Point Patterns and Raster Surfaces

CSDE Statistics Workshop

Matt Dunbar and Chris Fowler

April 26th 2011

University of Washington

Center for Studies in Demography and Ecology

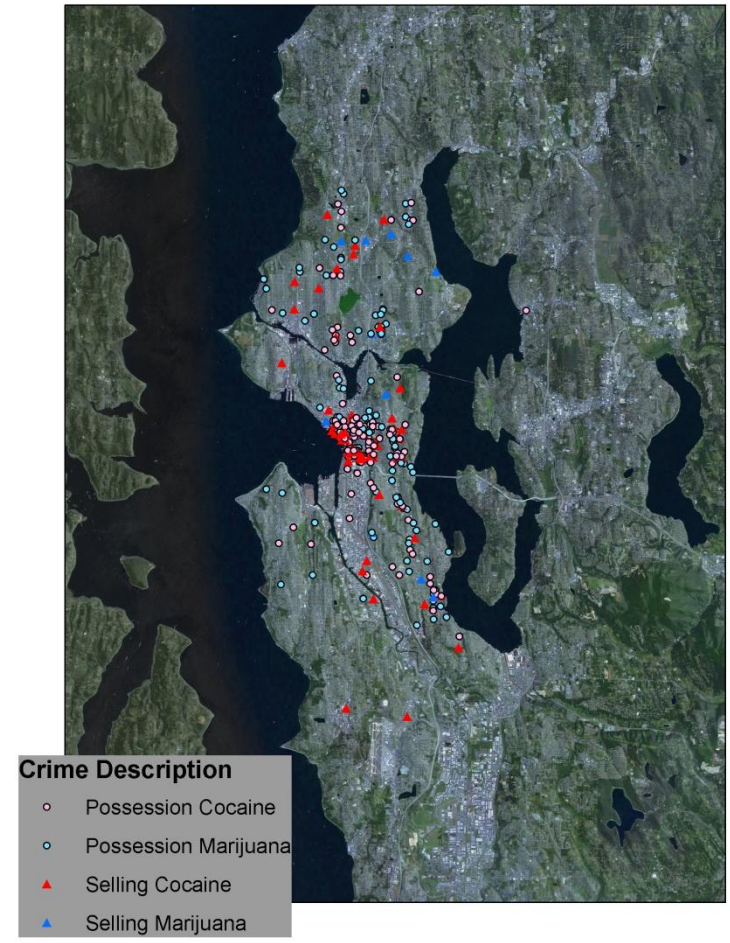
Outline for the Session

- Data Types:
  - Point Events, Point Samples, Raster
- Event Data
  - Density Surfaces, Point Pattern Analysis
- Point Sample Data
  - Interpolation
  - Geostatistics
- Going Further
  - Raster Map Algebra

DATA TYPES
-EVENT AND SAMPLE POINTS
-RASTER VS. VECTOR

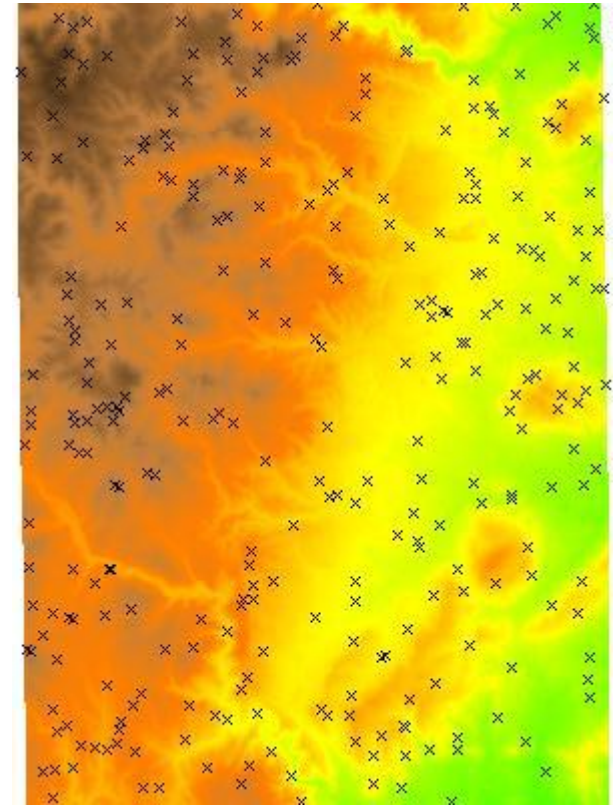
Data Types: Event Data

- x,y coordinates
- "marked" data contain some additional information.
  - Type of event, date of occurrence, etc…
- Understood as a complete set
  - Omissions can bias subsequent analysis



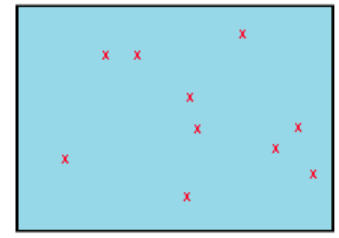
Data Types: Point Samples

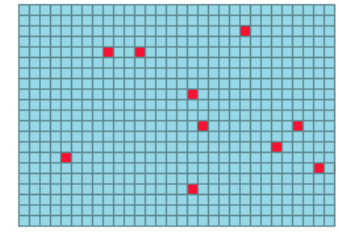
- x,y coordinates
- Measured value (attribute) at every location
  - Elevation, rainfall, pollution intensity, etc.
- Samples (incomplete) of continuous pattern



