CRP 4080: Introduction to Geographic Information Systems for planners

Lecture 9: Introduction to spatial statistics

Wenzheng Li, Ph.D.
City and Regional Planning
Fall 2024

Announcement

This week:

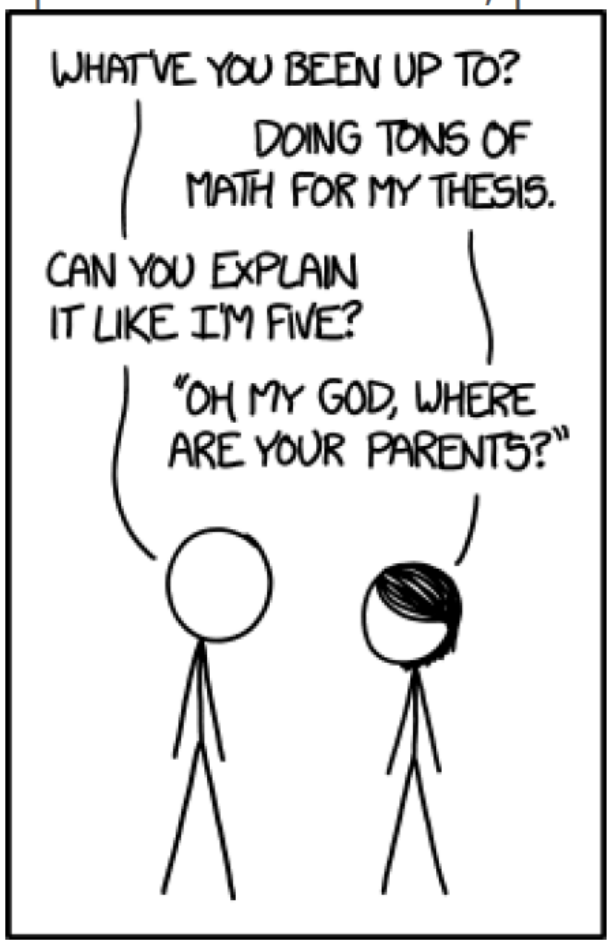
- Office hour this week:
  - Tuesday 3-6pm by appointment in Sibley 214
  - Thursday 5-6pm in Lab
- Lab 8 due this Friday

Next week:

- Gauri will lead Thursday's lab session
- Office hours:
  - Monday 5:00-6:30pm in Lab
  - Tuesday 3-6pm in Sibley 214

What are spatial statistics?

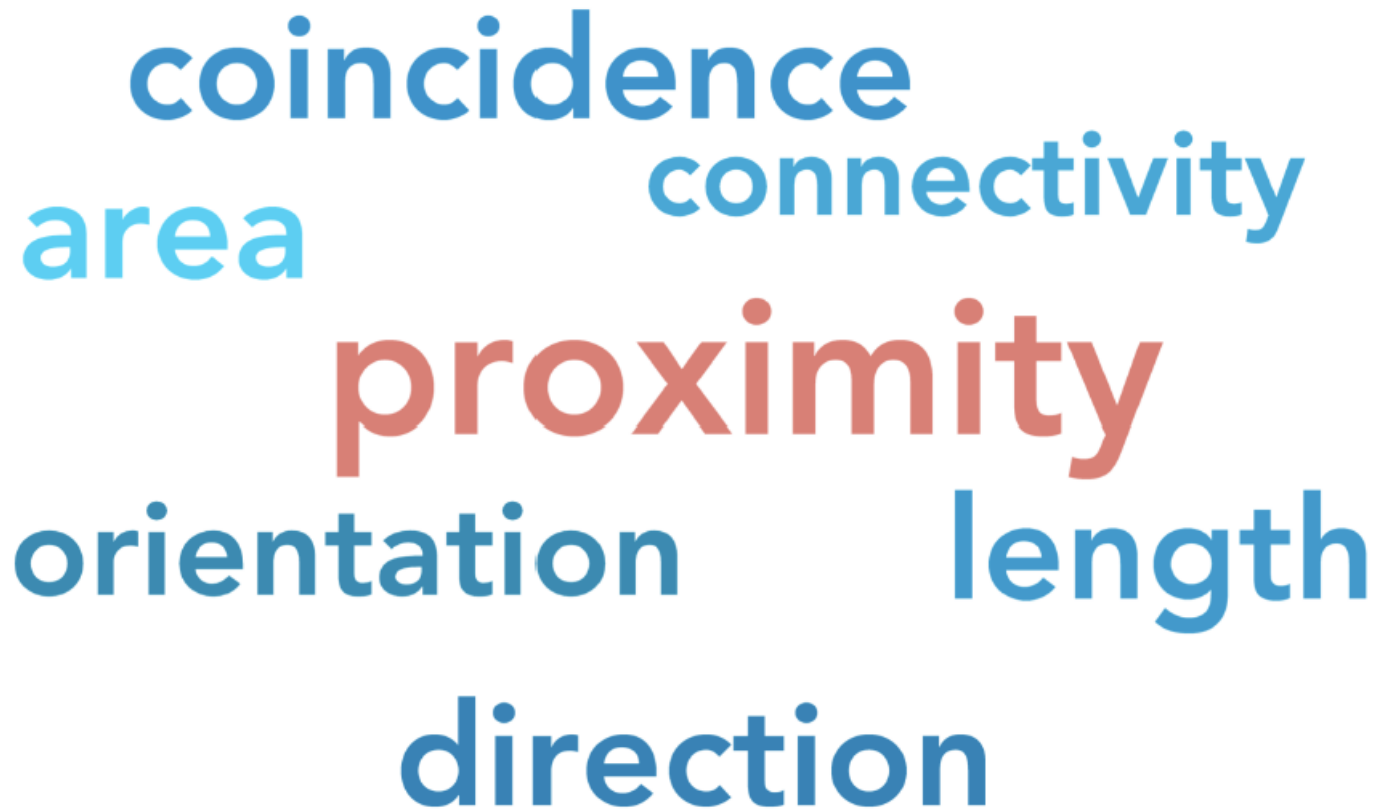
Spatial Statistics are a set of exploratory techniques for describing and modeling spatial distributions, patterns, processes, and relationships.



- Quantify pattern/relationships. Probability that a pattern/relationship actually exists (vs. random chance)
- Compare feature sets and track changes over time
- Its foundations are maths and inferential statistics.

The uniqueness about spatial statistics

Traditional statistics don't account for spatial relationships!!!



The uniqueness about spatial statistics

- Location: where things happen matters!
- **Dependence is the rule**: spatial interaction (contagion), spatial externalities (spillovers)

- scale matters: local vs. global — individual vs. total

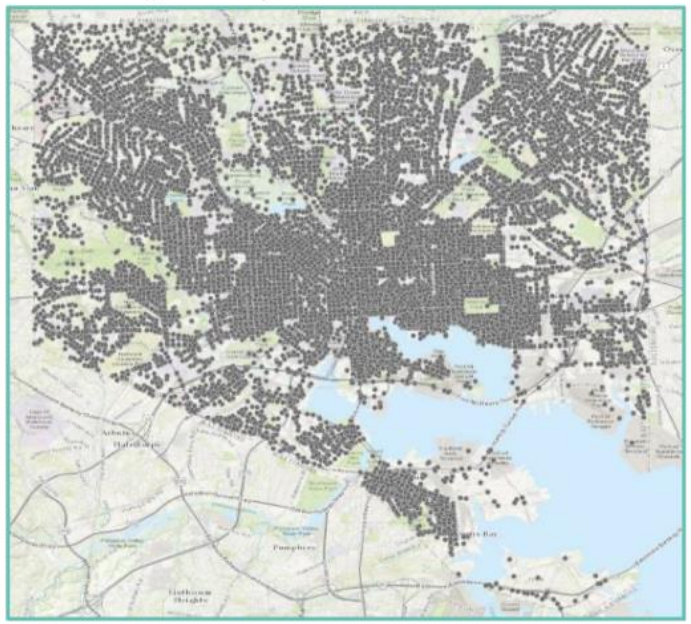
The outcome changes when the locations of the objects under study changes. CONTEXT MATTERS!

Geographic analysis with statistics

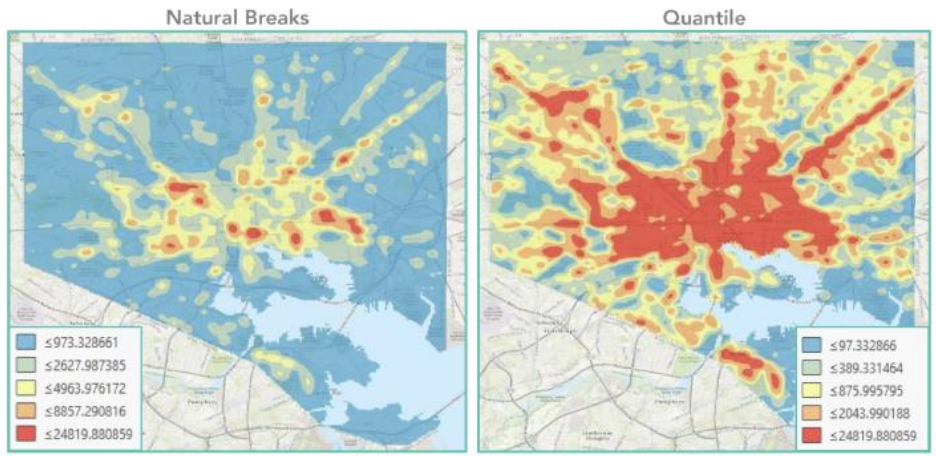
- Descriptive — A question: “What, Where?”
- Inferential statistics — A hypothesis: “How, Why?”

