Plan:

1. Explain the limitations of classical inference when working with geospatial data
2. Provide examples of each of these limitations to geospatial analyses

Spatial Statistics: The Why

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

Spatial Statistics

The statistical techniques we've discussed so far don't work well when considering spatial distributions...

Spatial Statistics

The statistical techniques we've discussed so far don't work well when considering spatial distributions…

…which means we have a chance to take a look at data and the relationship between the data in new and interesting ways
(distance, adjacency, interaction, and neighbor)
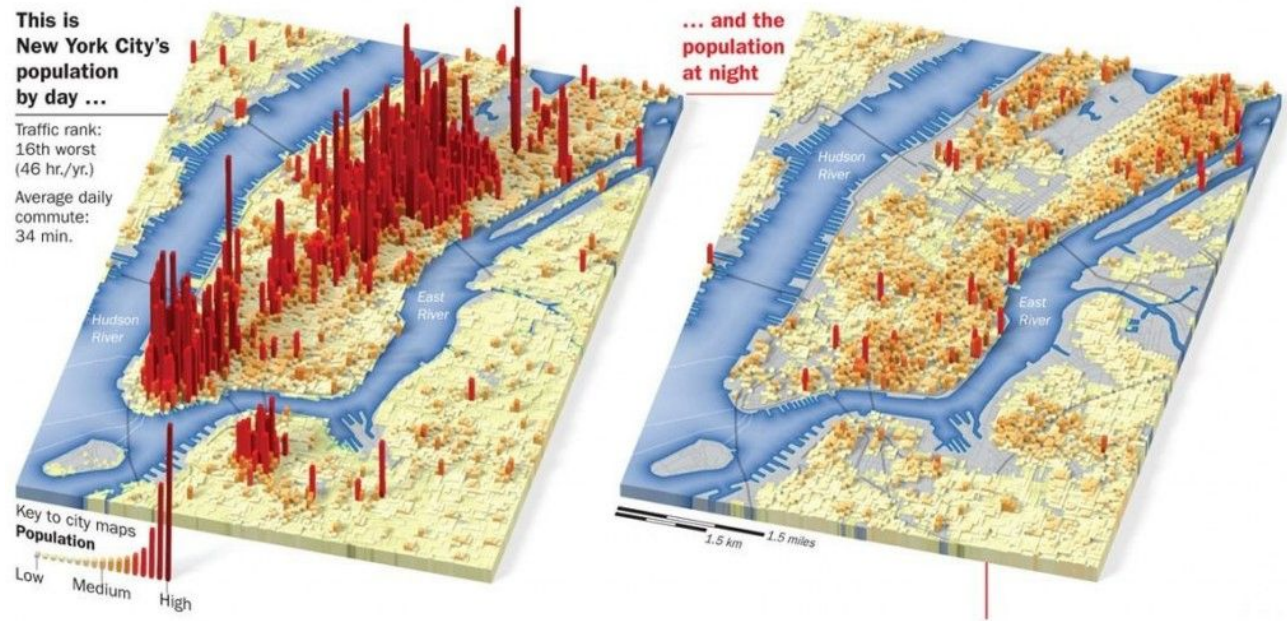
Spatial data violate conventional statistics:

Violations of conventional statistics:

- Spatial autocorrelation
- Modifiable areal unit problem (MAUP)
- Edge effects (Boundary problem)
- Ecology fallacy
- Nonuniformity of space

Spatial Autocorrelation

Data from locations near one another in space are more likely to be similar than data from locations remote from one another:

- Housing market
- Elevation change
- Temperature

This is New York City's population by day …

Traffic rank: 16th worst (46 hr./yr.)

Average daily commute: 34 min.

… and the population at night

Hudson River

East River

Hudson River

East River

Key to city maps
Population

Low   Medium   High

1.5 km   1.5 miles

Modifiable Areal Unit Problem (MAUP)

The aggregation units used are arbitrary with respect to the phenomena under investigation, yet the aggregation units used will affect statistics determined on the basis of data reported in this way.

