Exploratory Spatial Data Analysis

CSDE GIS Workshop

Matt Dunbar, PhD.
and Chris Fowler, PhD.

October 12th, 2011

Center for Studies in
Demography and Ecology

Outline

- Spatial and aspatial data distributions
  - Means, outliers, correlations
- Finding a spatial structure
  - Neighborhoods and scale
- Quantifying a spatial structure
  - Global and local measures of clustering

Why we explore our data

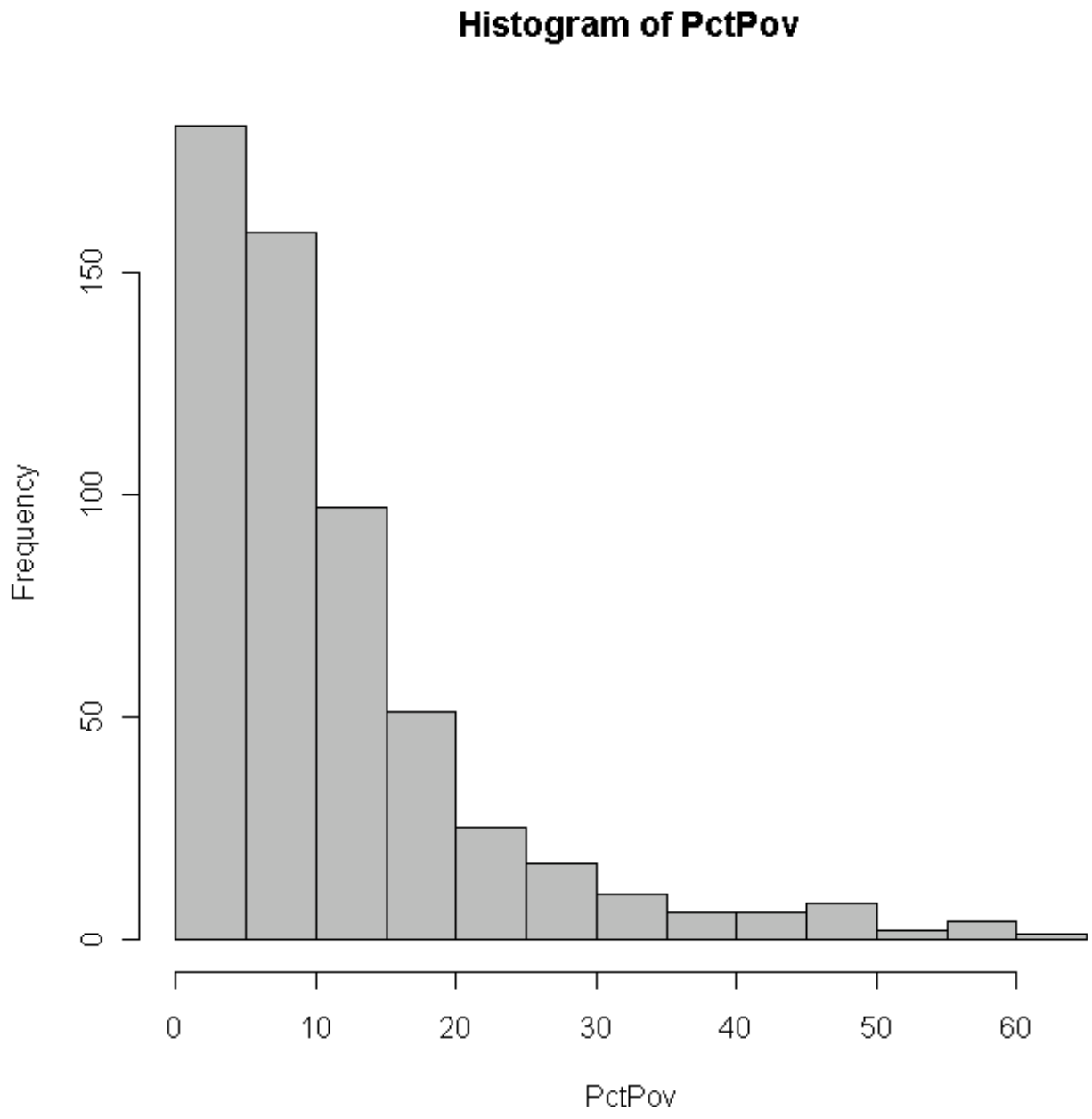
- General understanding
- Hypothesis formation
- Suitability for inclusion in statistical analysis
- Refinement of all of the above