The map as data

High Priority 911 Calls in Baltimore



The subjectivity of visual pattern analysis



Where are the hot spots? Where is the variation greater?

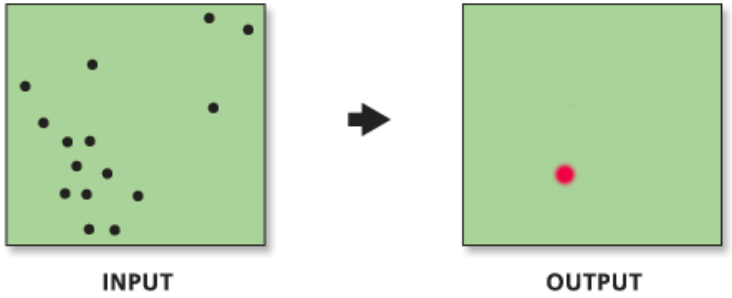
Measuring Geographic distribution: *Centrality*

- *Central Feature:* Identifies the most centrally located feature (having the shortest distance to all other features)

- *Mean Center:* Identifies the geographic center, or the center of concentration, for a set of features

Measuring Geographic distribution: *Centrality*

Mean center:



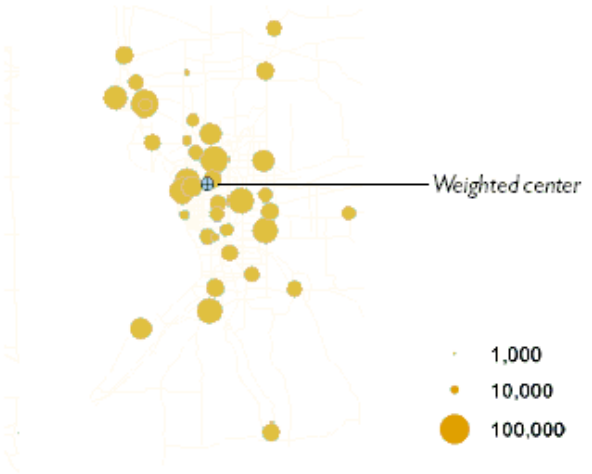
$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \quad , \quad \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

average x- and y-coordinate of all the features in the study area

Weighted Mean center:

$$\bar{X}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad , \quad \bar{Y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

weight at feature i.



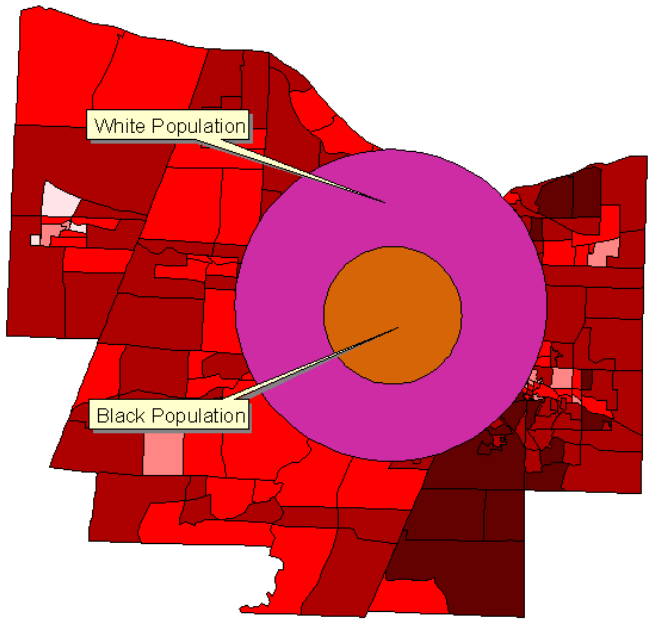
The center of grocery stores weighted using store size (in square feet).

Mean Center of Population for the United States: 1790 to 2020

U.S. Census Bureau, 2020 Census

Measuring Geographic distribution:
Dispersion and directionality

Standard Distance: Measures the degree to which features are concentrated or dispersed around the points of an input feature. Equivalent to the standard deviation in aspatial diagnostics. Recall that +/- standard deviation from the mean in a normal distribution encompasses 68% of the observations.



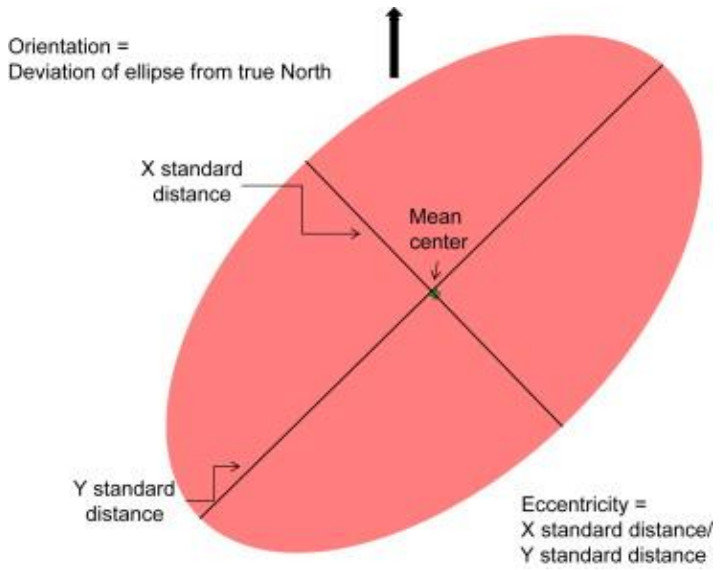
Measuring Geographic distribution: Dispersion and directionality

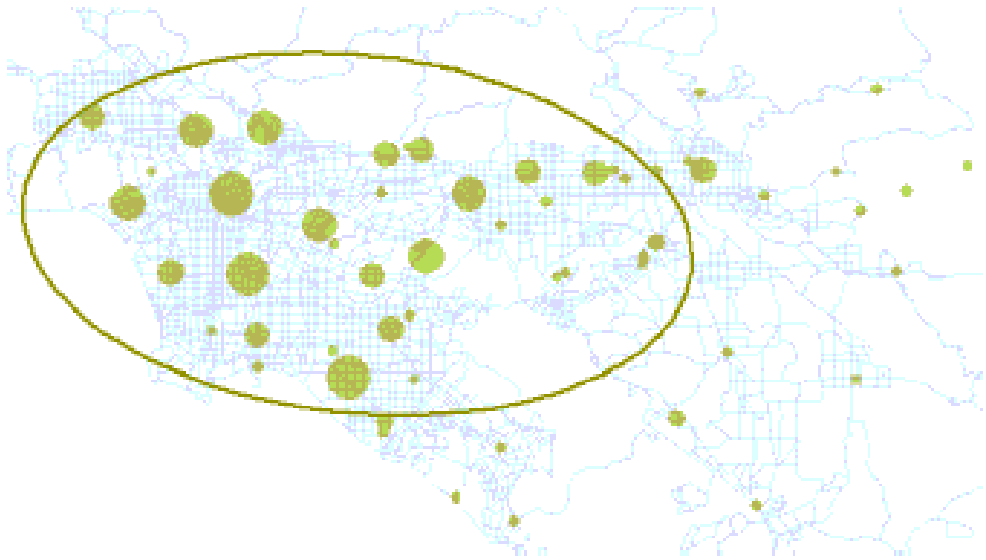
The Standard Deviational Ellipse is given as:

$$SDE_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}}$$

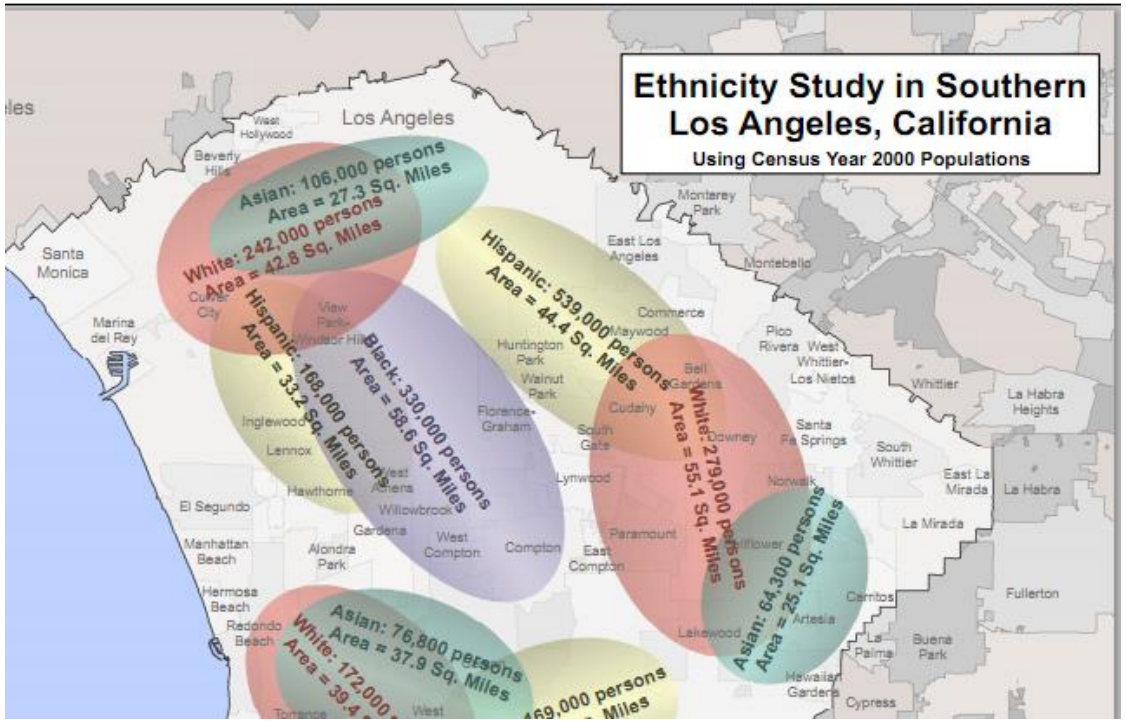
$$SDE_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}}$$