If the spatial units in a particular study were specified differently, we might observe very different patterns and relationships.

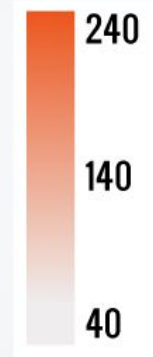# Modifiable Areal Unit Problem (MAUP)

modifiable area: Units are arbitrary defined and different organization of the units may create different analytical results.
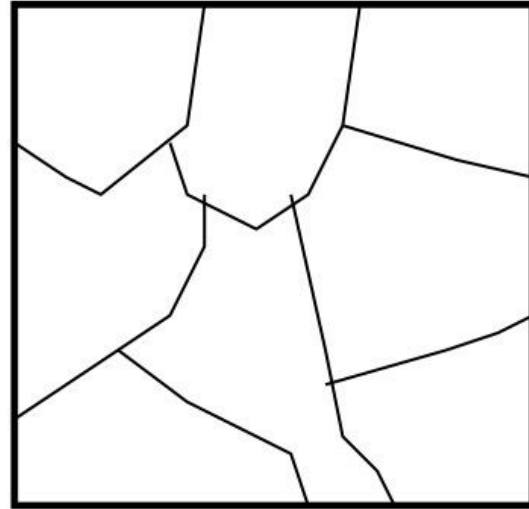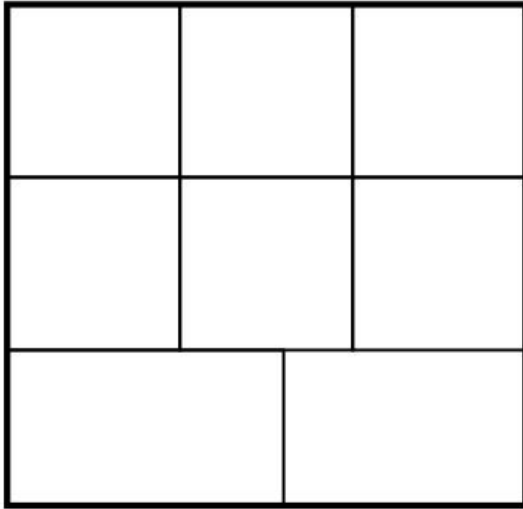
Multiple Sclerosis by Geography

CASE-CONTROL RATIO OF MS
A higher ratio indicates greater prevalence
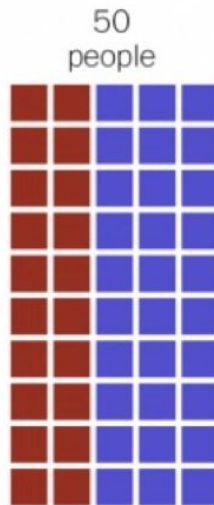
240
140
40

LEARN MORE AT WWW.INVW.ORG/MS

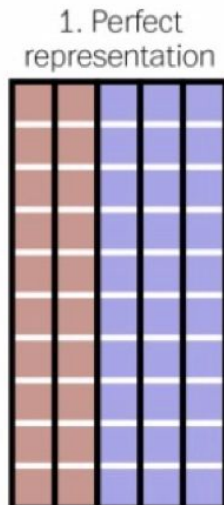PRODUCED BY: JASON ALCORN/INVESTIGATEWEST   SOURCE: BERETICH AND BERETICH (2009)

adapted from Brad Voytek

For example...gerrymandering

Gerrymandering, explained

Three different ways to divide 50 people into five districts

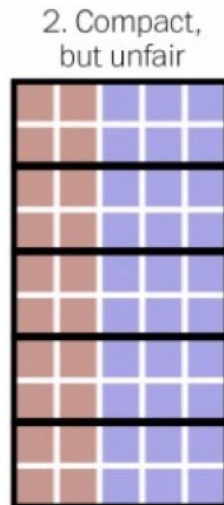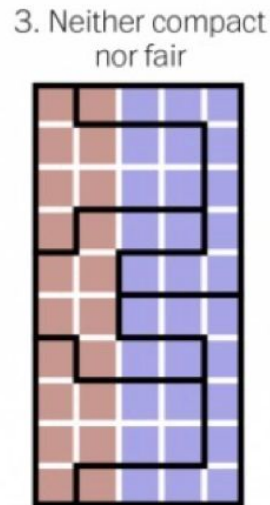50 people — 60% blue, 40% red