Poverty, Race, and the Location of Financial Institutions in Seattle

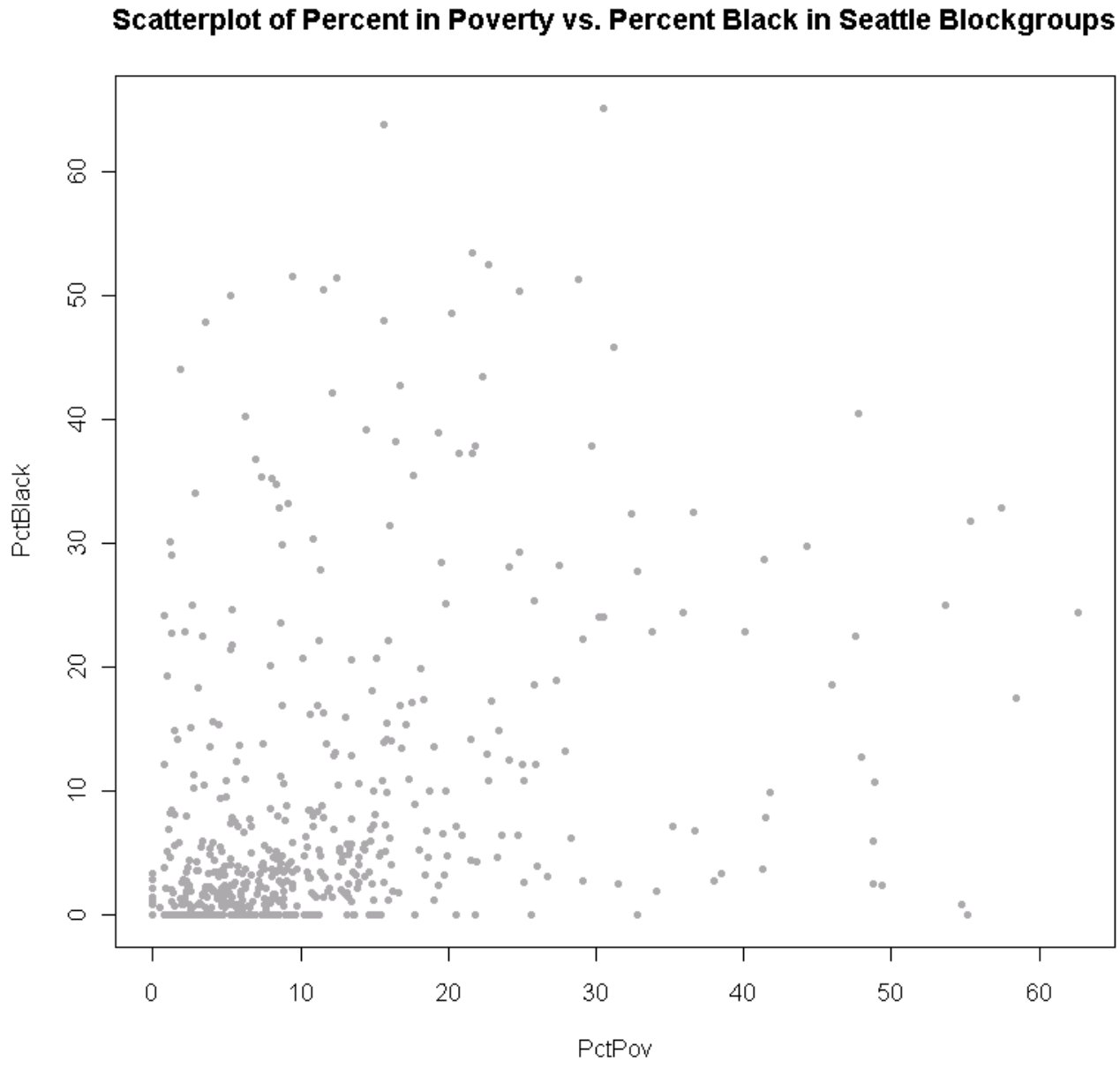
- What kinds of relationships exist among poverty status, race, and neighborhood access to financial services (banks as well as payday lending and pawnshops)?

- Socio-economic data from the 2000 Census at the block group level
- Geocoded, point data on financial institutions from June 2009.