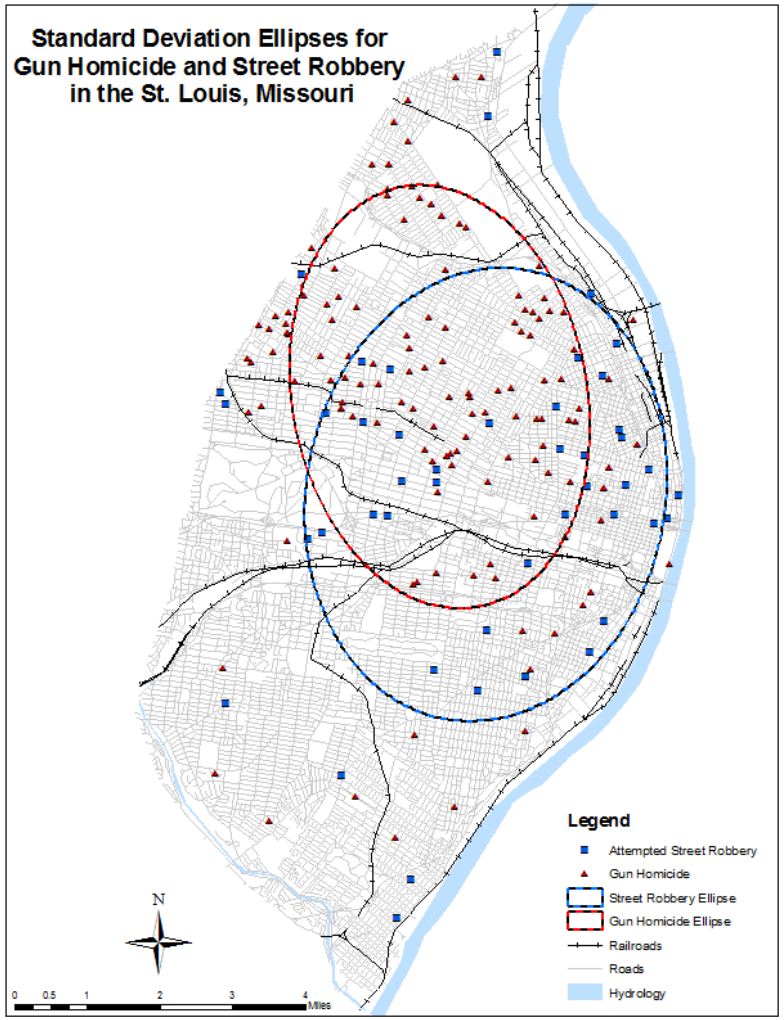
- Directional Distribution (*standard deviational ellipse)*: Measures whether a distribution of features exhibits a directional trend (whether features are farther from a specified point in one direction than in another direction).
- calculates the standard deviation of the x-coordinates and y-coordinates from the mean center to define the axes of the ellipse.





Measuring Geographic distribution: **Dispersion and directionality**

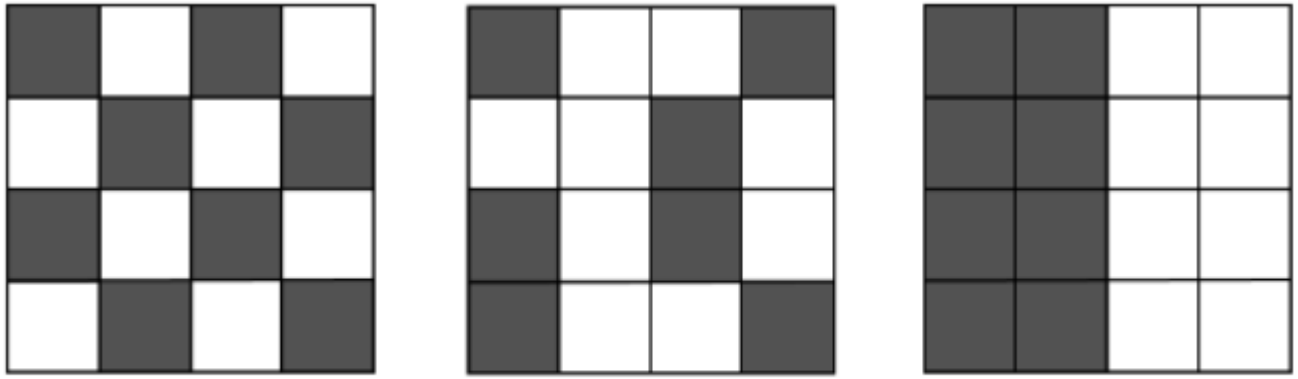




Anisotropic: directionally dependent (as opposed to isotropic)

Spatial Autocorrelation (or spatial dependence)

- When nearby values are more similar (or divergent) than we would expect given a ***spatially random process*** *->* *spatial* ***independence***
  - ***positive*** *spatial autocorrelation:* events clustered together
  - ***negative*** *spatial autocorrelation:* When events are dispersed (or Dissimilar values are clustered together)



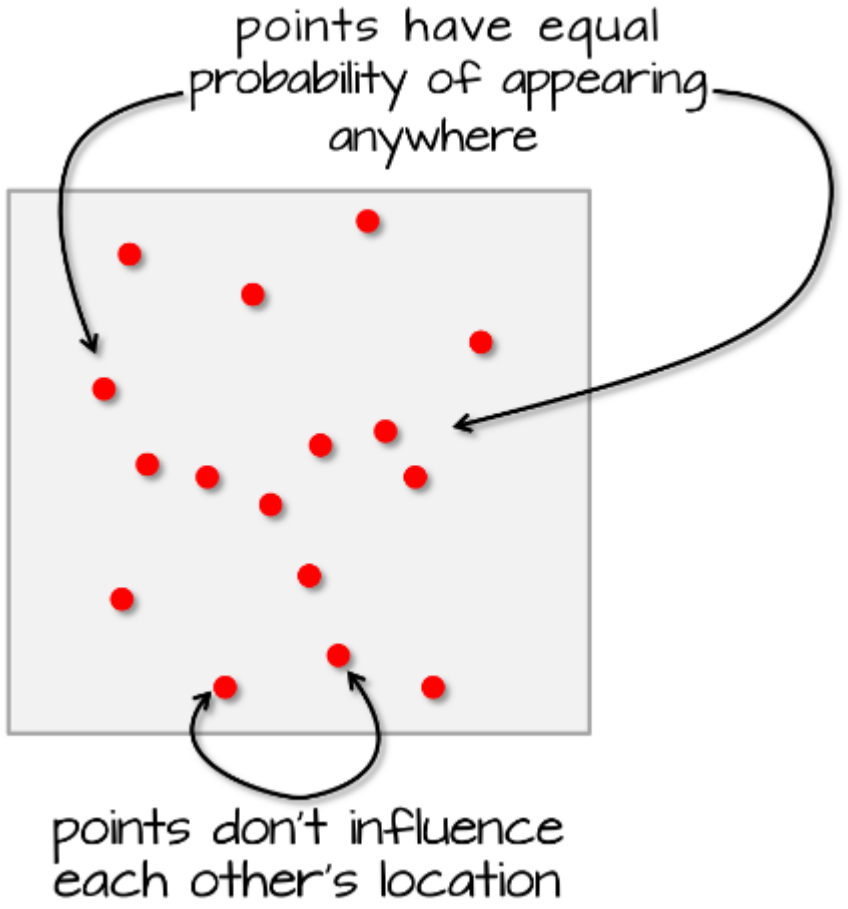
Negative spatial autocorrelation | No spatial autocorrelation | Positive spatial autocorrelation

Most social phenomena are ***positively*** *spatially autocorrelated*

Complete spatial randomness (CSR)

1. Equal probability: any feature has equal probability of being in any position

2. Independence: the positioning of any feature is independent of the positioning of any other feature.



Global indicator of Spatial Autocorrelation

- **Moran's I statistic**: a spatial statistic measures ***overall spatial autocorrelation*** of a dataset--whether similar values are clustered, dispersed, or randomly distributed
  - measures spatial autocorrelation based on both feature locations and feature values.
  - Calculation based on the variable X and the "***spatial lag***" of X formed by averaging all the values of X for the neighbors.
  - Neighbors' information are stored in a ***spatial weight matrix***.