1. Perfect representation — 3 blue districts, 2 red districts — BLUE WINS

2. Compact, but unfair — 5 blue districts, 0 red districts — BLUE WINS
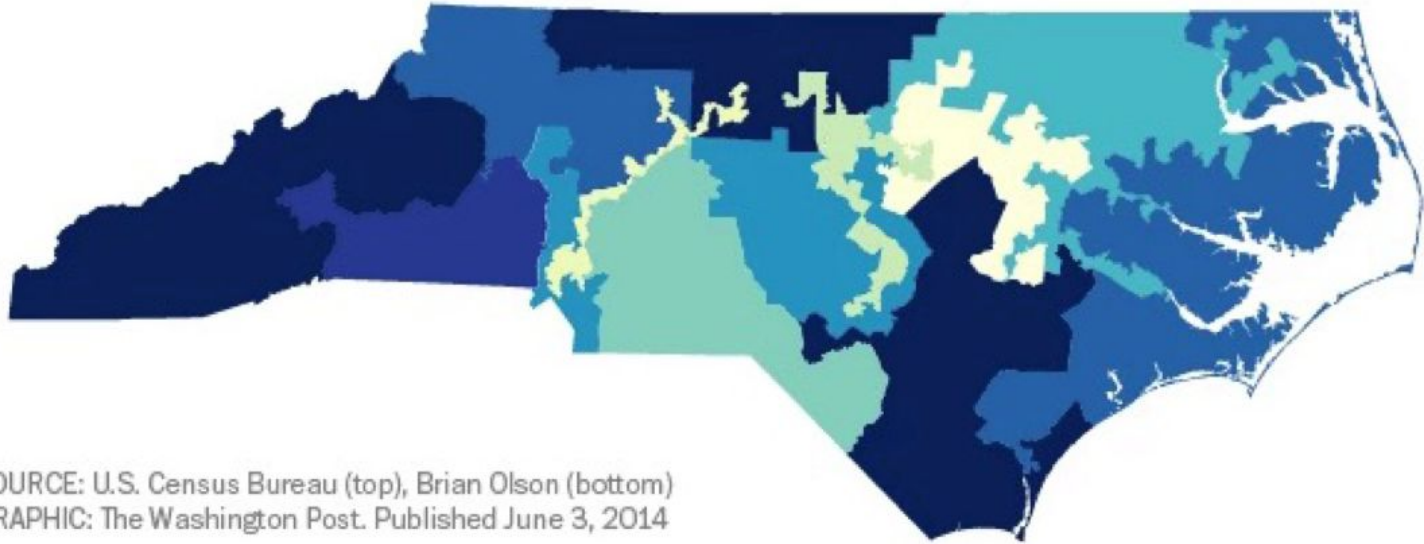
3. Neither compact nor fair — 2 blue districts, 3 red districts — RED WINS

For example...gerrymandering

North Carolina

CURRENT CONGRESSIONAL DISTRICTS

SOURCE: U.S. Census Bureau (top), Brian Olson (bottom)
GRAPHIC: The Washington Post. Published June 3, 2014

For example...gerrymandering

North Carolina

DISTRICTS REDRAWN TO OPTIMIZE COMPACTNESS

SOURCE: U.S. Census Bureau (top), Brian Olson (bottom)
GRAPHIC: The Washington Post. Published June 3, 2014

Modifiable Areal Unit Problem (MAUP)

Potential problems in almost every field that utilizes spatial data.

In the 2000 U.S. presidential election, Al Gore, with more of the population vote than George Bush, but failed to become president. (Same in 2016)

A different aggregation of U.S. counties into states could have produced a different outcome (switch just one northern Florida county to Georgia or Alabama would have produced a different outcome).