Understanding data distributions



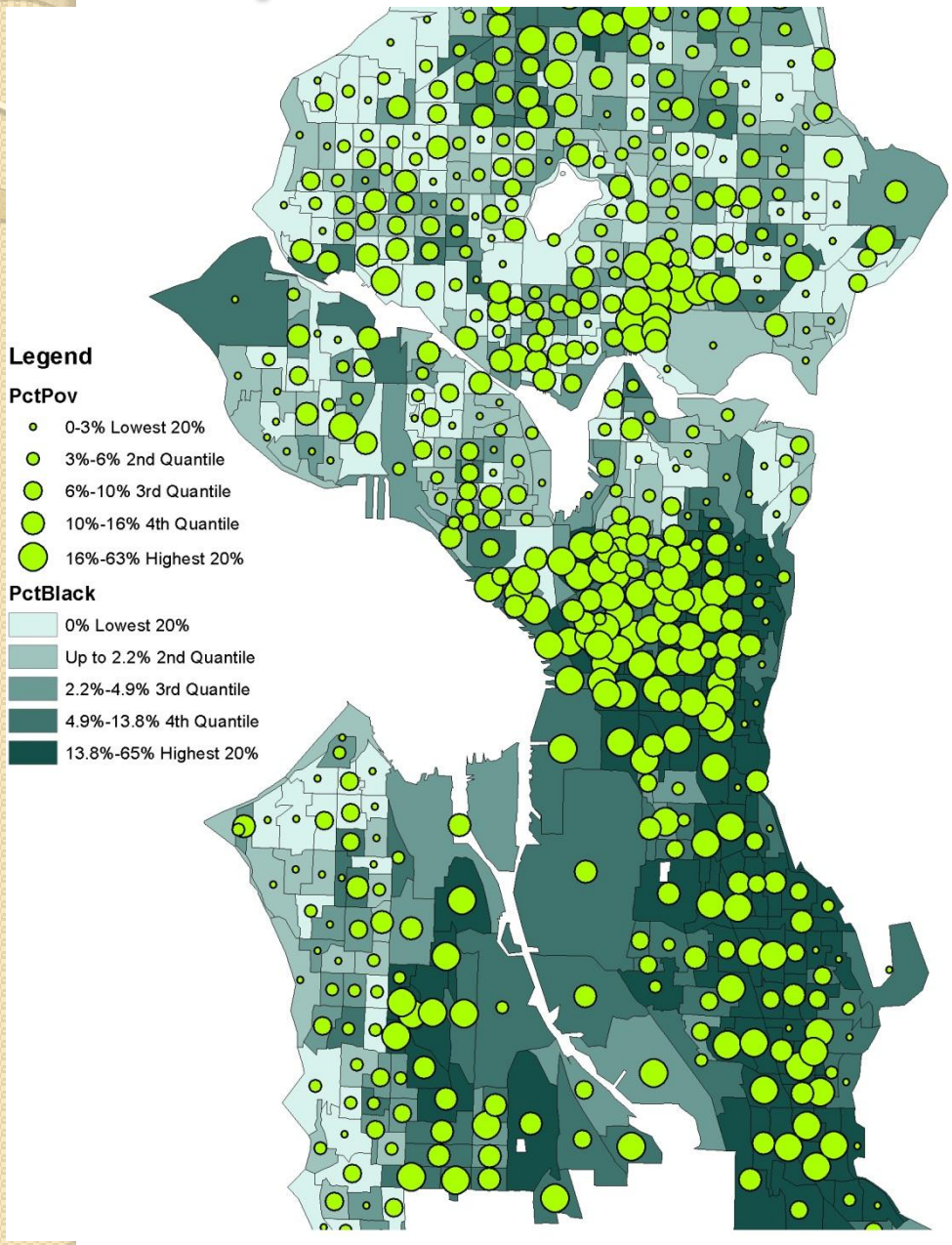
Locating outliers



Spatial data distributions

But an aspatial distribution may hide spatial patterns

Spatial Correlations



Also interested in determining relationships (b/w variables) within a given location.

Lab Goals:

- Explore data distributions in ArcGIS
- Generate geographic and population mean centers
- Generate spatial standard deviations and estimates of direction and compactness

Summary and Discussion

- We need to understand our data in both spatial and aspatial terms
- Outliers, means, medians, compactness, direction
  - all have spatial representations

- What can we say about our data at this point?
- What questions have been raised by our examination?

Outline

- Spatial and aspatial data distributions
  - Means, outliers, correlations
- Finding a spatial structure
  - Neighborhoods and Scale
- Quantifying a spatial structure
  - Global and local measures of clustering

What is "spatial structure"

- 1) A claim about the role spatial processes play in our analysis
  - Proximity, distance, connectivity, interconnectedness, movement, borders, boundaries, etc
- 2) A claim about the scale at which these processes have meaning

- Our claims about these two elements will determine what it is where we expect to see it
  - Clustering, dispersion, regionality, discontinuity, trending

Spatial processes

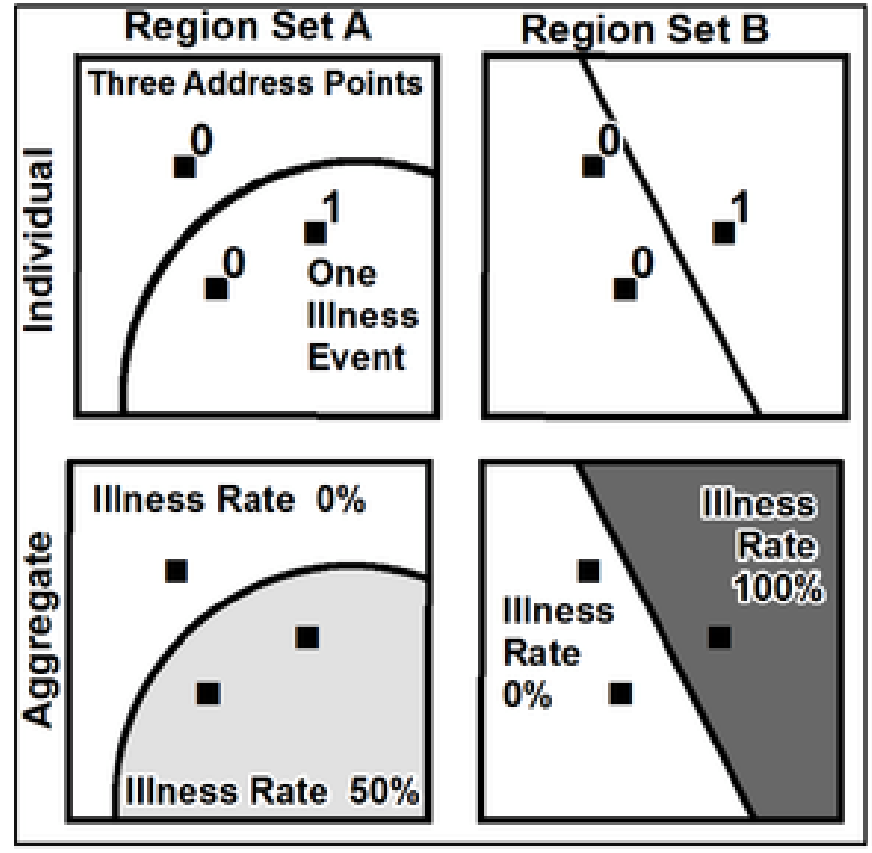
- Patterns of influence/interdependence
  - "spatial dependence"
  - Ex. Stepped up crime prevention in my neighborhood alters crime levels in nearby neighborhoods as well
- Broad scale patterns of similarity based on history, climate, etc…
  - "spatial heterogeneity"
  - Ex. People living in northern areas more likely to play hockey than those in southern regions
- Diffusion
  - Technically spatial dependence as above, but with potentially different observed outcomes
  - Ex. Language/dialect drift

Spatial scale

- Individuals share information and learn from those they associate with
- Neighborhoods get a reputation that is self-fulfilling
- Regions may operate under a uniform regulatory structure
- Countries may sharply modify costs of doing business across a border
- World regions face similar climate conditions

A note on scale

- MAUP, The Modifiable Areal Unit Problem
  - The scale of our analysis may condition our results