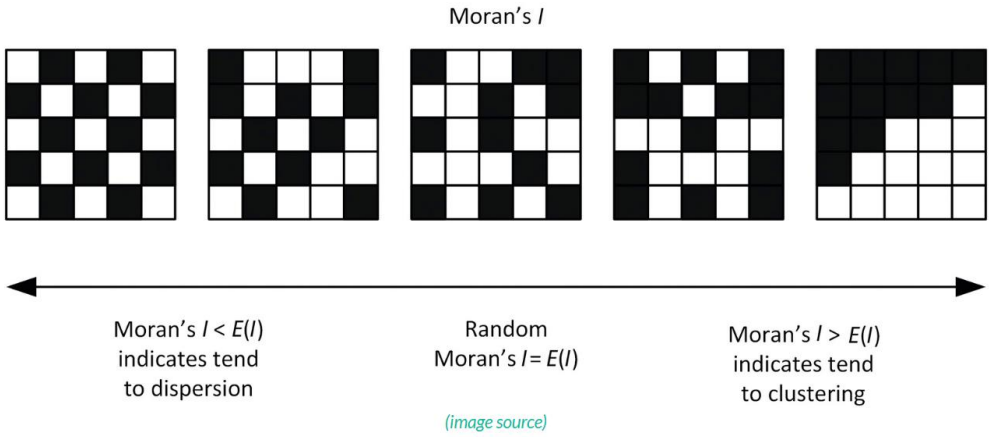
Moran's *I* is defined as

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

where N is the number of spatial units indexed by i and j;

X is the variable of interest; $\bar{X}$ is the mean of X;

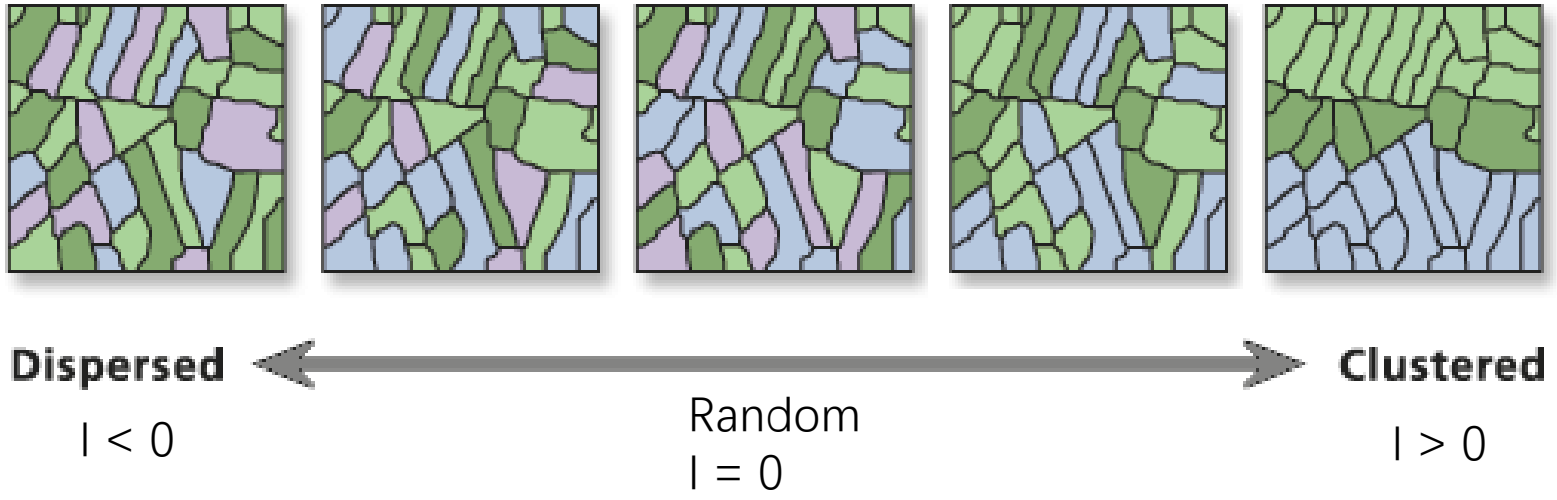
and w_{ij} is an element of a matrix of spatial weights.



(image source)

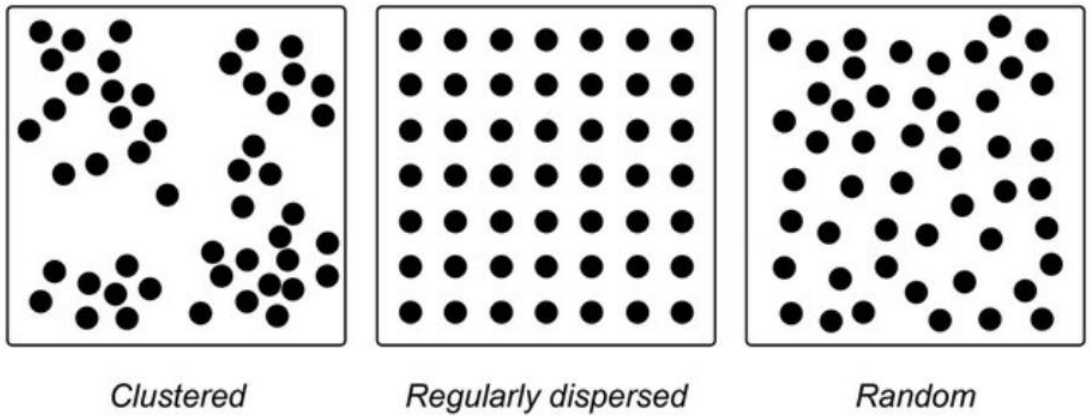
Interpreting Moran’s I

- Moran’s I evaluates whether the pattern is clustered, dispersed, or random
- a positive Moran's I index value indicates tendency toward clustering while a negative Moran's I index value indicates tendency toward dispersion. (ranges from -1 to 1).
- Moran's I = 0 (the attribute being analyzed is randomly distributed among the features in your study area (*Complete Spatial randomness*)

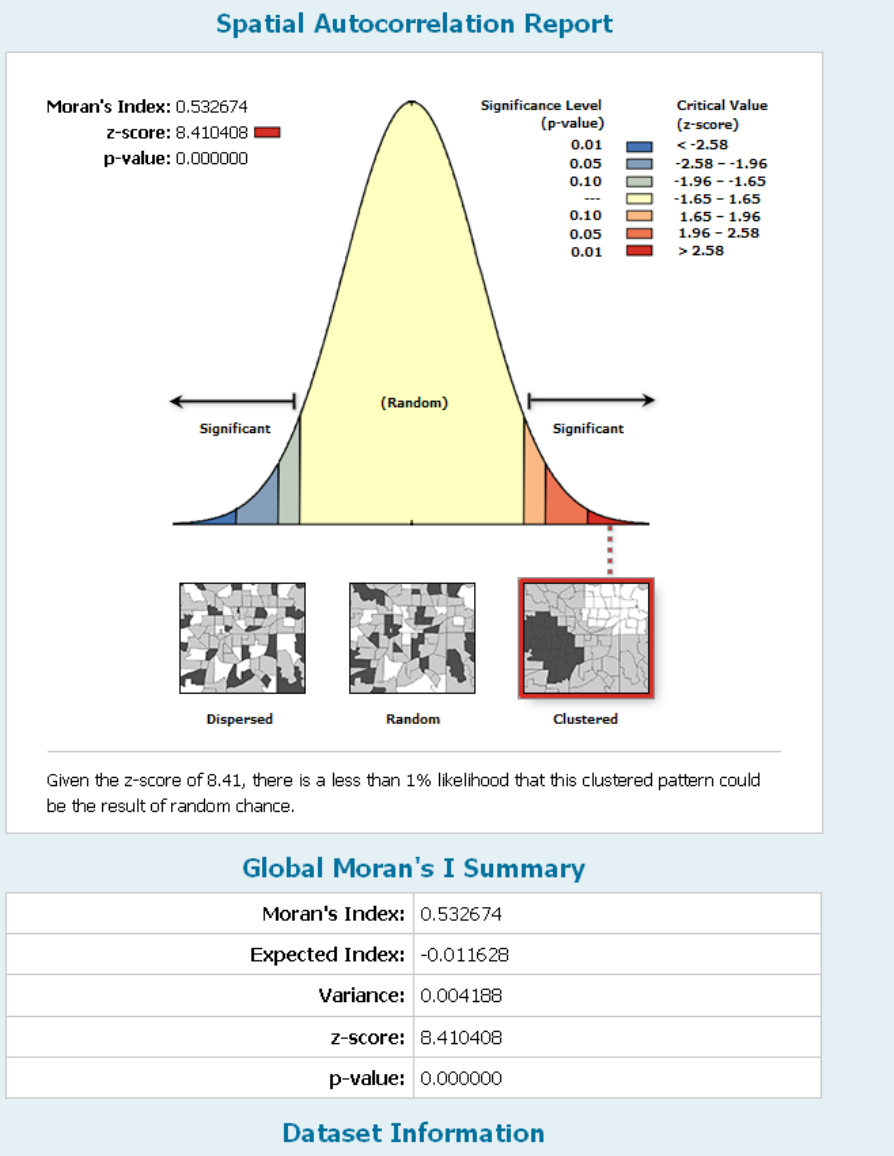


Inferential Spatial Statistics
Null and Alternative Hypotheses

- *Null Hypothesis:*
  - The spatial pattern is **random**
  - *CSR (complete spatial randomness):* The theoretical notion that our data actually represent just one realization of a very large number of possible outcomes
- *Alternative Hypothesis:*
  - The spatial pattern is **not random**
  - It may be *clustered* or *dispersed*



Interpreting Moran’s I



Moran’s Index: 0.532674 — positive Moran's I value indicates a strong clustering pattern, meaning similar values (e.g., high or low) tend to be near each other.

