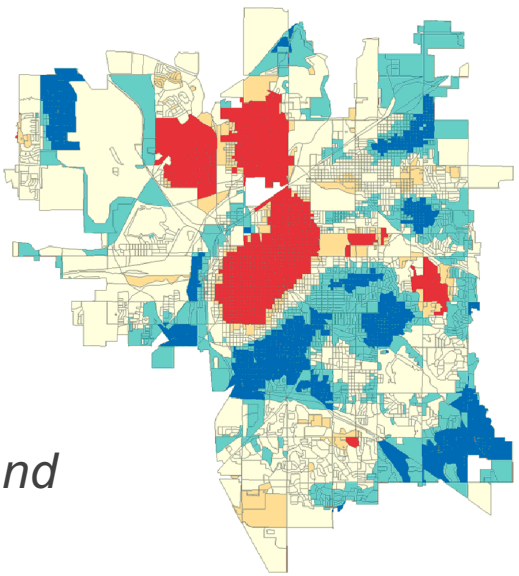
Spatial Autocorrelation



Dr. Beatriz Martínez López, DVM, MPVM, PhD.
Center for Animal Disease Modeling and Surveillance (CADMS) &
Dept. of Medicine & Epidemiology,
UC Davis

Date	Day	Time	Item	Description
2 Apr	Tue	2:10-3:00pm	Lec	Lec1A: Introduction & Health applications of spatial epidemiology
4 Apr	Thu	1:10-4:00pm	Lec/Lab	Lec1B: Spatial data characteristics and metadata Lab1: Introduction to ArcGIS
9 Apr	Tue	2:10-3:00pm	Lec	Lec2A: Coordinate systems and projections *Final project details today* ***Lab #1 due by 5pm***
11 Apr	Thu	1:10-4:00pm	Lec/Lab	Lec2B: Scale, MAUP, ecological fallacy Lab2: Coordinate systems and projections
16 Apr	Tue	2:10-3:00pm	Lec	Lec3A: Relational databases and joins ***Lab #2 due by 5pm***
18 Apr	Thu	1:10-4:00pm	Lec/Lab	Lec3B: Geocoding Lab3: Geocoding, relational databases and joins
23 Apr	Tue	2:10-3:00pm	Lec	Lec4A: Cartographic symbology ***Lab #3 due by 5pm***
25 Apr	Thu	1:10-4:00pm	Lec/Lab	Lec4B: Introduction to spatial statistics Lab4: Cartography and univariate statistics
30 Apr	Tue	2:10-3:00pm	Lec	Lec5A: Buffers ***Lab #4 due by 5pm***
2 May	Thu	1:10-4:00pm	Lec/Lab	Lec5B: Global tests for spatial clustering Lab5: Proximity analysis: buffering, nearest neighbors and point pattern analysis
7 May	Tue	2:10-3:00pm	Lec	Lec6A: Spatial autocorrelation ***Lab #5 and Final Project Outline due by 5pm***
9 May	Thu	1:10-4:00pm	Lec/Lab	Lec6B: Local tests for spatial clustering Lab6: Clustering and hot spot analysis
14 May	Tue	2:10-3:00pm	Lec	Lec7A: Clustering in space and time ***Lab #6 due by 5pm***
16 May	Thu	1:10-4:00pm	Lec/Lab	Lec7B: Clustering: review and practical applications Lab7: SatScan lab
21 May	Tue	2:10-3:00pm	Lec	Lec8A: Spatial exposures: interpolation with IDW, spline, Kriging and areal interpolation ***Lab #7 due by 5pm***
23 May	Thu	1:10-4:00pm	Lec/Lab	Lec8B: Spatial health outcomes: exploratory regression, ordinary least squares and geographically weighted regression Lab8: Interpolation and regression
28 May	Tue	2:10-3:00pm	Lec	Lec9A: Disease Mapping ***Lab #8 due by 5pm***
30 May	Thu	1:10-4:00pm	Lec/Lab	Lab9: Disease Mapping
4 Jun	Tue	2:10-3:00pm	Lec	Lec10A: Spread of infectious diseases ***Lab #9 due by 5pm***
6 Jun	Thu	1:10-4:00pm	Lec/Lab	Presentation of Final Projects
11 Jun	Tue	5:00pm	Project	***Final project due***

