Module 4

Spatial Statistics: Spatial Pattern and Spatial Autocorrelation

Spatial Statistics

- Geographers are very interested in studying, understanding, and quantifying the patterns we can see on maps
- **Q:** What kinds of map “patterns” can you think of?
  - There are so many that are possible – what phenomena create patterns that we can view on a map?

Spatial Statistics

- In this module we will deal with two different situations where it is possible to view and test map patterns
  - 1. point patterns
  - 2. area patterns
- We will deal with point patterns first

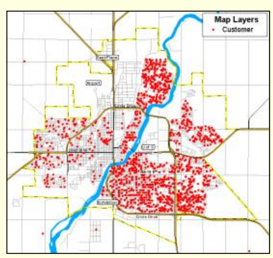
Point Patterns: Nearest Neighbor

- “Nearest neighbor analysis” deals specifically with point patterns
  - Focus of method: determine whether a point pattern is clustered or dispersed
  - Origin of method: plant ecology (spread of plants and seeds)

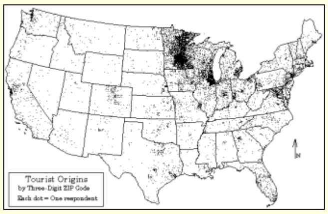
Point Patterns: Nearest Neighbor

- **Why study point patterns?**
  - Trying to understand processes
    - Agglomeration/grouping
    - Diffusion/spreading
    - Competition (between different types of “points”, such as plants, families, or businesses)
  - Looking at pattern change and pattern comparison – differences in patterns between distinct regions and times

Point Example: Customer Map

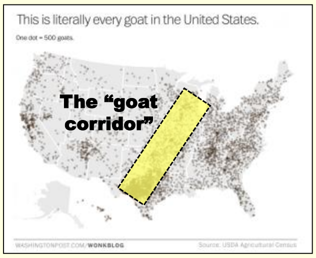


Point Example: Visitor Map



Residential locations for visitors to a Minnesota tourist attraction

Point Example: Visitor Map

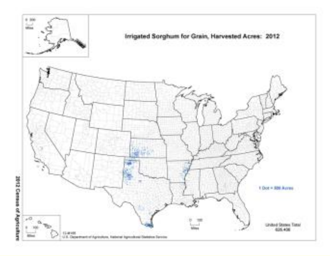


Point Example: School Map

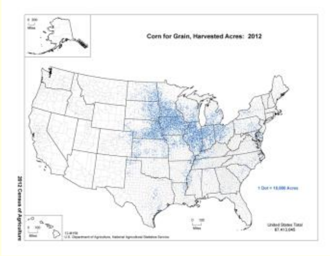
School Closures and Openings in Dallas

Point Example: Crop Map

Point Example: Crop Map

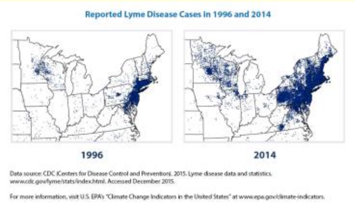


Point Example: Crop Map



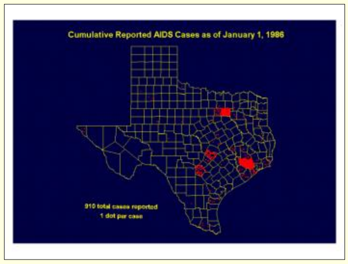
UNT

Point Example: Crop Map



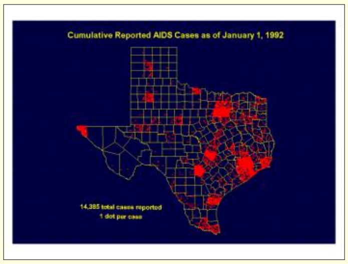
UNT

Point Example: Disease Spread



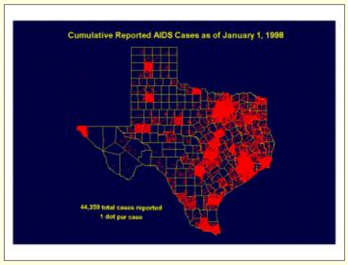
UNT

Point Example: Disease Spread



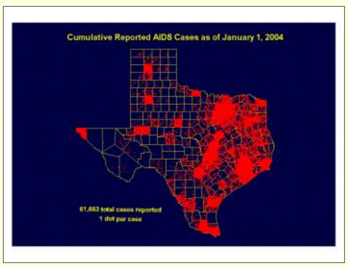
UNT

Point Example: Disease Spread



UNT

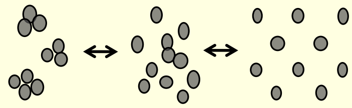
Point Example: Disease Spread



UNT

Nearest Neighbor Analysis

- **Basic Idea**
  - Compare an observed point pattern to a theoretical distribution
  - Think of patterns as ranging along a continuous scale from "clustered" to "uniform"