Significance Level: This report displays the distribution curve, with significant results shown at the tail ends. The clustering pattern falls in the right tail.

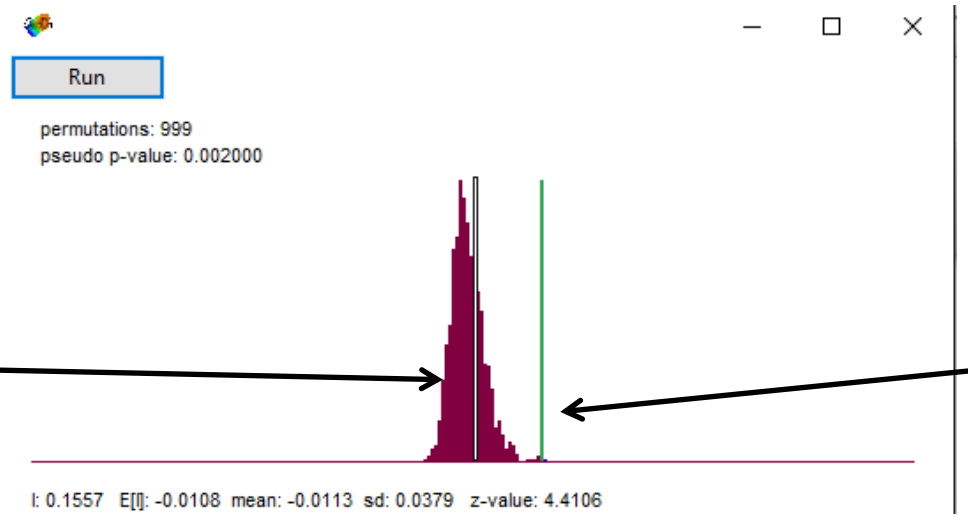
- Z-Score: 8.410408 — A high z-score (well above 1.96) implies that the clustering is far from random (at least at the 10% significance level).
- P-Value: 0.000000 — A very low p-value (< 0.01) indicates that there is less than a 1% likelihood that this observed clustering pattern is due to random chance (statistical significance).

Expected Index and Variance: The expected Moran's I index (-0.011628) and variance (0.004188) serve as baselines to compare the observed Moran's I, The observed value is considerably higher than expected under random spatial distribution.

Spatial Statistical Hypothesis Testing(Moran's I)

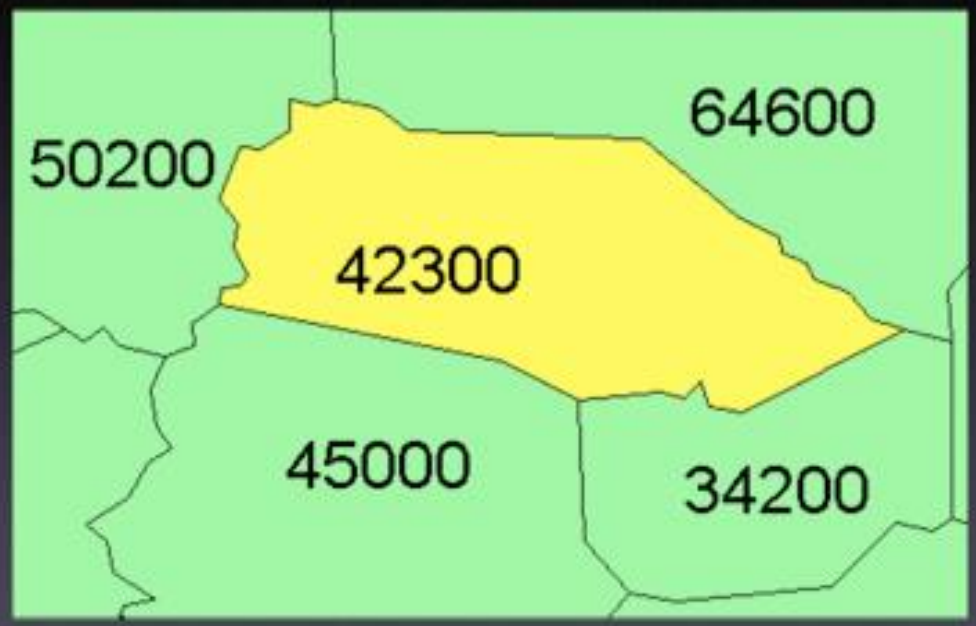
- Because of the complexity of spatial processes, it is often difficult to derive a test statistic with a known probability distribution
- Instead, we often use a ***Monte Carlo simulation***: We take multiple samples from a **random** spatial pattern. Moran's I is calculated for each sample, and then a frequency distribution is drawn
- This *simulated sampling distribution* is used to measure the probability of obtaining our actual observed spatial statistic

Empirical frequency distribution from 1000 random patterns ("samples")



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

Spatial Lag Example



Average Neighbor Land Values

1/4x50200 + 1/4x45000 + 1/4x34200 + 1/4x64600

Spatial weights matrix

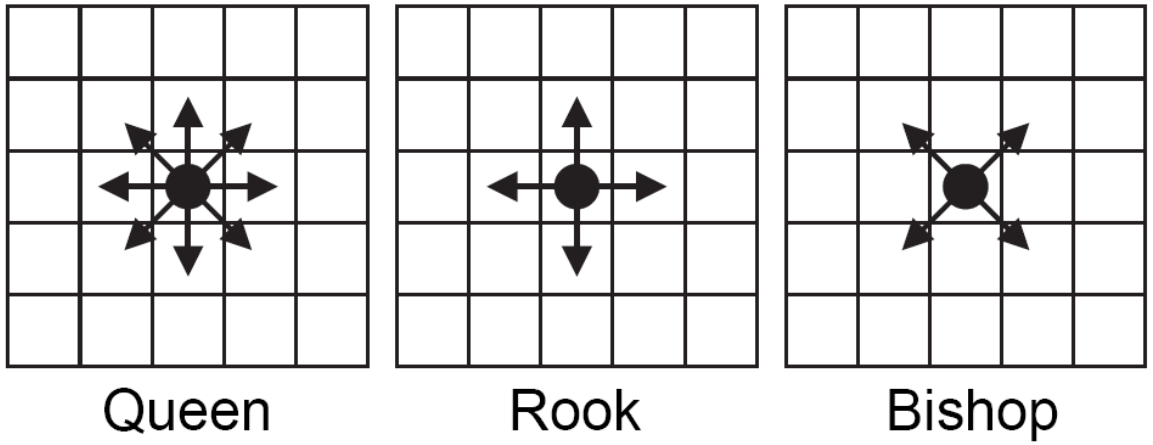
- A spatial weights matrix is a representation/quantification of the spatial structure/relationship of your data, ***guiding each target point to find its neighbors.***
  - ***Who are my neighbors?***
  - ***How important is each neighbor?***

- At the most basic level, two strategies for creating weights to quantify the relationships among data features:
  - **binary weight strategies**: a feature is either a neighbor (1) or it is not (0).
  - **Varied weights strategies**: weights to reflect the importance of each neighbor..