Statistical Problem

- In traditional statistics, two of the foremost assumptions/conditions are that samples must be random and independent
  - Randomness implies no systematic patterns beyond those explained by a statistical model
  - Independence implies that the outcome of one event does not impact the probability of the outcome for a second event (e.g. rolling dice)
- Spatial data almost always violate these fundamental requirements
  - e.g. the temperature in Davis is likely to be similar to the temperature in Sacramento
  - Attributable to spatial autocorrelation

Spatial Autocorrelation

- If there is any systematic pattern in the spatial distribution of a variable, it is said to be spatially autocorrelated
- It measures the extent to which the occurrence of an event in a spatial unit alters the probability of events in neighboring units
- It is, conceptually as well as empirically, a two-dimensional form of redundancy

Spatial Autocorrelation

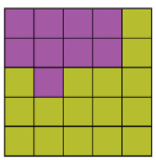
- With such redundancy in the data, each new data point provides less info than expected based on *n*, the sample size
  - Therefore, we should assess the degree of spatial autocorrelation before doing any conventional statistics at all
    - Diagnostic measures include: Join count statistics, Moran's I, Geary's C, Ripley's K, variogram cloud, etc.

Why should we assess spatial autocorrelation?

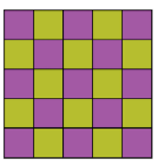
- A potential consequence of ignoring the spatial dependence in a statistical analysis is to underestimate errors and to overestimate statistical significance levels
- Increases the risk of making a Type I error
  - Reject H_0 when it is true (False Positive)

Spatial Autocorrelation

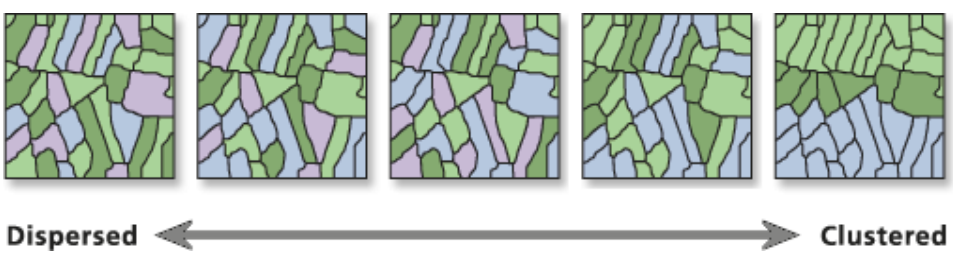
- Positive Autocorrelation (clustering)
  - Most common – situations where nearby observations are likely to be similar to one another (e.g. elevation)



- Negative Autocorrelation (dispersion)
  - Less common – observations from nearby locations are likely to be different from one another (e.g. standard eight-by-eight checkerboard pattern)



- Zero Autocorrelation
  - No spatial effect is discernible (random)