Clustered **Random** **Uniform**

Nearest Neighbor Analysis

- **Goals of Nearest Neighbor Analysis**
  - 1. Measure pattern: is it nonrandom?
  - 2. Is it significantly nonrandom?
  - 3. What can we say about the ordering process: clustered or dispersed?

Nearest Neighbor Analysis

- **Nearest neighbor index (R ratio)**

$$R = \frac{\overline{d}_{obs}}{\overline{d}_{ran}}$$

R = degree of clustering

$$\overline{d}_{obs} = \frac{\sum d_i}{N}$$

d_i = distance to nearest neighbor of point i, and

N = # points

$$\overline{d}_{ran} = \frac{1}{2\sqrt{\rho}}$$

ρ = density of points per unit area

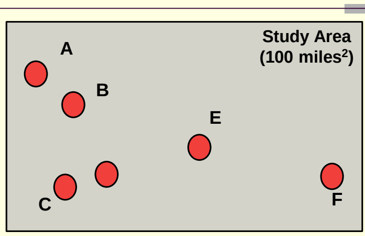
Nearest Neighbor Analysis

- **Identifying nearest neighbors: two methods**
  - *1. visually:* use this method on a simple map
  - *2. calculated:* use this method on a complex map; implement using a distance matrix derived from a GIS analysis of the point data
    - ArcGIS, for example, implements this analysis as a standard part of its geostatistical package

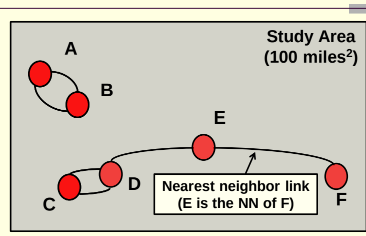
Nearest Neighbor Analysis

- **Following is a very basic example of the application of NNA**
  - Dealing with a simple distribution of 6 dots (places) on a map, with an identification of the nearest neighbor for each dot

Nearest Neighbor Analysis

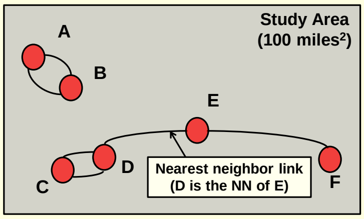


Nearest Neighbor Analysis



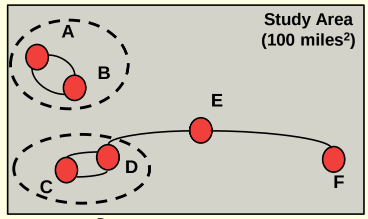
Nearest neighbor links

Nearest Neighbor Analysis



Nearest neighbor links

Nearest Neighbor Analysis



A and B
C and D } Reflexive nearest neighbors

Nearest Neighbor Analysis

Distance Matrix (Miles)

	A	B	C	D	E	F
A	0	2	8	6	7	8
B	2	0	6	4	5	6
C	8	6	0	3	7	6
D	6	4	3	0	4	6
E	7	5	7	4	0	5
F	8	6	6	6	5	0

Nearest Neighbor Analysis

Distance Matrix (Miles)

	A	B	C	D	E	F
A	0	2	8	6	7	8
B	2	0	6	4	5	6
C	8	6	0	3	7	6
D	6	4	3	0	4	6
E	7	5	7	4	0	5
F	8	6	6	6	5	0

Lowest values in each column
(nearest neighbors)

Nearest Neighbor Analysis

Distance Matrix (Miles)

	A	B	C	D	E	F
A	0	2	8	6	7	8
B	2	0	6	4	5	6
C	8	6	0	3	7	6
D	6	4	3	0	4	6
E	7	5	7	4	0	5
F	8	6	6	6	5	0

$\bar{d}_{obs}$ = (2+2+3+3+4+5)/6 = 3.17 miles

$\bar{d}_{ran}$ = 1/(2 $\sqrt{(6 \text{ points} / 100 \text{ miles}^2)}$) = 2.04 miles