Unattributed image. Downloaded from Wikipedia 1/24/11

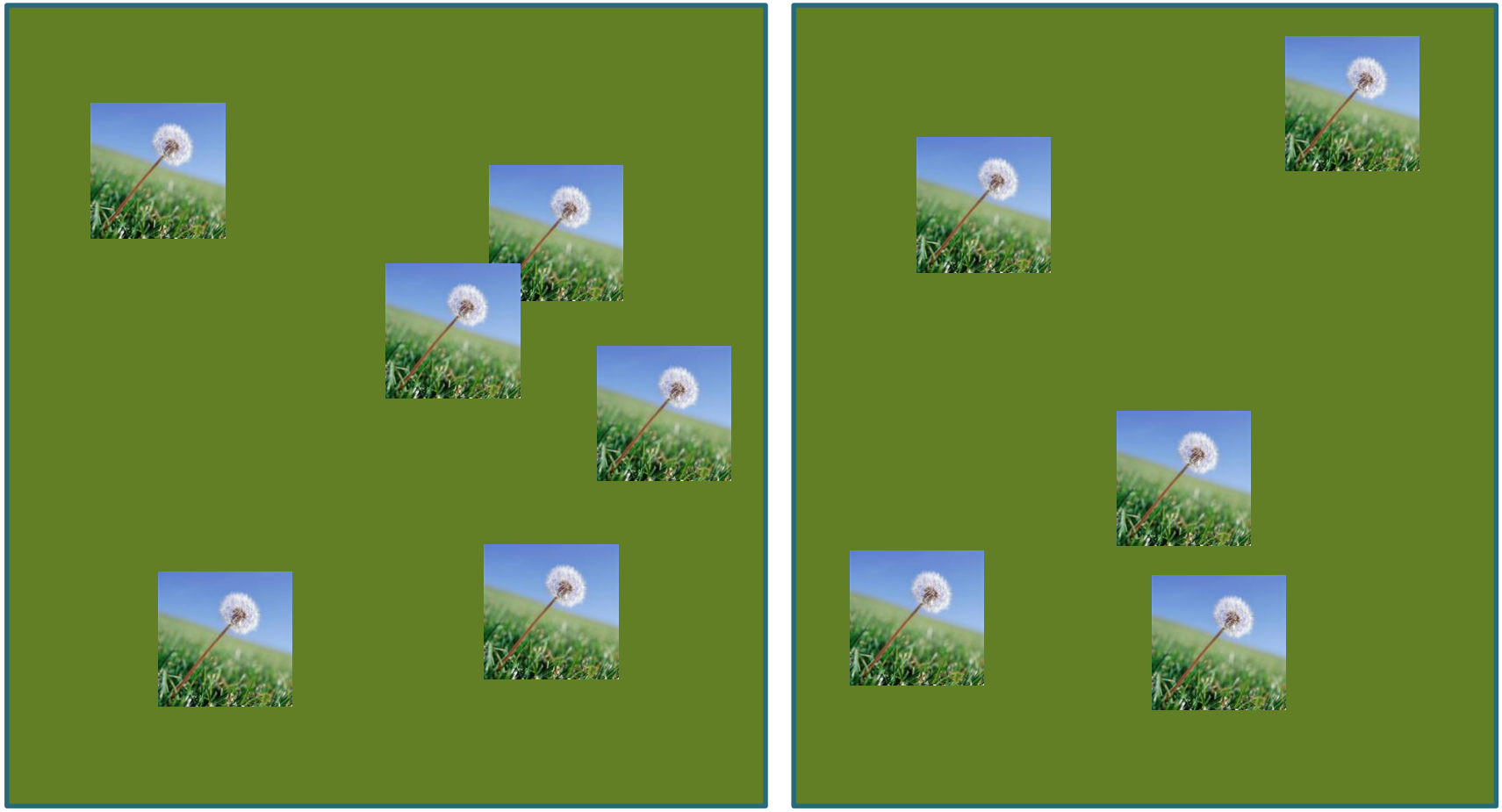
Outcomes

- Discontinuity
  - Spatial outliers, layering with borders
- Clustering and dispersion
  - Lots of tests measuring clustering, hot spots, spatial autocorrelation, etc…
- Regionality
  - Local measures of clustering
- Trending
  - Covered in spatial regression workshop

Finding a spatial structure

- Tobler's First Law "Everything is related to everything else, but near things are more related than distant things" (1970)
- But how related and why?
- How near is "near?"

The wrong structure may artificially generate evidence for a process that is not really meaningful or obscure evidence of a process that is.