ILA LÓ

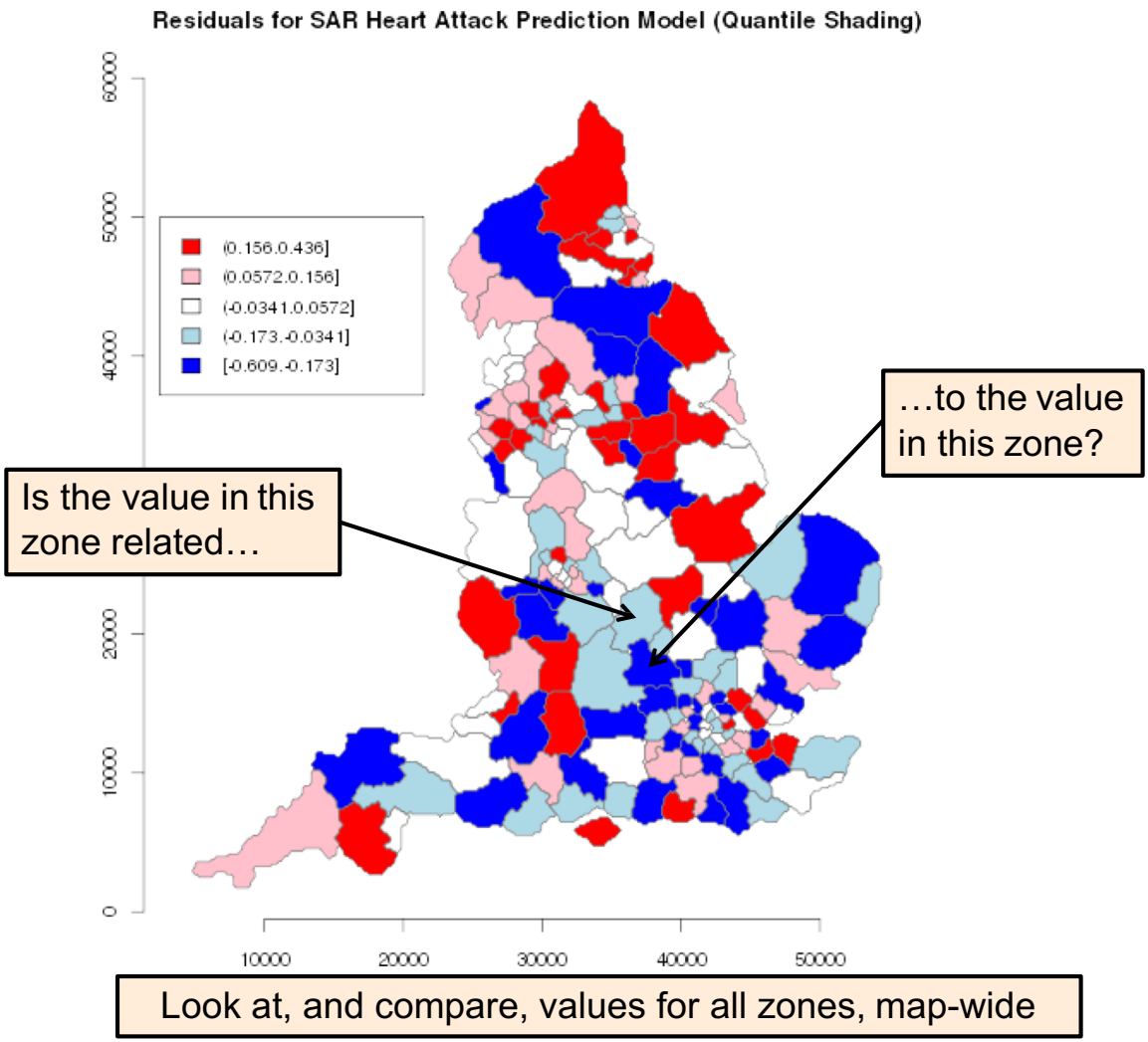
Tests for spatial autocorrelation

- Methods for aggregated data
  - Moran's I.
- Methods for point data
  - Moran's I.
  - Cuzick and Edwards' k-nearest neighbor test
  - Ripley's K-function

check you residuals from Moran's I → if you have patterns you need to do something to "correct" the data

Moran's I

- One of the oldest indicators of spatial autocorrelation (Moran, 1950). Still a *de facto* standard for determining spatial autocorrelation
- Applied to **polygons** or **points** with **continuous** variables associated with them
- Compares the **value** of the variable at any one location/zone with the value at all other locations or zones