Modifiable Areal Unit Problem (MAUP)

Potential problems in almost every field that utilizes spatial data.

In the 2016 U.S. presidential election, Hillary Clinton, with more of the population vote than Donald Trump, but failed to become president.

A different aggregation of U.S. counties into states could have produced a different outcome.

Edge Effects (The Boundary Problem)

Analyzing A vs B ignores similarities
between the two based on their
shared boundary
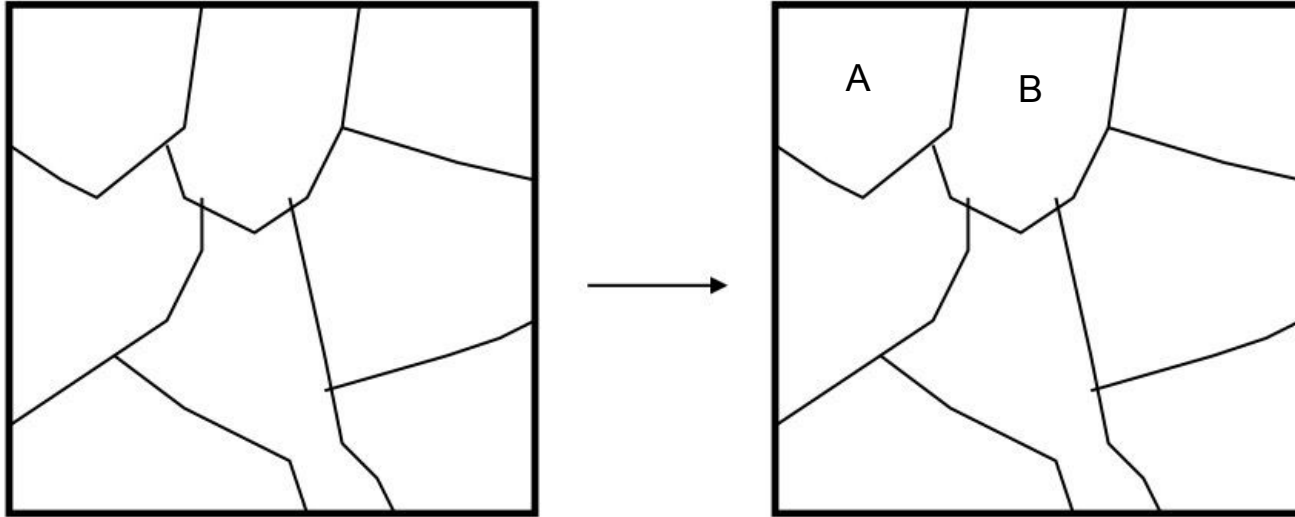
Ecological Fallacy

The Ecological Fallacy is a situation that can occur when a researcher or analyst makes an inference about an individual based on aggregate data for a group.
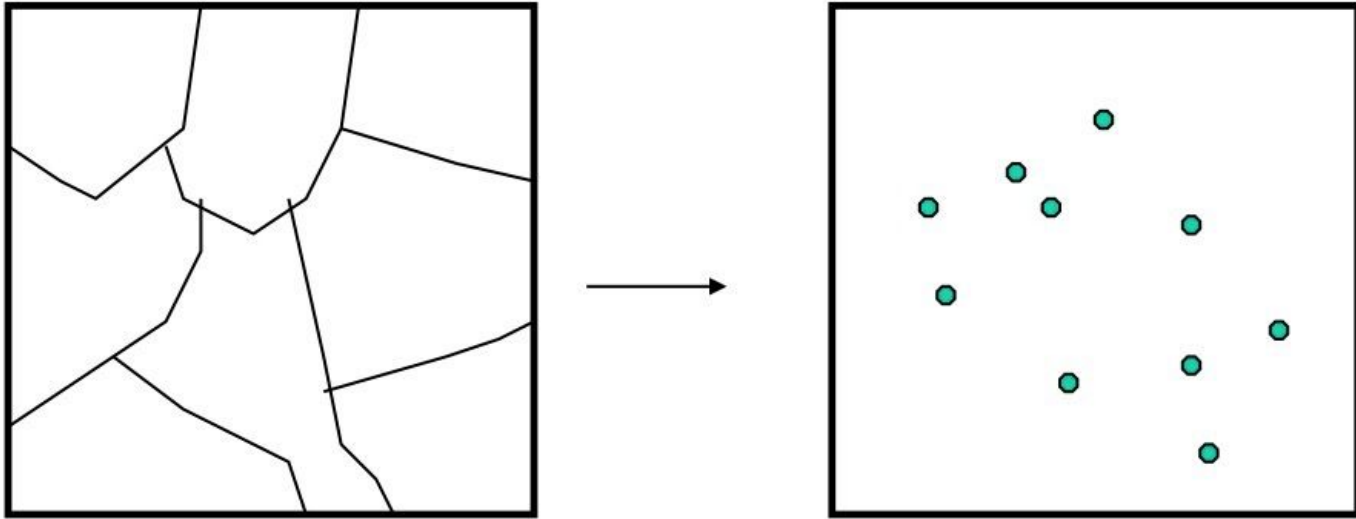
Ecological Fallacy