Nearest Neighbor Analysis

So, finish the R calculation

$$R = \frac{\bar{d}_{obs}}{\bar{d}_{ran}} = \frac{3.17}{2.04} = 1.55$$

Interpretation:

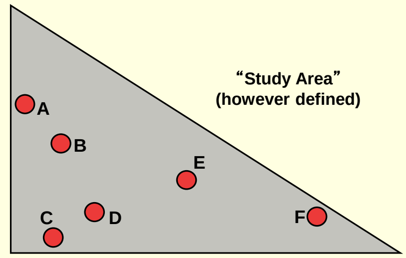
R=1.0: random
R=0.0: clustered
R=2.1491: dispersed (uniform)

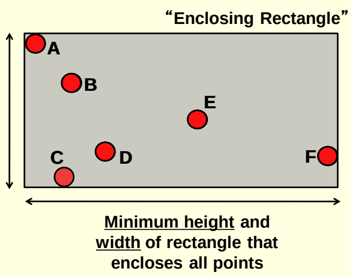
maximum possible

UNT

Nearest Neighbor Analysis

- **Problem with NNA: boundaries**
  - how we define the study area can impact the calculated R value
  - Options:
    - 1. use your initial "study area" (hopefully defined from the start using some physical, human, or ecological criteria that make sense for your study)
    - 2. use an objective measure like the "enclosing rectangle" (the smallest rectangle that encloses all study points) to give an "unbiased" definition (and use the same area throughout the study)





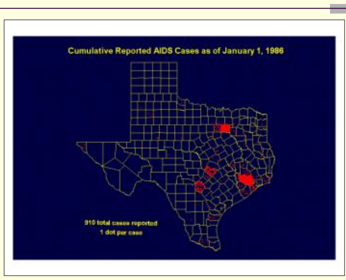
UNT

Nearest Neighbor Analysis

- **Problem with NNA: boundaries**
  - Caution: use the same area in every period for a time-based study (examining changes in a study area) – difficult to compare R values when the study area changes in some way

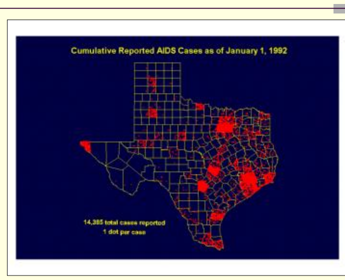
UNT

Nearest Neighbor Analysis

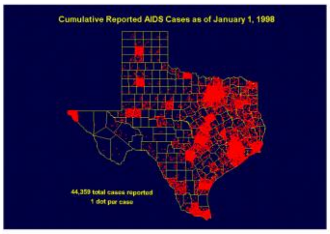


UNT

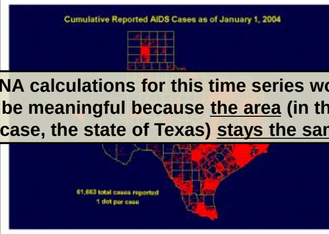
Nearest Neighbor Analysis



Nearest Neighbor Analysis



Nearest Neighbor Analysis



NNA calculations for this time series would be meaningful because <u>the area</u> (in this case, the state of Texas) <u>stays the same</u>

Nearest Neighbor Analysis

- **Key Question: Significance**
  - Is the pattern <u>significantly</u> different from random?
  - H_0: pattern is random (always have same H_0)
- **Options come in on your H_1 (choose one)**
  - H_1: pattern is <u>not random</u> (two-tailed test)
  - H_1: pattern is <u>not random</u> and is <u>clustered</u> (one-tailed test)
  - H_1: pattern is <u>not random</u> and is <u>dispersed</u> (one-tailed test)

Nearest Neighbor Analysis

- **Key Question: Significance**
  - Test Statistic ("Geary's C"):

$$C = \frac{(\bar{d}_{obs} - \bar{d}_{ran})}{SE_{\bar{d}}}$$

Standard error of the NN distance

Nearest Neighbor Analysis

- **Key Question: Significance**
  - Standard error calculation

$$SE_{\bar{d}} = \frac{0.26136}{\sqrt{N \times \rho}}$$

N = # points

ρ = density of points per unit area

Nearest Neighbor Analysis

- **Key Question: Significance**
  - Doing the calculation with the "6 dots on a map" example and the associated values calculated earlier

$$SE_{\bar{d}} = \frac{0.26136}{\sqrt{N \times \rho}} = \frac{0.26136}{\sqrt{6 \times 0.06}} = 0.43$$