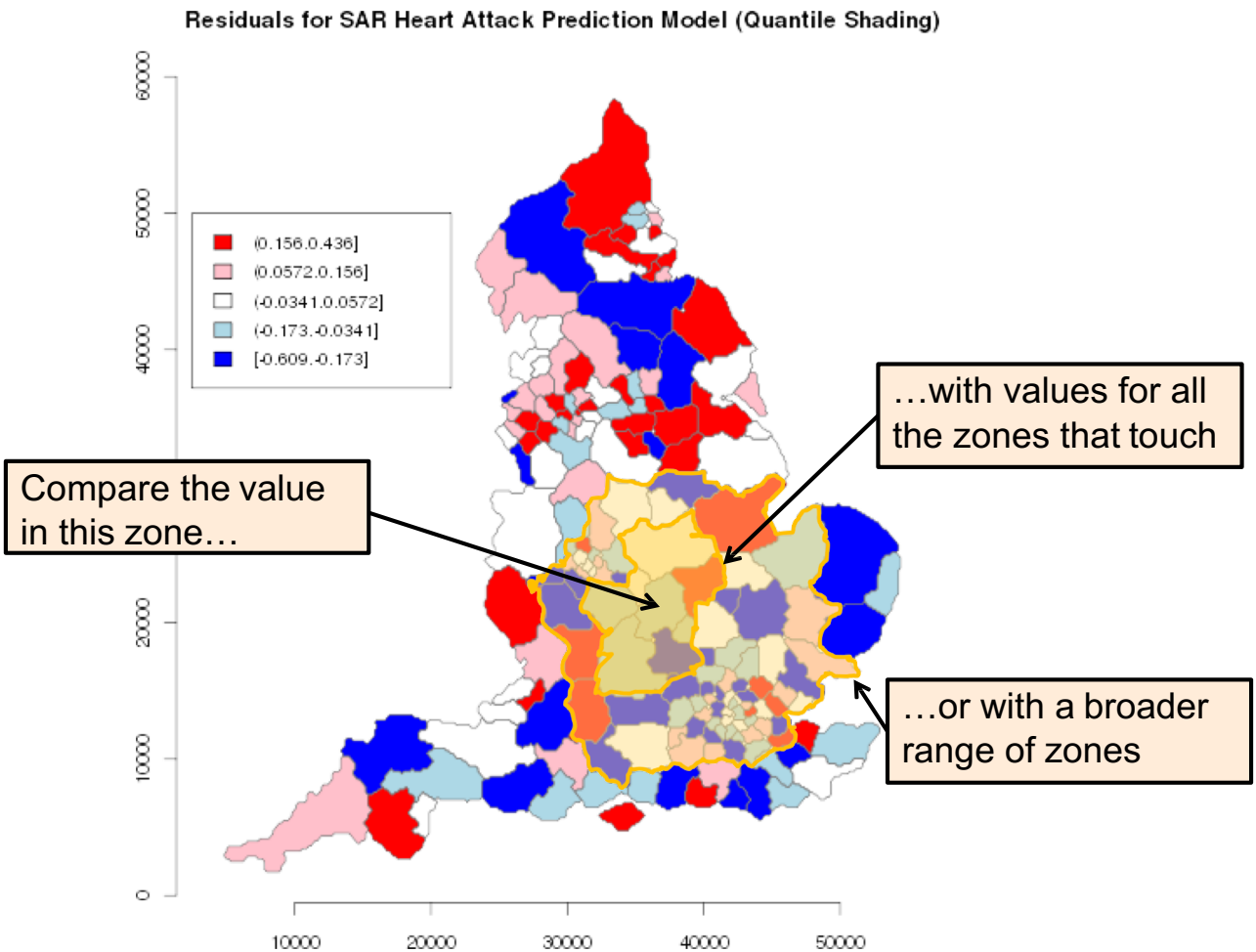

Residuals for SAR Heart Attack Prediction Model (Quantile Shading)
(0.156,0.436]
(0.0572,0.156]
(-0.0341,0.0572]
(-0.173,-0.0341]
[-0.609,-0.173]
60000
50000
40000
20000
10000
0
10000
20000
30000
40000
50000
Is the value in this zone related…
…to the value in this zone?
Look at, and compare, values for all zones, map-wide

The Moran's I Equation

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

- Where
  - n is the number of spatial features
  - x_i is the variable value at a particular location
  - x_j is the variable value at another location
  - $\bar{x}$ is the mean of the variable, area-wide
  - ω_{ij} is a weight applied to the comparison between one location, *i*, and another location, *j*

↳ if you are my neighbor = 1
if you are not " " = 0

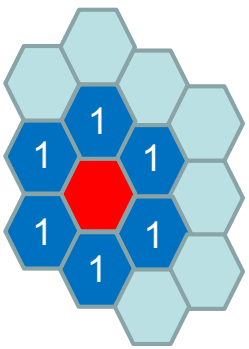

Residuals for SAR Heart Attack Prediction Model (Quantile Shading)
(0.156,0.436]
(0.0572,0.156]
(-0.0341,0.0572]
(-0.173,-0.0341]
[-0.609,-0.173]
60000
50000
40000
20000
10000
0
10000
20000
30000
40000
50000
Compare the value in this zone…
…with values for all the zones that touch
…or with a broader range of zones

The Spatial Weights Matrix

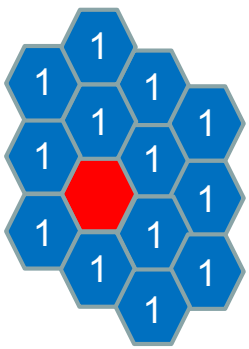
- May be called the adjacency matrix or contiguity matrix for contiguity-based definitions
  - **For strict contiguity (touching only)**: if zone *j* is adjacent to zone *I*, then cell (*i,j*) receives a weight of 1, otherwise it is a zero
  - **The broader option** would make ω_{ij} a distance-based weight, calculated, for example, on the inverse of the distance between locations *i* and *j* ($1/d_{ij}$)
    - Bigger distance →smaller number (weight)
    - This means that close, but non-touching zones can also be accounted for

