, abs, etc.)**

WY: Covariance

- Set **W** to preferred spatial weights matrix
- (rooks, queens, distance decline, etc.)
- Set **Y** to $(x_i - \mu)\ (x_j - \mu)$
- Set scale to $n/W\ \Sigma(x_i - \mu)^2$
- $I = n\ \ \Sigma\ \Sigma\ W_{ij}\ (x_i - \mu)\ (x_j - \mu) / W\ \Sigma(x_i - \mu)^2$
  where W is sum of all W_{ij} and $i \neq j$

This is Moran's I.

WY: Multiplicativ

- Set **W** to 1/0 spatial weights matrix
- 1 within *d*; 0 outside of *d*
- Set **Y** to $(x_i + x_j)$
- Set scale to $\Sigma\ W_{ij}(d) / \Sigma\ (x_i)$
- $G(d) = \Sigma\ W_{ij}(d)\ (x_i + x_j) / \Sigma\ (x_i)$ *and* $i \neq j$

This is Getis and Ord's G.

WY: Difference

- Set **W** to preferred spatial weights matrix
- Set **Y** to $(x_i - x_j)^2$
- Set scale to $(n-1)/2W \sum (x_i - \mu)^2$
- $c = (n - 1) \sum \sum W_{ij} (x_i - y_{ij})^2 / 2W \sum (x_i - \mu)^2$
  where W is sum of all W_{ij} and $i \neq j$.

This is Geary's c.

WY: Difference

- Set **W** to 1/0 weights matrix; 1 within *ad* and 0 otherwise; *a* is an integer; d is a constant distance
- Set **Y** to $(x_i - x_j)^2$
- Set scale to *1/2*
- $\chi(ad) = 1/2 \sum \sum W_{ij} (x_i - x_{ij})^2$

This is the semivariance

Isotropic Moran's I