Finding a spatial structure

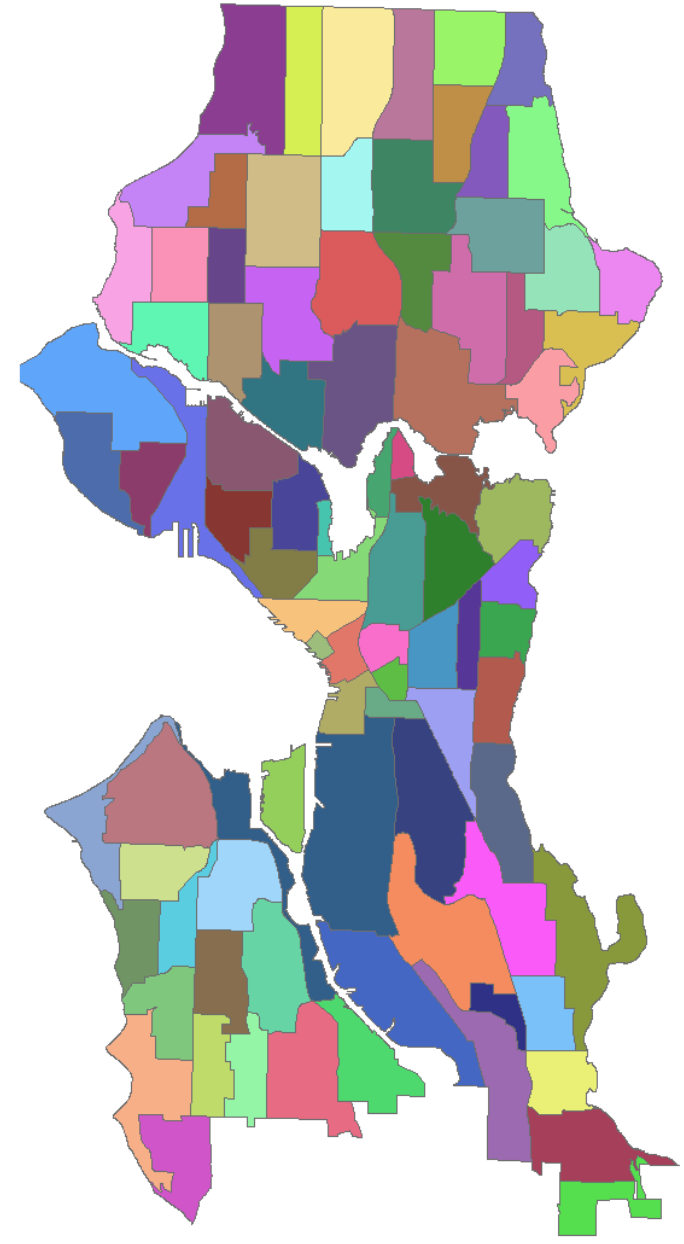
- Define it empirically
- Define it based on theory
- Sad but true…define it based on limitations of the data

- In practice….a little bit of each

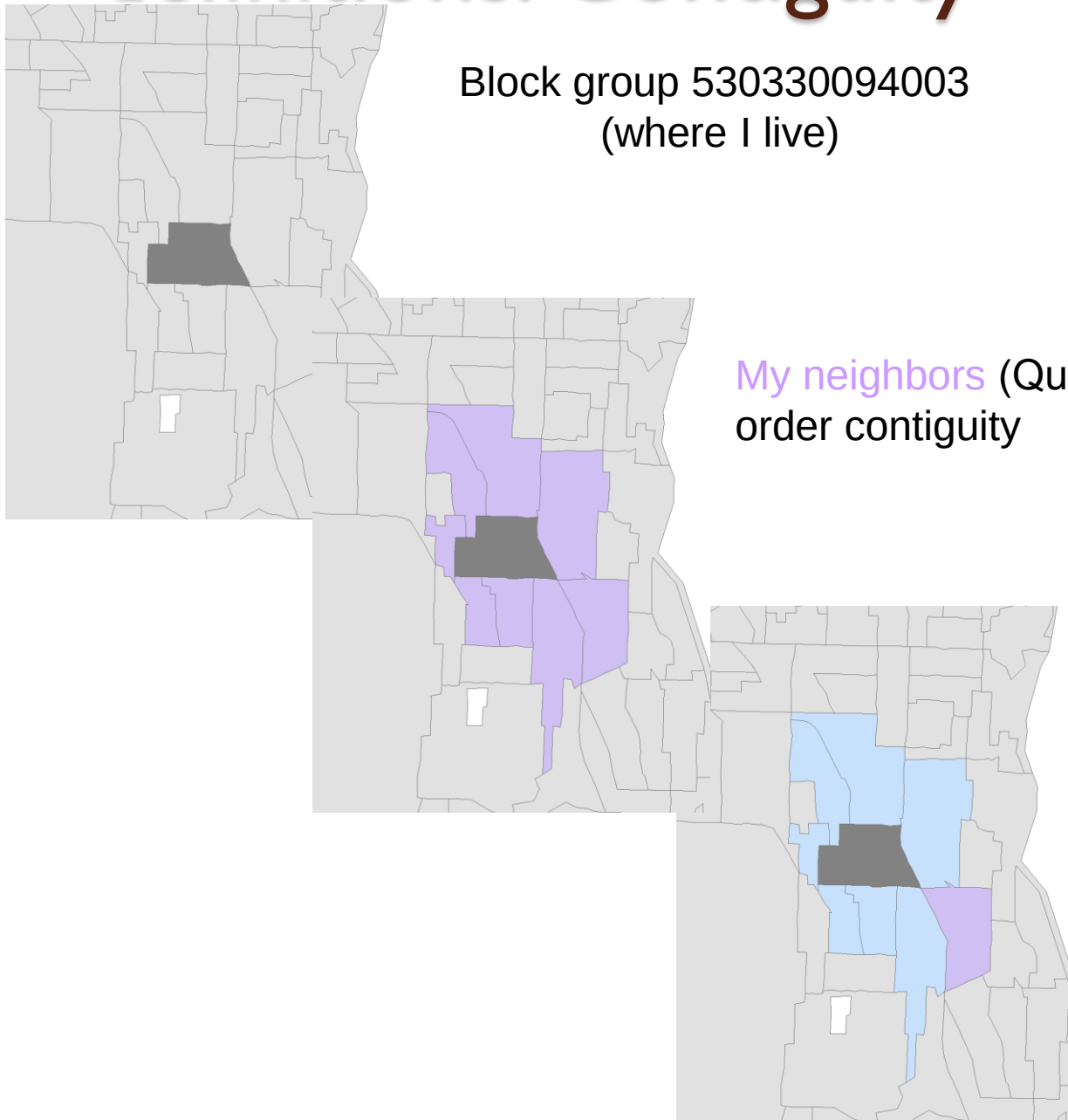
- Whatever you choose, your choice will represent a very powerful assumption going forward.

Poverty and access to financial services in Seattle

- Aspatial version: Test for a correlation between presence of fringe services and high percent minority population in a block group.
- But might we care about space?
- If we do care about space, at what scale is it likely to be meaningful?