The Spatial Weights Matrix

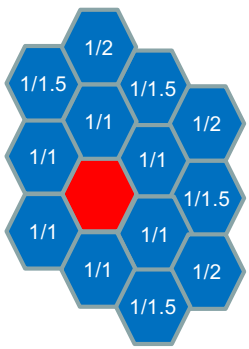
1st order contiguity



2nd order contiguity



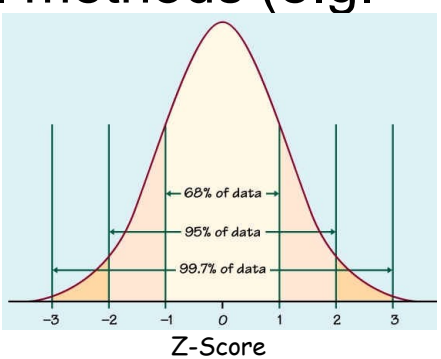
Inverse distance



- ω_{ij} functions to define the "reach" of the comparisons used in the Moran's I calculation
  - Could limit the equation's comparisons to contiguous (spatially touching) zones only
  - Or, it could widen the equation's comparisons to a broader range of zones

Moran's I Interpretation

- Similar to a correlation coefficient, it varies between -1.0 and +1.0
  - When autocorrelation is high, the coefficient is high
  - A high value indicates positive autocorrelation or clustering
- Significance is tested using a Z score
- If you have significant autocorrelation in your data, then you can use other spatial methods (e.g. hot spot analysis)



The test statistic (z)

$$z_I = \frac{I - E[I]}{\sqrt{V[I]}}$$

- Where

  I = Observed Moran's I

  $E[I]$ = Expected Moran's I = $-1/(N-1) \sim 0$

  $\sqrt{V[I]}$ = Standard error of Moran's I

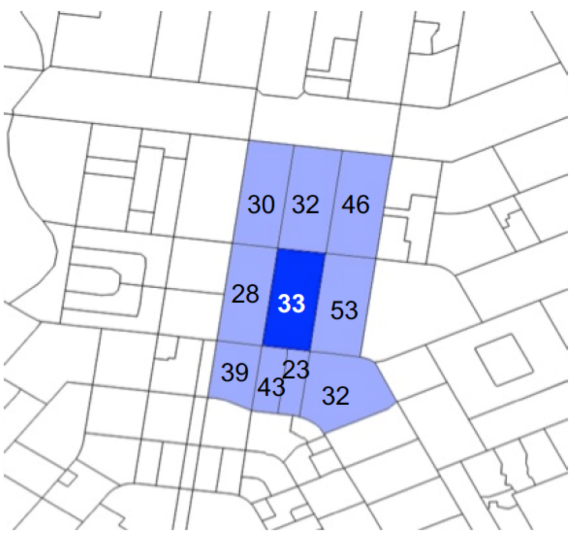
Example: Calculating Moran's I

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

- $\bar{x}$ = area-wide mean = 35
- Contiguity-based weights, so 1 for each pairing with immediate neighbors

$1\times(33-35)\times(32-35)$ $1\times(33-35)\times(43-35)$
$1\times(33-35)\times(46-35)$ $1\times(33-35)\times(39-35)$
$1\times(33-35)\times(53-35)$ $1\times(33-35)\times(28-35)$
$1\times(33-35)\times(32-35)$ $1\times(33-35)\times(30-35)$
$1\times(33-35)\times(23-35)$