Example: we might observe a *strong relationship between income and crime at the county level*, with lower-income areas being associated with higher crime rate.

Conclusion:
- Lower-income persons are more likely to commit crime
- Lower-income areas are associated with higher crime rates
- Lower-income counties tend to experience higher crime rates

Ecological Fallacy

Issues:

Inferences drawn about associations between the characteristics of an aggregate population and the characteristics of sub-units within the population are wrong. That is: *results from aggregated data (e.g. counties) cannot be applied to individual people*

What should we do?
Be aware of the process of aggregating or disaggregating data may conceal the variations that are not visible at the larger aggregate level
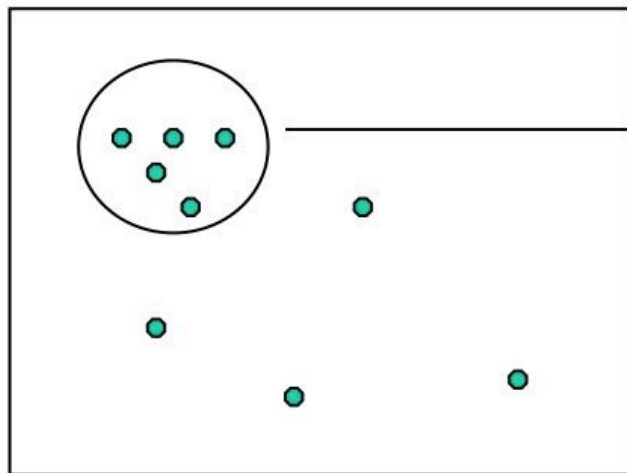
Nonuniformity

Crime locations

Area with high crime rates?
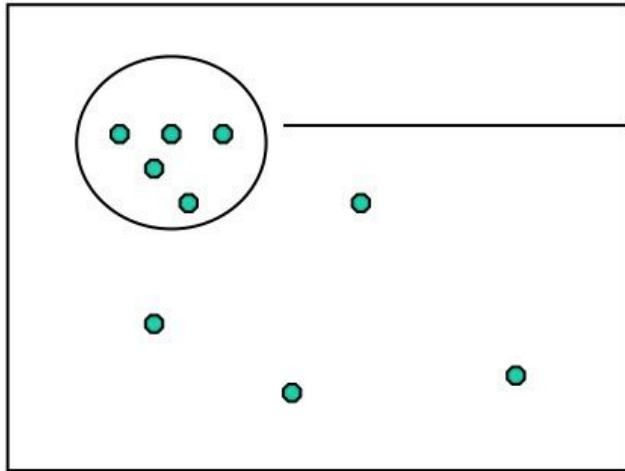
Nonuniformity

Area with high crime rates?

Crime locations

Conclusion: Bank robberies are clustered ....but only because banks are clustered!