The nuts and bolts of neighborhood definitions: Contiguity



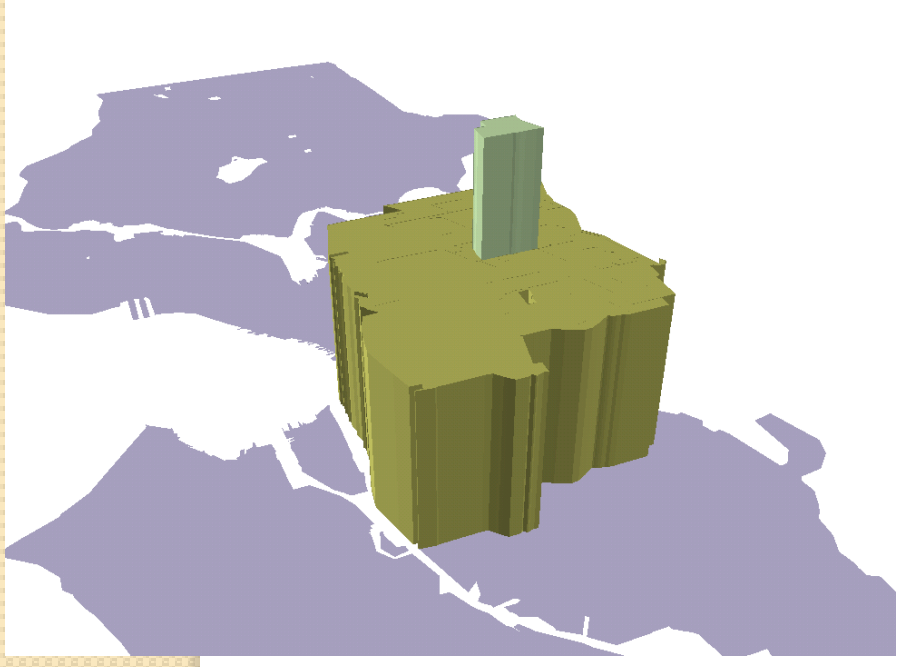
Block group 530330094003
(where I live)

My neighbors (Queen's 1st order contiguity

My neighbors (Rook's 1st order contiguity

Nuts and bolts: Distance and Inverse Distance

DISTANCE



A neighbor is any location whose centroid is within 1 mile of my block group's centroid

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

INVERSE DISTANCE

Things to think about when defining a spatial structure

- Islands
- Differences in observation size
- Differences in observation shape
- Border effects
- Study area shape

Which Weight Matrix Should I Use?

- Many other configurations are possible
  - ArcGIS allows for editing by hand to further customize these choices.
- Decision needs to be based on theory
  - What kinds of relationships do you think are important in your data.
  - Possible to employ more than one, but beware of excessive complexity.

Outline

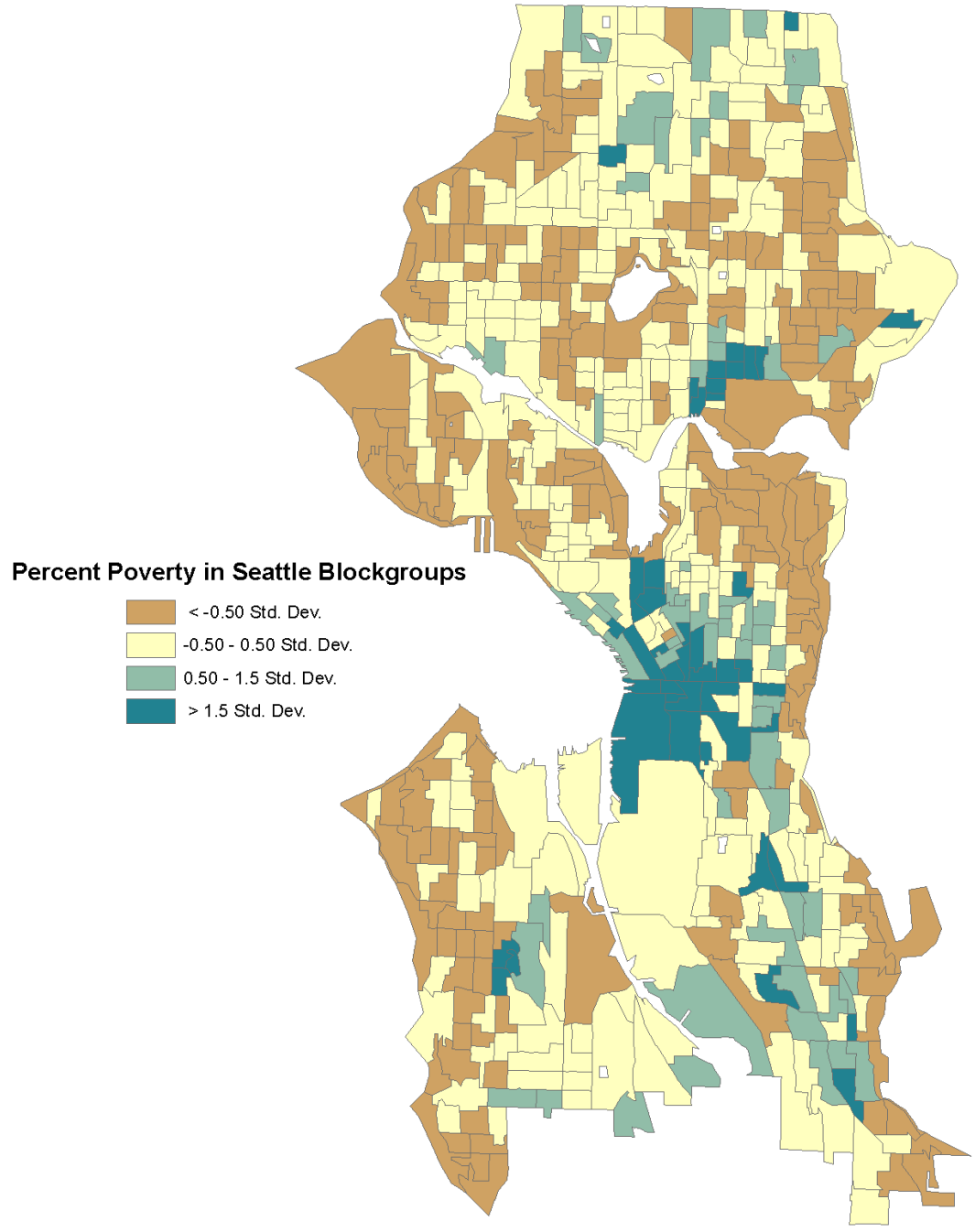
- Spatial and aspatial data distributions
  - Means, outliers, correlations
- Finding a spatial structure
  - Neighborhoods and Scale
- Quantifying a spatial structure
  - Global and local measures of clustering