center poly – mean of all poly; contiguous poly – mean poly

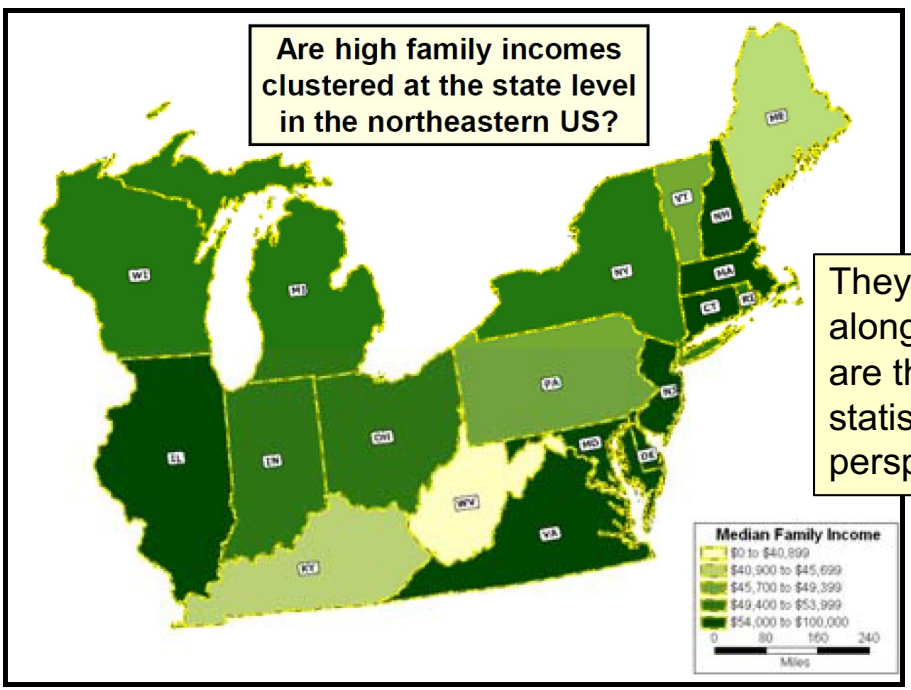


Moran I Assumes:
- Z-score : Normal distribution of the data. you need to transform the data 1st.

Moran's I Example

- Spatial autocorrelation of incomes in the northeastern US
- For simplicity, using state-level data

small sample size of polygons.
Edge effect can also be an issue.



They look clustered along the coast, but are they from a statistical perspective?

Moran's I Example

- Run Moran's I tool on 'Median Family Income'
- Results:
  - Moran's I: 0.090668
  - Expected value: -0.055556
  - Std Error: 0.159114

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

Z = 0.918989

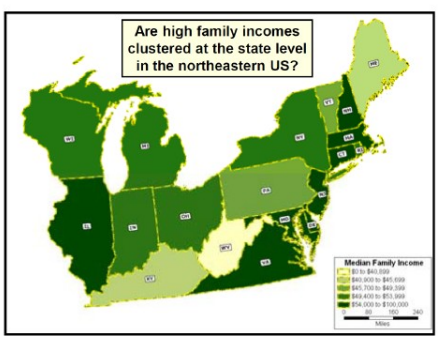
Moran's I Example

- Are high family incomes correlated?
  - Use the 0.05 level and a one-tailed test (z=1.645)
  - H_0: the pattern is random
  - H_1: The pattern is not random and is clustered

1.645 > 0.918986
So we accept H_0 – incomes are random and not clustered

Moran's I: Caution!

- Autocorrelation is related to sample size
  - The bigger the better
  - With small samples (few points or zones), the statistics can break down



only 19 zones !!

Compare Methods for Measuring the Pattern of Feature Locations

Test	Data Type	Evaluate	Result	Caveats
Moran's I	Continuous data, associated with polygons or points	Similarity of values between neighboring features	Statistic to evaluate spatial autocorrelation globally	Doesn't indicate whether clustering is for high or low values
Nearest Neighbor Index	Point patterns	Calculating the average distance between features	Nearest neighbor index, z-score	Results may be biased if there are many features near edge of study area
Ripley's K Function	Point patterns	Counting the number of features within defined distances	K-function, Uses multiple simulations to create a random distribution envelope	Patterns are suspect at larger distances due to edge effects

Acknowledgments

- "The ESRI Guide to GIS Analysis: Volume 2: Spatial Measurements and Statistics," by Andy Mitchell
- ArcGIS Help and ESRI online course materials
- "Geographic Information Analysis," by O'Sullivan and Unwin
- "Moran's I and Geary's C" by Arthur J. Lembo, Jr. of Cornell University
- "GIS Tutorial II: Spatial Analysis Workbook" by David W. Allen
- "Spatial Statistics: Spatial Pattern and Spatial Autocorrelation," University of North Texas
- "GIS Tutorial for Health," by Kristen S. Kurland and Wilpen L. Gorr
- Este Geraghty, MD, MS, MPH, ESRI
- Hugh Howard, PhD, American River College

QUESTIONS?

beamartinezlopez@ucdavis.edu