Referred to as the “conceptualization of spatial relationships” in Arc

Spatial weights matrix

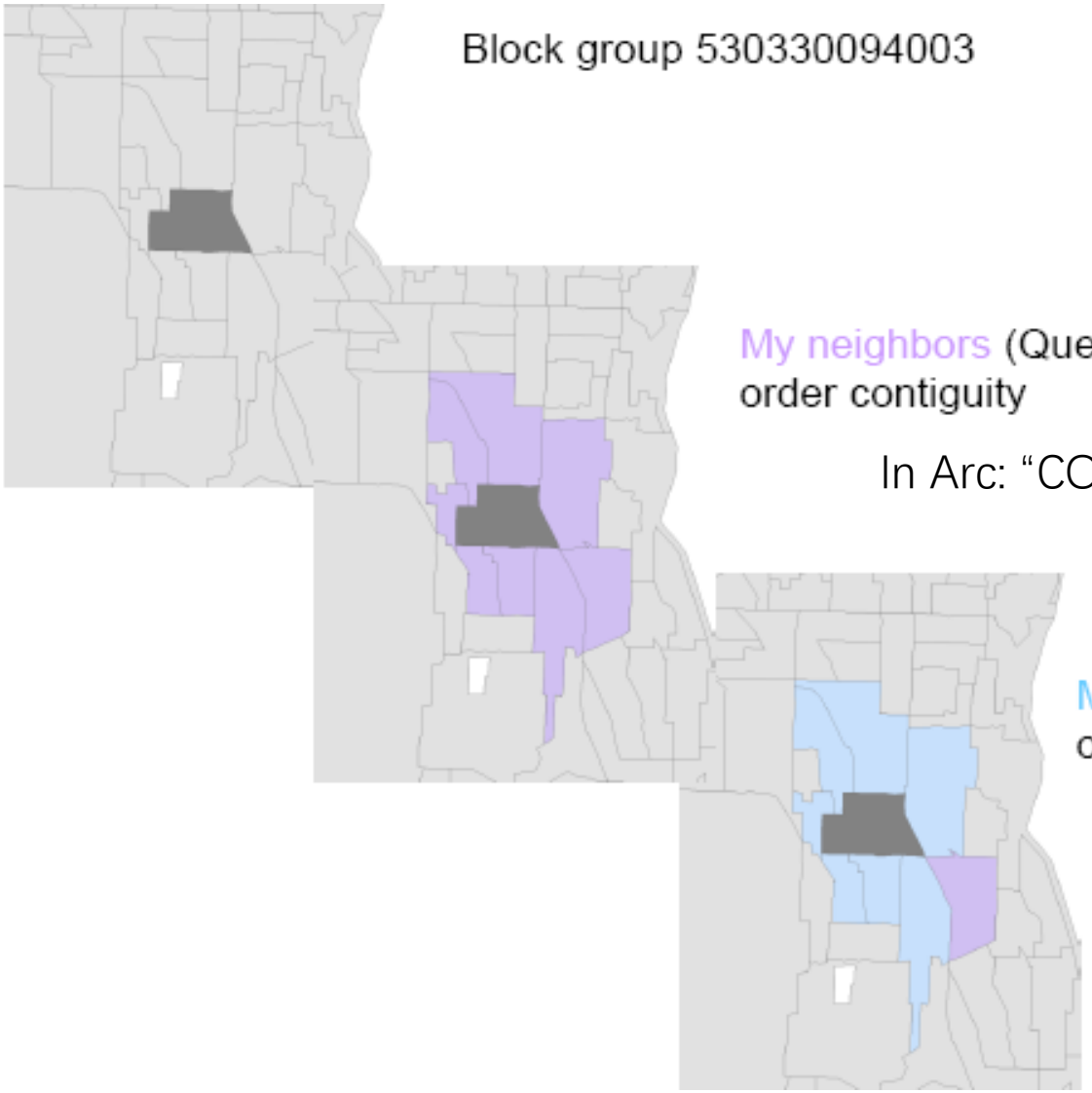
Binary: polygon contiguity



In Arc: Queen = “CONTIGUITY_EDGES_CORNERS”
Rook = “CONTIGUITY_EDGES_ONLY

- 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
- you can define a range of contiguity matrices: 1st nearest, 2nd nearest, 3rd nearest, etc.

Defining a neighborhood: Contiguity



Block group 530330094003

My neighbors (Queen's 1st order contiguity

In Arc: "CONTIGUITY_EDGES_CORNERS"

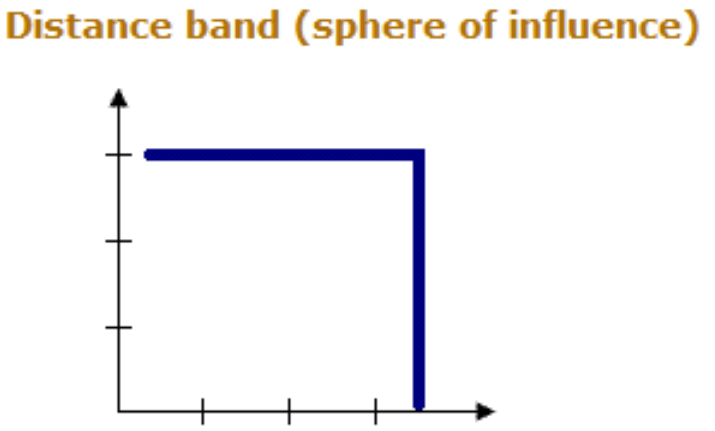
My neighbors (Rook's 1st order contiguity

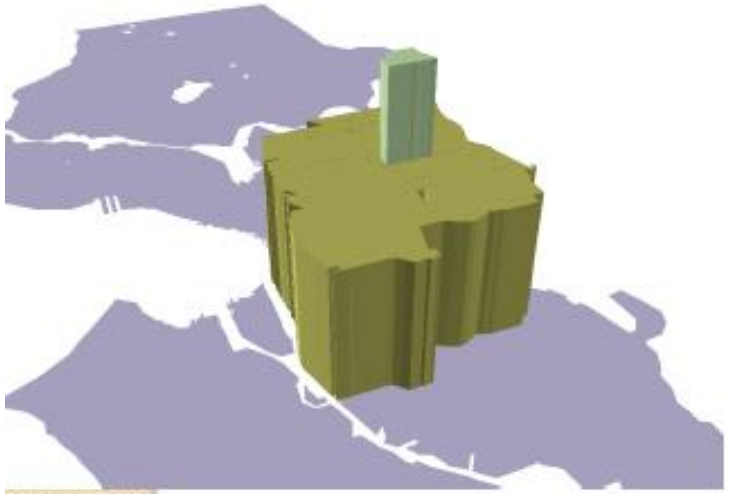
In Arc: "CONTIGUITY_EDGES_ONLY"

Spatial weights matrix

Binary: fixed distance band

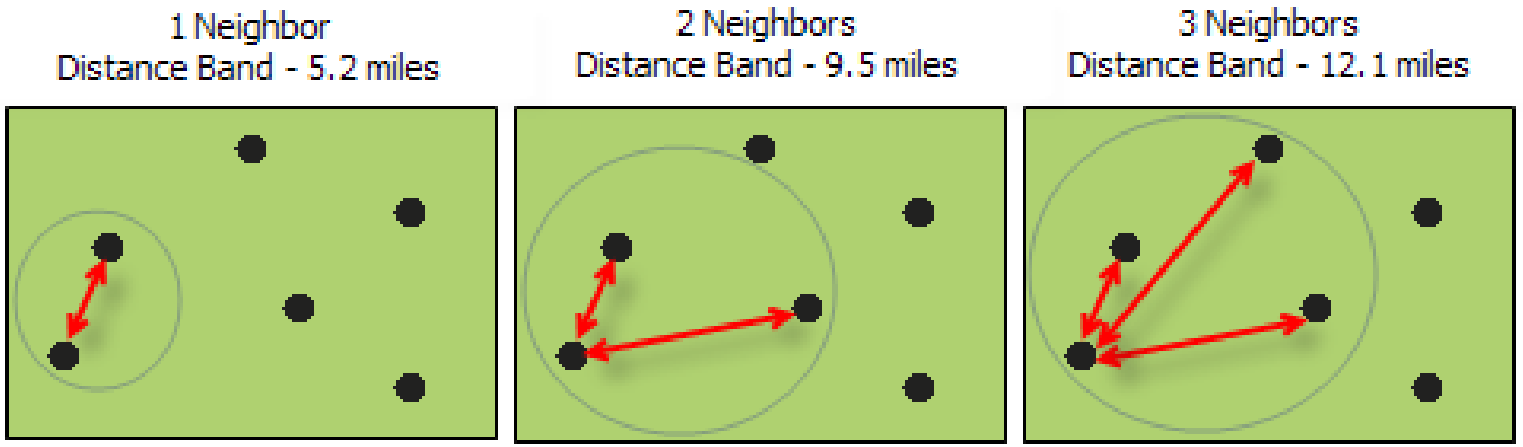
- neighboring features are defined within a certain threshold distance and are weighted equally.
- Features outside the specified distance don't influence calculations (their weight is zero).
- If you are studying commuting patterns and know that the average journey to work is 15 miles—you may want to use a 15-mile fixed distance for your analysis.





Spatial weights matrix
Binary: k-nearest neighbors

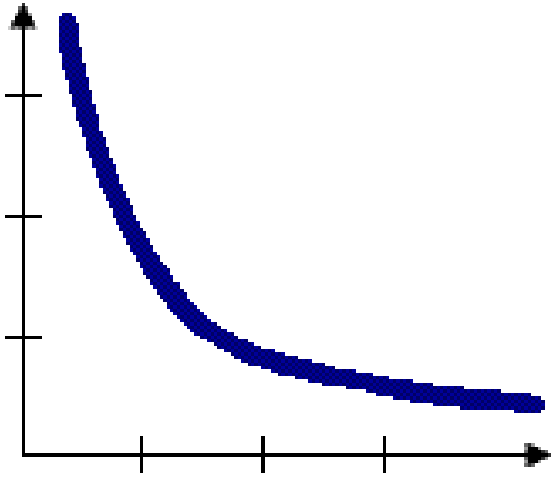
- Weights based on each feature is assessed within the spatial context of a fixed number of its closest neighbors.
  - In locations where feature density is high, the spatial context of the analysis will be smaller.
  - ensures there will be some neighbors for every target feature



Spatial weights matrix

Weighted strategies: Inverse distance

- All features impact/influence all other features, but the farther away something is, the smaller the impact it has
- Appropriate with continuous data or to model processes where the closer two features are in space, the more likely they are to interact/influence each other.