  - Therefore, the value of the test statistic C is

$$C = \frac{\bar{d}_{obs} - \bar{d}_{ran}}{SE_{\bar{d}}} = \frac{3.17 - 2.04}{0.43} = 2.63$$

Nearest Neighbor Analysis

- **Key Question: Significance**
  - Compare the calculated *C* value with the critical value for the statistic (one tailed test, 0.05 level)

C10 Critical Values of a Standard Normal Deviate z

	Significance level (one-tailed)				
	0.1	0.05	0.01	0.005	0.001
z	1.282	1.645	2.326	2.576	3.090
−z	−1.282	−1.645	−2.326	−2.576	−3.090

	Significance level (two-tailed)				
	0.1	0.05	0.01	0.005	0.001
z	1.645	1.960	2.576	2.813	3.291
−z	−1.645	−1.960	−2.576	−2.813	−3.291

  - From the table, C_{Crit}=1.645 (remember C_{Calc}=2.63)
  - So, reject H_0 (C_{crit}<C_{Calc}): the pattern is significantly uniform and not random

Areal Patterns

- **Now, move along to our other situation where we can test spatial patterns: area patterns and "spatial autocorrelation"**
  - "Serial autocorrelation": what's the next number in the following series
    - 1, 3, 2, 4, 3, 5, 4, ___?
  - The next number is perfectly predictable because the series follows a sequence
  - Any one number is not independent of the other events in the series

Areal Patterns

- **"Spatial autocorrelation" is the same idea, extended from one to two dimensions**
  - **Are the area features on a map independent of each other?**

2	1	5	6
3	0	9	2
1	7	___	2
4	8	0	5

No autocorrelation
(independent values)

1	2	3	3
1	2	3	4
1	3	___	4
2	3	3	4

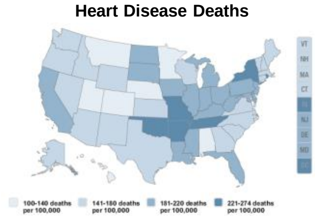
Autocorrelation
(dependent values)

Areal Patterns

- **Autocorrelation is undesirable in many research contexts outside of geography**
  - Often, researchers in other fields outside of geography want/ need independent data and try to eliminate autocorrelation
  - In geography, we are interested in pattern
  - We want to find spatial autocorrelation