Global vs. Local Cluster Measures

- Global
  - Do the data *as a whole* exhibit a spatial pattern or is the distribution random?
  - Moran's I
- Local
  - Identifies the specific units that are correlated (positively or negatively) with their neighbors
  - LISA (Local Indicators of Spatial Autocorrelation)

If Seattle's Poor were distributed at random…

would we expect the map to look like this?



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
  - Come back for the next two workshops to learn more!

Moran's *I* in depth

$$I = \left(\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}}\right)\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Covariance Term

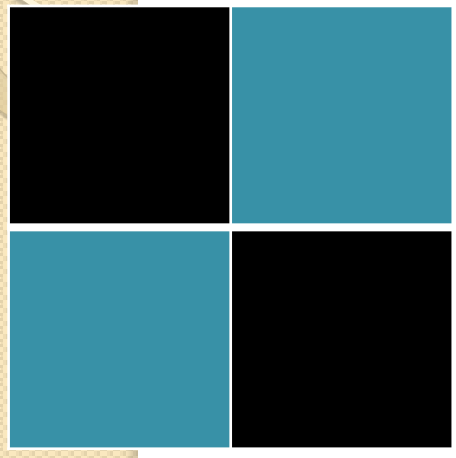
Normalization Term

- The top portion of the equation multiplies our weights matrix by the covariance.
- Zeroes in the weights matrix (non-neighbors) will cancel out any effects but those of the relevant observations
- The bottom portion controls for the overall variance in the data set

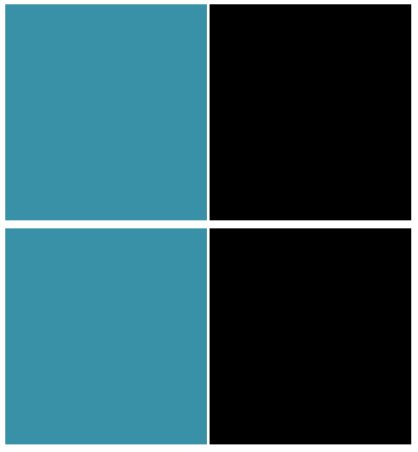
Interpreting Moran's *I*

- To what degree are observed values similar to the average of neighboring observed values—scaled to account for variance in data as a whole
- Index: Varies from -1 (evenly distributed) to 1 (clustered)
- Expected Index: establishes the effective "0" value
- P-value: The probability that this configuration could be the result of a **random spatial process**

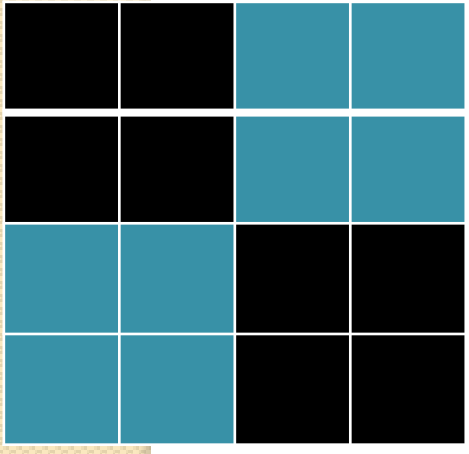
Results will vary by scale and weights matrix



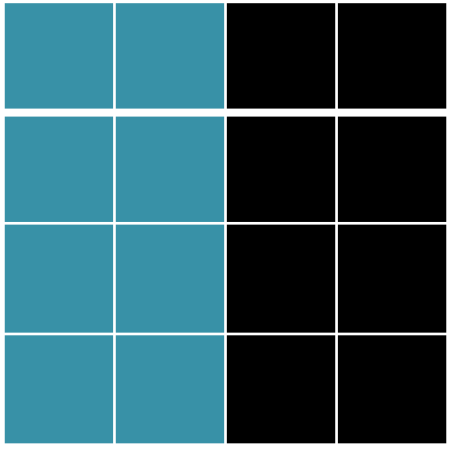
Rook: -1
Queen: -.33



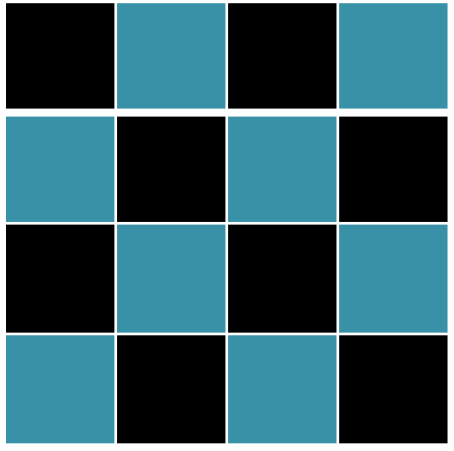
Rook: 0
Queen: -.33



Rook: .33
Queen:.24



Rook: .67
Queen:.52



Rook: -1.0
Queen:-.14

Global vs. Local

- Global
  - Do the data *as a whole* exhibit a spatial pattern or is the distribution random?
  - Moran's I
- Local
  - Identifies the specific units that are correlated (positively or negatively) with their neighbors
  - LISA (Local Indicators of Spatial Autocorrelation)