Neighbors are weighted based on the inverse of the log of their distance from my block group's centroid

Which to choose?

- you should select a conceptualization that best reflects how features actually interact with each other (giving thought, of course, to what it is you are trying to measure). Theory should inform this decision!

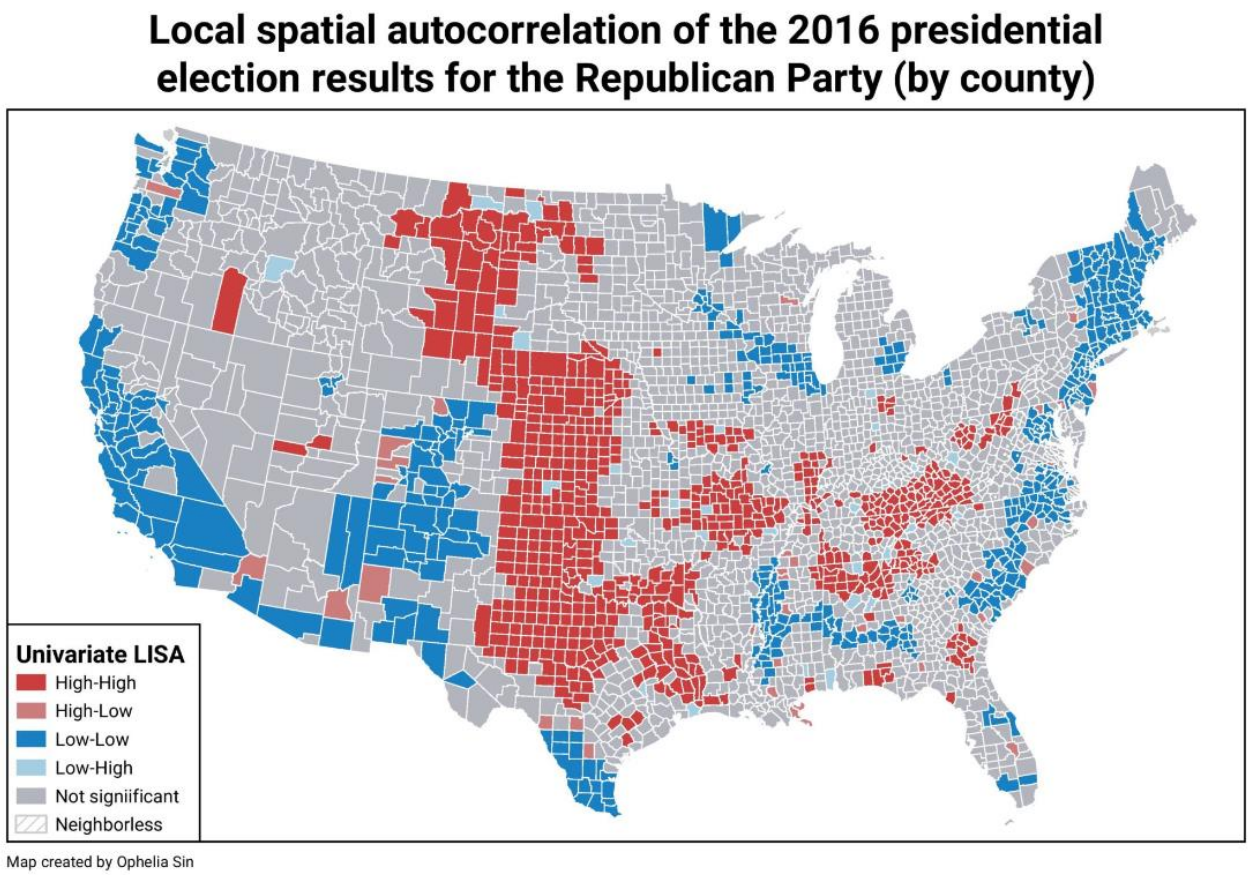
- If you are measuring clustering of a particular species of seed-propagating tree in a forest, for example, some form of inverse distance is probably most appropriate. If you are measuring social processes at the neighborhood level, polygon contiguity is a good choice

Local measures of clustering?

- We have a global sense of clustering from Moran's I, but lack information about its local variation
  - Where are the clusters? Which areas contribute the most to our global statistics? Which areas not at all?

- Local Indicator of spatial autocorrelation.
  - The Local Moran statistic (Anselin, 1995): as a way to identify local clusters and local spatial outliers.
  - Calculates Local Moran's I for each spatial unit and evaluating the statistical significance for each based on its similarity to its neighbor
  - Method*:* LISA (Local Indicators of Spatial Autocorrelation) or Cluster and outlier analysis in ArcGIS Pro

Local Indicator of spatial autocorrelation interpretation

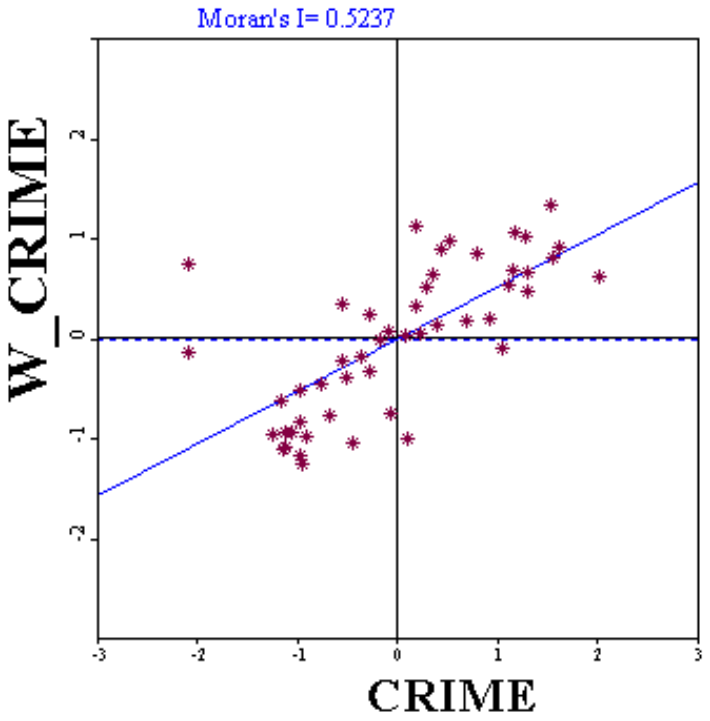


- High-High (red): Counties in red are Republican strongholds where high Republican voting is surrounded by other high Republican-voting counties, indicating clusters of strong Republican support.
- Low-Low (blue): Counties in blue show clusters of low Republican voting surrounded by other low-Republican counties, representing Democratic-leaning regions,
- High-Low (light red): These counties have high Republican voting but are surrounded by counties with lower Republican voting, indicating isolated Republican support in less supportive regions.
- Low-High (light blue): These counties show low Republican voting but are surrounded by higher Republican-voting counties, typically isolated Democratic-leaning areas within Republican regions.
- Not Significant (gray): Gray counties show no statistically significant clustering pattern, suggesting a mix or lack of strong spatial autocorrelation.

LISA 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: **X** and **lag-X**
- The slope of the regression line is Moran's I

Spatial outliers: **Low/High**
negative SA



Spatial Clusters: **High/High**
positive SA

Each quadrant corresponds to one of the four different types of spatial association