Areal Patterns

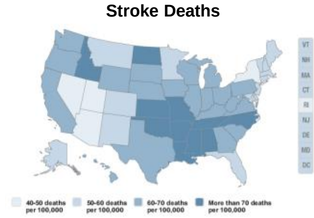
Heart Disease Deaths



Source: Centers for Disease Control and Prevention

Areal Patterns

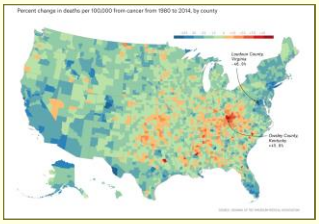
Stroke Deaths



Source: Centers for Disease Control and Prevention

Areal Patterns

Change in Cancer as a Cause of Death

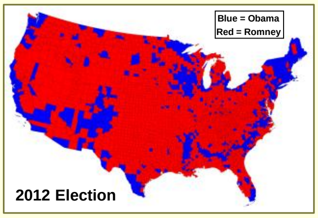


Areal Patterns

Areal Patterns

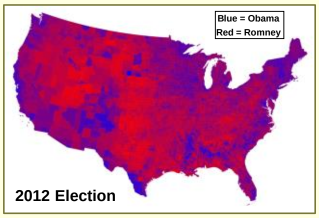
Areal Patterns

Presidential Election by County



Areal Patterns

Presidential Election by County: Another View

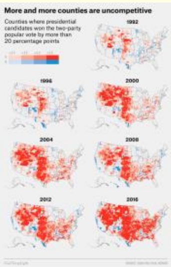


Areal Patterns

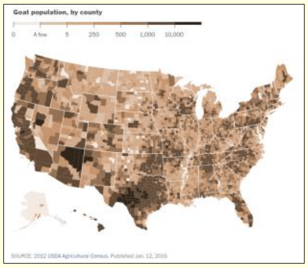
The 1992 – 2016 Presidential Elections

Uncompetitive Counties

National Map Time Series



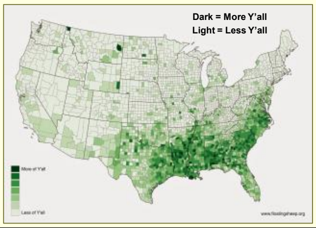
Areal Patterns



Areal Patterns

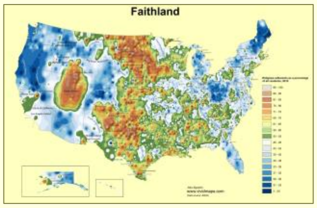
Areal Patterns

Use of “Y’all” on Twitter



Areal Patterns

Religious Adherents as % of Population

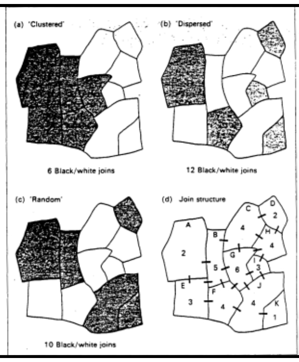


Areal Patterns

- **Spatial autocorrelation techniques can deal with two different kinds of data**
  - **1. dichotomous data:** area values are “yes/no”, “above average/below average”, or any other situation where there are <u>two possibilities</u> only
  - **2. continuous data:** area values can fall anywhere in a range of numbers

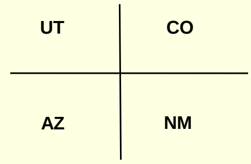
Dichotomous Data: Join Count

- **We will focus on the <u>dichotomous</u> situation first**
- **We measure spatial autocorrelation with dichotomous data using a methodology called “<u>join count statistics</u>”**
  - **Idea:** look at the “joins” connecting our two classes or kinds of areas, and where the two area classes are distributed within the study map



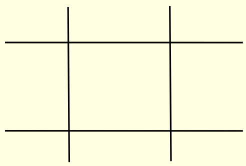
Dichotomous Data: Join Count

- **Key question: how do we define a "join"?**



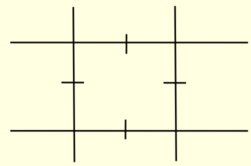
Dichotomous Data: Join Count

- **Possible Join Definitions**



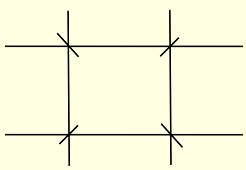
Dichotomous Data: Join Count

- **Rook's Case**



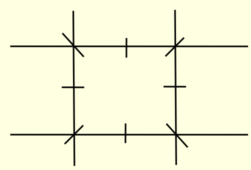
Dichotomous Data: Join Count

- **Bishop's Case**



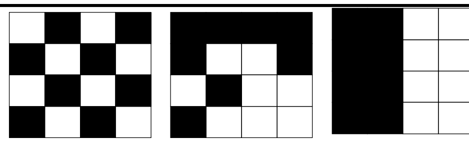
Dichotomous Data: Join Count

- **Queen's Case**



Dichotomous Data: Join Count

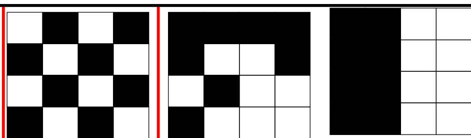
- **Rook's case is the most commonly used**
  - The others are certainly good options, depending on the nature of what you are studying
- **As we've seen, dichotomous variables are mapped in two colors (B & W)**
  - So, the *join* (or border) can be classified as WW, BB, or BW
  - BW joins form the basis for our join count statistic calculations, but looking at all types is good for understanding



A B C

Join Counts

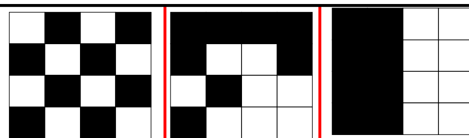
Map	BB	WW	BW	Tot.
A	0	0	24	**24**
B	5	7	12	**24**
C	10	10	4	**24**



A B C

Join Counts

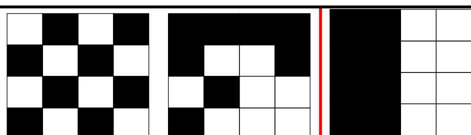
Map	BB	WW	BW	Tot.
A	0	0	24	**24**
B	5	7	12	**24**
C	10	10	4	**24**