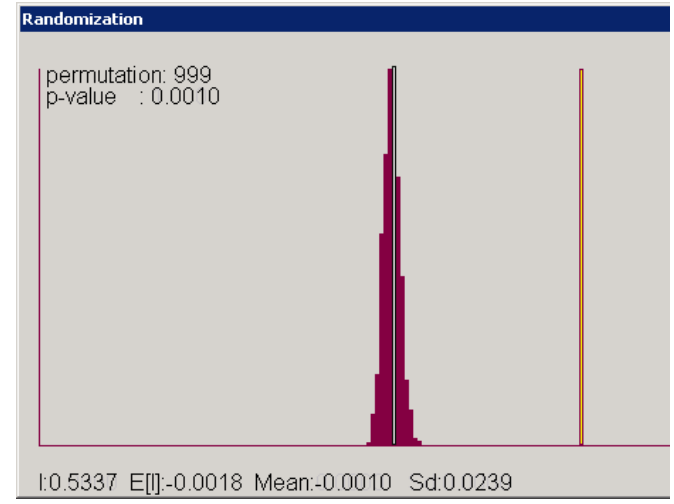
Local Measures of Clustering

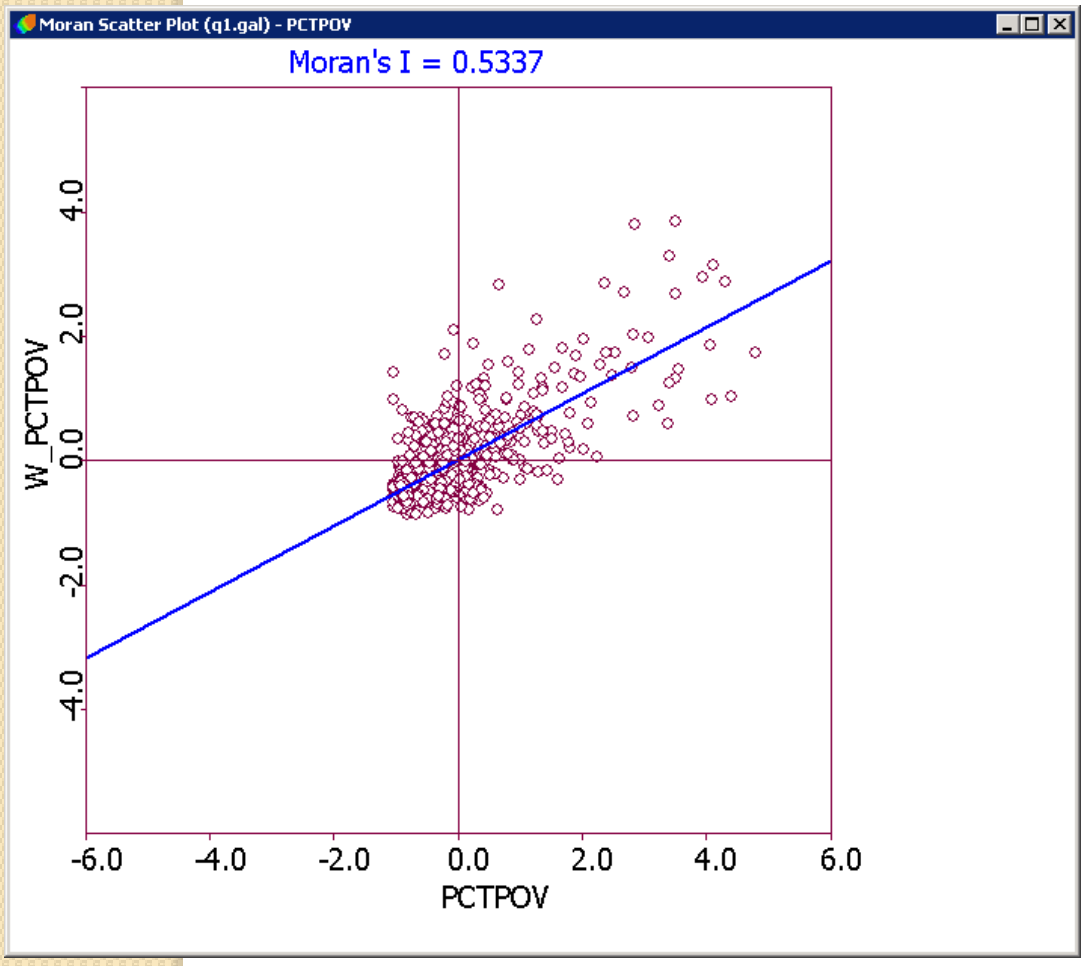
- We have a global sense of clustering from Moran's *I*, but lack information about its local variation.
- Where are the clusters?
- Which areas contribute the most to our global statistic? Which areas not at all?

A brief departure from ArcGIS

- Open GeoDa
  - http://geodacenter.asu.edu/
  - Free, small, excellent for exploring spatial data
- For our purposes it is also a good way of explaining the different components of LISA

LISA





Values at Location	Values of Neighboring Locations: Low	Values of Neighboring Locations: High
High	High/Low	High/High
Low	Low/Low	Low/High

Lab: Global and Local Moran's *I*

- Look at global measures of clustering
- Conduct LISA analysis
- Decompose measure to see where clustering is strongest and where counterfactuals are.