Spatial Clusters: **Low/Low**
positive SA

Spatial outliers: **High/Low**
negative SA

US Income Convergence Example

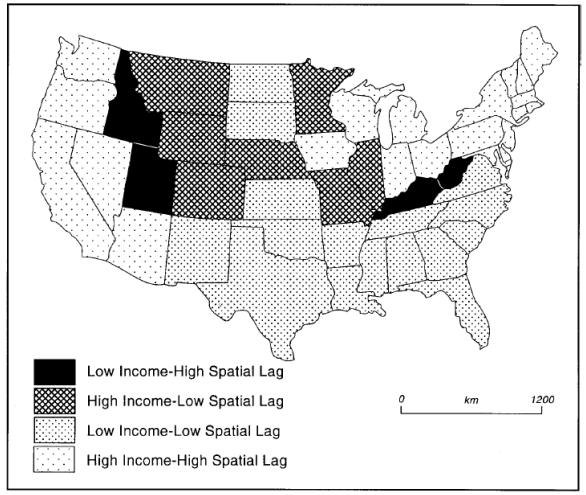


Fig. 4. Local Moran statistics per capita income, 1929

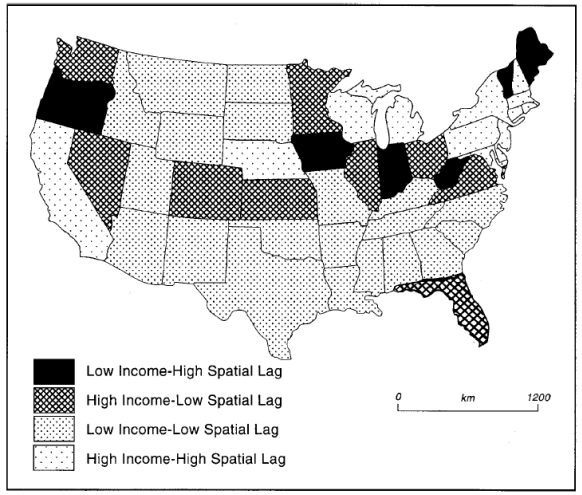
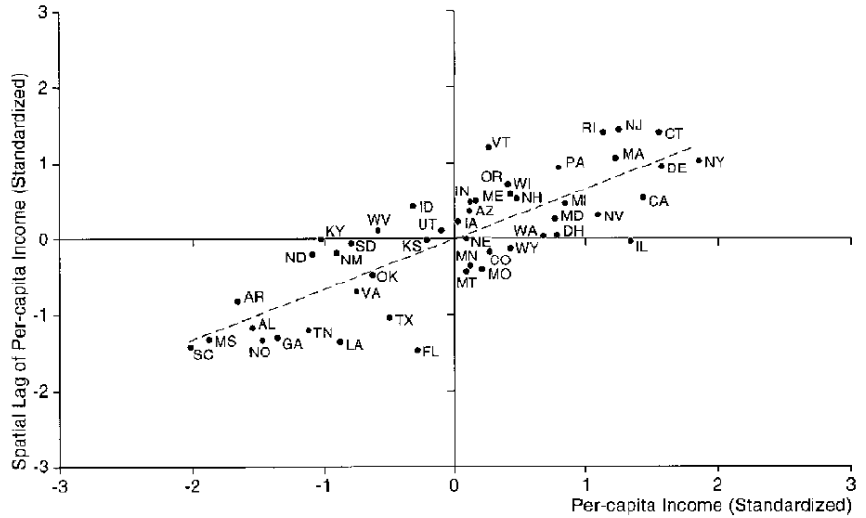
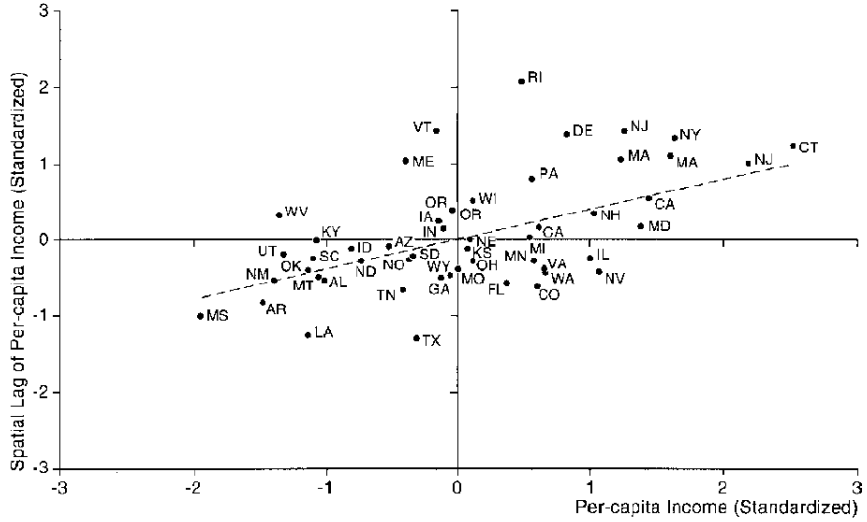


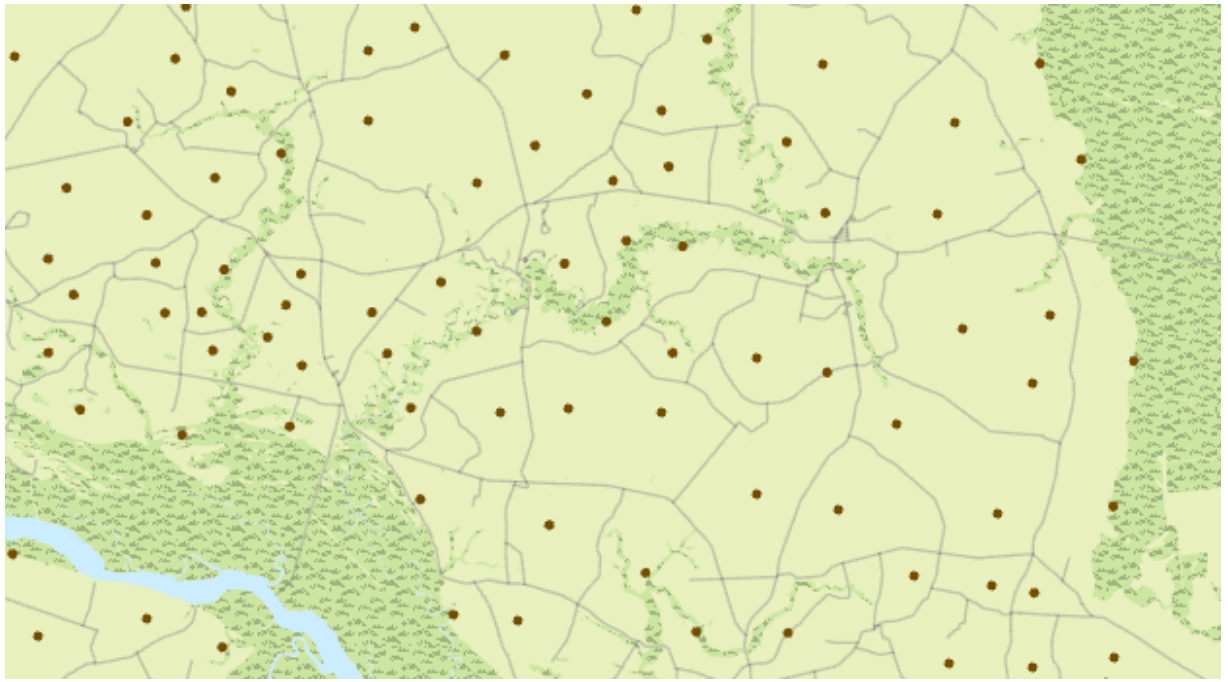
Fig. 5. Local Moran statistics per capita income, 1994





Source: Rey, S and B. Montouri, US Regional Economic Convergence: A Spatial Econometric Perspective, Regional Studies, 33 (2) 143-156

Standardization



- **Sampling design bias** refers to data sampling that may not reflect the underlying spatial distribution of the data that you want to analyze. This bias can be a result of ease of access, weather, political boundaries, or other factors that can affect data collection.
- **Standardization** adjusts the analysis to accommodate for different types of potential bias in your data.

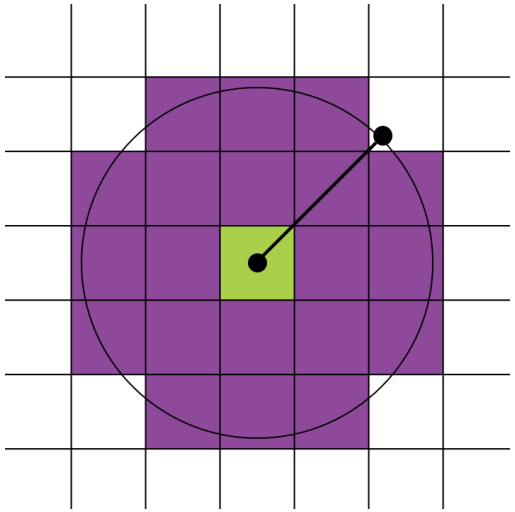
Row Standardization

- Row standardization is recommended whenever the distribution of your features is potentially biased due to features having different numbers of neighbors
- **Row standardization** will scale all weights so they are between 0 and 1, creating a *relative*, rather than *absolute*, weighting scheme. Each weight is divided by its row sum (the sum of the weights of all neighboring features).
- Common if you are working with polygon features representing administrative boundaries.

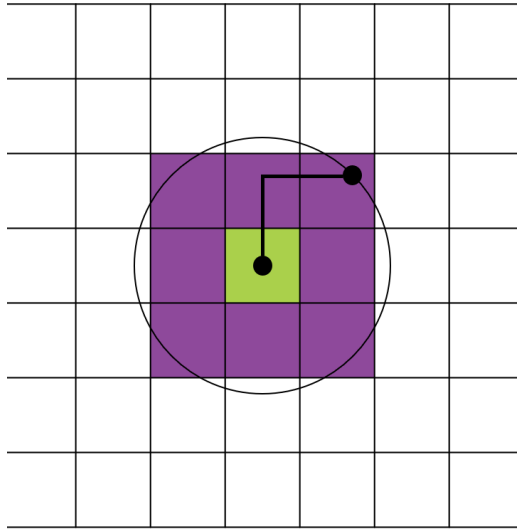
Distance Methods

Specifies how distances are calculated from each feature to neighboring features.

- EUCLIDEAN DISTANCE: The straight-line distance between two points (as the crow flies)



- MANHATTAN DISTANCE: The distance between two points measured along axes at right angles (city block); calculated by summing the (absolute) difference between the x- and y-coordinates



Why does clustering matter?

- Evidence of a spatial process at work
  - An end in itself, evidence of clustering can support a wide range of hypotheses about what is happening in your data
  - Can indicate the presence of problems in the data set for the purposes of statistical analysis (Spatial dependence is a violation of the independence of errors assumption of OLS regression analysis)

- Testing for spatial autocorrelation in your data is important. Unfortunately, identifying and quantifying the extent of spatial autocorrelation doesn't tell you what's *causing it*