A B C

Join Counts

Map	BB	WW	BW	Tot.
A	0	0	24	**24**
B	5	7	12	**24**
C	10	10	4	**24**



A B C

Join Counts

Map	BB	WW	BW	Tot.
A	0	0	24	**24**
B	5	7	12	**24**
C	10	10	4	**24**

Dichotomous Data: Join Count

- **Join Count Statistics: Rationale**
  - 1. Need to know the expected number of joins for a random distribution
  - 2. Compare this "random" value with the observed distribution

Dichotomous Data: Join Count

- **Same deal here as with NNA when we are thinking about possible join count hypotheses**
  - **H_0**: The pattern is random (null hypothesis)
  - **H_1**: The pattern is <u>not random</u> (two-tailed test)
  - **H_1**: The pattern is <u>not random</u> and is <u>clustered</u> (one-tailed test)
  - **H_1**: The pattern is <u>not random</u> and is <u>dispersed</u> (one tailed test)

Dichotomous Data: Join Count

- **Two options for measuring and testing area patterns for dichotomous data using "join count statistics"**
  - **Free sampling/normalization:** you know facts about a larger area, and these apply to your study area (in other words, the study area is a subset of a larger region)
  - **Non-free sampling/randomization:** the facts you use in this test only come from the study area itself (this is the most common of the two, so we will focus on this one here)

Dichotomous Data: Join Count

- **In both cases, the test statistic *z* is found with the equation**

$$z = \frac{O_{BW} - E_{BW}}{\sigma_{BW}}$$

O_{BW} = Observed # BW joins
E_{BW} = Expected # BW joins
σ_{BW} = Stand. dev. in BW joins

- **There is a different E_{BW} and σ_{BW} calculation for "free sampling" or "non-free sampling", so <u>decide in advance</u> what you will do so you don't confuse the two (see Ebdon reading for both calculations)**

Dichotomous Data: Join Count

- **Basic idea with the join count statistical test**
  - Calculate the *z* value for your study area
  - Compare the calculated *z* value with the sampling distribution

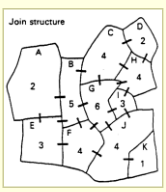
C10 Critical Values of a Standard Normal Deviate z

	Significance level (one-tailed)				
	0.1	0.05	0.01	0.005	0.001
z	1.282	1.645	2.326	2.576	3.090
−z	−1.282	−1.645	−2.326	−2.576	−3.090

	Significance level (two-tailed)				
	0.1	0.05	0.01	0.005	0.001
z	1.645	1.960	2.576	2.813	3.291
−z	−1.645	−1.960	−2.576	−2.813	−3.291

Dichotomous Data: Join Count

- **Some definitions**
  - **L** = # joins for each zone (in the example to the right, area "A" has 2 joins)
  - **J** = total # of joins among all zones in the study area (19 in the same example)
  - **B =** # of "black" zones
  - **W =** # of "white" zones
  - **n =** total # of zones (11 in the example)



Dichotomous Data: Join Count

- **For non-free sampling (most common)**
  - Equations to use include

$$z = \frac{O_{BW} - E_{BW}}{\sigma_{BW}}$$ From a few slides ago

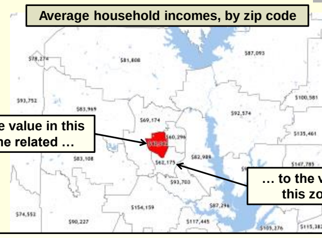
$$E_{BW} = \frac{2JBW}{n(n-1)}$$ Expected # BW joins

σ_{BW} = See p. 153 of Ebdon reading — Stand. dev. in BW joins

Continuous Data: Moran's I

- ***Moran's I* is one of the oldest indicators of spatial autocorrelation (Moran, 1950)**
  - Still the standard for determining spatial autocorrelation
  - Can actually be applied to either zones or points with continuous data variables
  - Compares the value of the variable at any one location/zone with the value at all other locations or zones

Continuous Data: Moran's I



Continuous Data: Moran's I

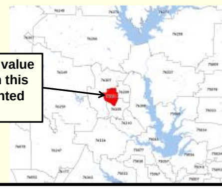
The *Moran's I* Equation:

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

- Where
  - N is the number of cases (zones/locations)
  - X_i is the variable value at a particular zone/location
  - X_j is the variable value at another zone/location
  - $\bar{X}$ is the mean of the variable, map-wide
  - W_{ij} is a matrix: set of weights applied to the comparison between one location (i) and another location (j)

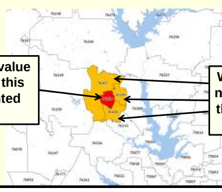
Continuous Data: Moran's I

- W_{ij} functions to define the "reach" of the comparisons used in the *Moran's I* calculation
  - 1. Could limit the equation's comparisons to contiguous (spatially touching) zones only



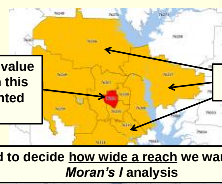
Continuous Data: Moran's I

- W_{ij} functions to define the "reach" of the comparisons used in the *Moran's I* calculation
  - 1. Could limit the equation's comparisons to contiguous (spatially touching) zones only



Continuous Data: Moran's I

- W_{ij} functions to define the "reach" of the comparisons used in the *Moran's I* calculation
  - 2. Or it could broaden the equation's range of comparisons to a wider range of zones



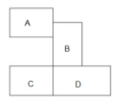
Continuous Data: Moran's I

- ***W_{ij}* is what we call a "contiguity matrix" or an "adjacency matrix"**
  - For strict contiguity (touching only): if zone *j* is adjacent to zone *i* (i.e. the zones border each other), then the cell (i, j) receives a weight of 1, otherwise it is a 0
  - A broader option could make W_{ij} a distance-based weight, based for example on the inverse of the distance between locations i and j ($1/d_{ij}$)
    - This means that close-by, but non-touching, zones can also be accounted for

Continuous Data: Moran's I

- ***W_{ij}* is what we call a "contiguity matrix" or an "adjacency matrix"**

Four-Zone Map



Corresponding Contiguity Matrix

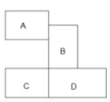
	A	B	C	D
A	0	1	0	0
B	1	0	1	1
C	0	1	0	1
D	0	1	1	0

Using Queen's Case: Strict Contiguity

Continuous Data: Moran's I

- ***W_{ij}* is what we call a "contiguity matrix" or an "adjacency matrix"**

Four-Zone Map



Corresponding Contiguity Matrix

	A	B	C	D
A	0	1	0.3	0.7
B	1	0	1	1
C	0.3	1	0	1
D	0.7	1	1	0

Using Queen's Case: Weighted Option

Continuous Data: Moran's I

- **Similar to the correlation coefficient, *Moran's I* varies between -1.0 and +1.0**
  - **Positive *Moran's I*:** clustering
  - **Negative *Moran's I*:** dispersion
    - Note, the expected *Moran's I* value of a perfectly random distribution would be

$\boxed{\textbf{-1/(N-1)}}$ where N = number of cases: close to zero, but slightly negative

Continuous Data: Moran's I

- **The same two options apply as with join count:**
  - 1. free sampling/normalization
  - 2. non-free sampling/randomization
- **The test statistic (also same idea exactly as with join count)**

$$z = \frac{I - E_I}{\sigma_I}$$

I = Observed Moran's I
E_I = Expected Moran's I
σ_I = Stand. dev. in Moran's I

Continuous Data: Moran's I

- **Six steps in the manual use of Moran's I**
  - 1. Calculate *I* (as outlined in previous slides)
  - 2. Calculate E_I

$$E_I = \frac{-1}{n-1}$$

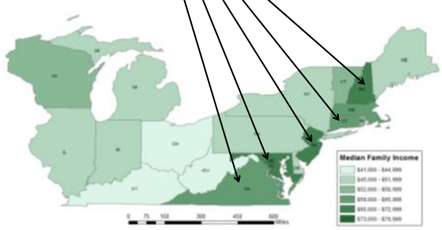
  - 3. Calculate σ_I (see p. 160 of Ebdon reading)
  - 4. Find *z* (remember that $z = (I - E_I)/\sigma_I$)
  - 5. Test significance
  - 6. Make a decision

Moran's I Example

- **Practical Moran's I application example**
  - "Spatial autocorrelation of incomes in the northeastern US"
  - To keep the example simple, we'll do the analysis at the state level
  - The following illustrates the Moran's I analysis procedure from *ArcGIS*, but other GIS packages offer similar capabilities
  - *TransCad* and *CrimeStat* are two other GIS packages that also offer powerful spatial statistics capabilities, including Moran's I

Basic question: are family incomes clustered at the state level in the northeastern US?

They look clustered (highest values along the coast), but are they actually clustered from a statistical perspective?



Six Steps of a Formal Statistical Test:

1. **H_0:** random pattern, **H_1:** clustered pattern
2. **Test statistic:** as already mentioned, the test statistic for the *Moran's I* calculation is the standard normal deviate *z*
3. **Significance level:**

- *Select p = 0.05:* a fairly usual level for scientific research, but not overly rigorous
- *One-tailed test:* because we already suspect the pattern might be clustered

---

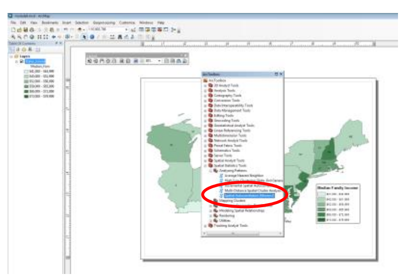
Six Steps of a Formal Statistical Test:

4. **Determine Critical value**
5. **Compute the test statistic**

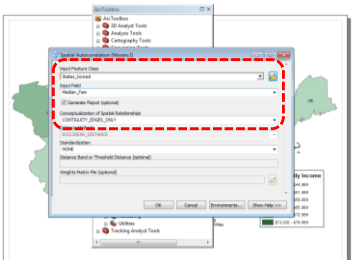
ArcGIS looks after both of these steps automatically

So starting with our basic state income map, we can complete our Moran's I calculation in ArcGIS…

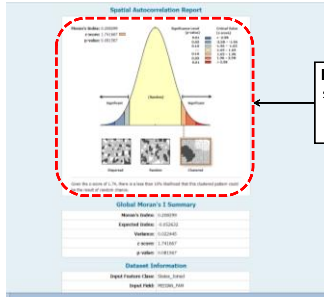
In ArcGIS: Locate the Spatial Autocorrelation Tool



Next: Select the layer, field, and spatial relationship

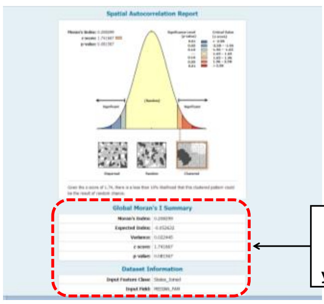


Lastly: View your results in the generated report



Helpful results summary with clarifying graphics

Lastly: View your results in the generated report

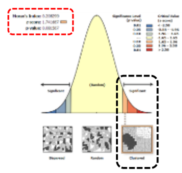


All the key stats and details from your analysis

Six Steps of a Formal Statistical Test:

6. **Make a decision**

Our ArcGIS report shows that our calculated z-score indicates a significance level (p-value) of 0.08



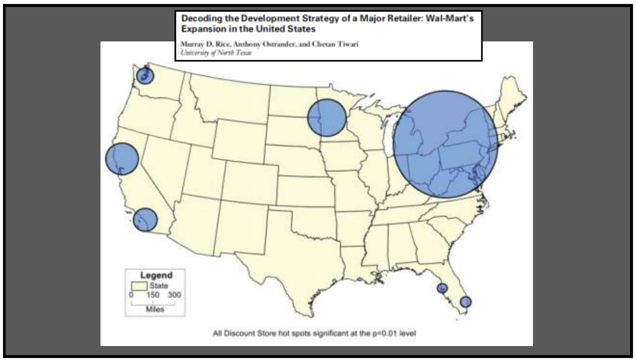
This result indicates a tendency toward clustering, but not at our selected 0.05 significance level (step 3): so we must accept H_0 and call the distribution random

Spatial Autocorrelation

- **Last point on spatial autocorrelation**
  - One caution for both forms of spatial autocorrelation relates to sample size
  - The bigger, the better
  - Use caution with small samples (small numbers of zones), as the statistics break down

Conclusion: Other Approaches to Spatial Pattern Measurement

- **If you're interested in further exploring more methods in this area, here are two more possibilities**
  - 1. Hot Spot Analysis (Getis-Ord Gi*): identification of statistically-significant clusters of high activity
    - Hot spot analysis defines the location and size of significant clusters



Conclusion: Other Approaches to Spatial Pattern Measurement

- **If you're interested in further exploring more methods in this area, here are two more possibilities**
  - 2. Local Moran's I Clustering: representation of local concentrations of a variety of statistically significant spatial behaviors within a larger study area
    - Areas of particularly high activity levels
    - Areas of particularly low activity levels