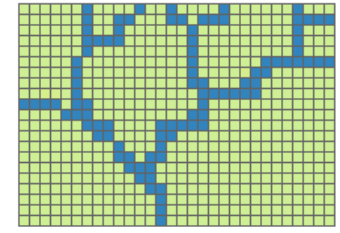
Data Types: Raster

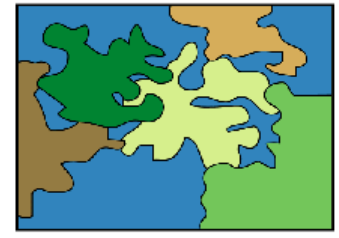
- Compared to vector: different data structure & analysis techniques
- Regularly spaced (continuous) grid
- Rapid computation and modeling due to it's simple structure
- Does not store discrete exact positions

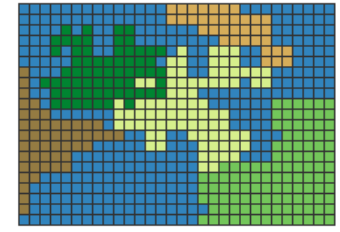

Point features


Raster point features


Line features


Raster line features


Polygon features


Raster polygon features

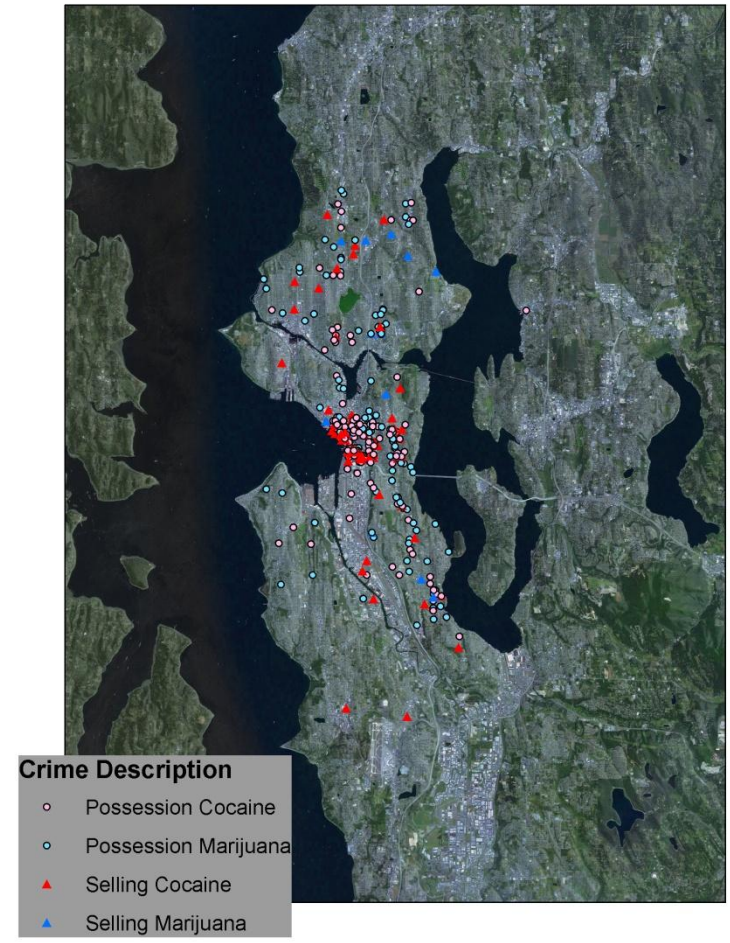
EVENT POINTS, DENSITY SURFACES & POINT PATTERN ANALYSIS

Section Outline

- Basic Assumptions
- Density Surfaces
  - Cell size
  - Neighborhood
  - Weighting mechanism
- Point Pattern Analysis
  - Distance analysis
  - Model Fitting

Data Types: Event Data

- x,y coordinates
- "marked" data contain some additional information.
  - Type of event, date of occurrence, etc…
- Understood as a complete set
  - Omissions can bias subsequent analysis



Event Data: Crime Statistics

- Developing techniques for focusing police activity

J Geograph Syst (1999) 1:385–398

Hotbeds of crime and the search for spatial accuracy

J.H. Ratcliffe, M.J. McCullagh

School of Geography, University of Nottingham, University Park Nottingham NG7 2RD, UK
(e-mail: michael.mccullagh@nottingham.ac.uk)

Received: 9 October 1998 / Accepted: 22 September 1999

Abstract. One of the most important aspects of spatial crime analysis is the identification of hotspots: areas of the highest crime concentration. This paper advances a methodology for hotspot detection based on a global moving window approach combined with the use of local statistics to define the hotspot limit. This technique generates hotspots that both follow the urban morphology of the crime distribution and ensures their spatial segregation. The hypothesis that police officers can construct an accurate perception of crime distribution from exposure to daily policing practices is used to demonstrate an application in the use of hotspot analysis. Significant regions generated from recorded crime data are compared with perceived local hotspots catalogued from surveys with police officers. Results from this study show two discrete types of hotspot, here termed hotpoints and hotbeds. The morphology of these crime hotpoints and hotbeds is discussed and possible causes documented.

Key words: Crime, hotspots, local statistics, police, mapping, GIS

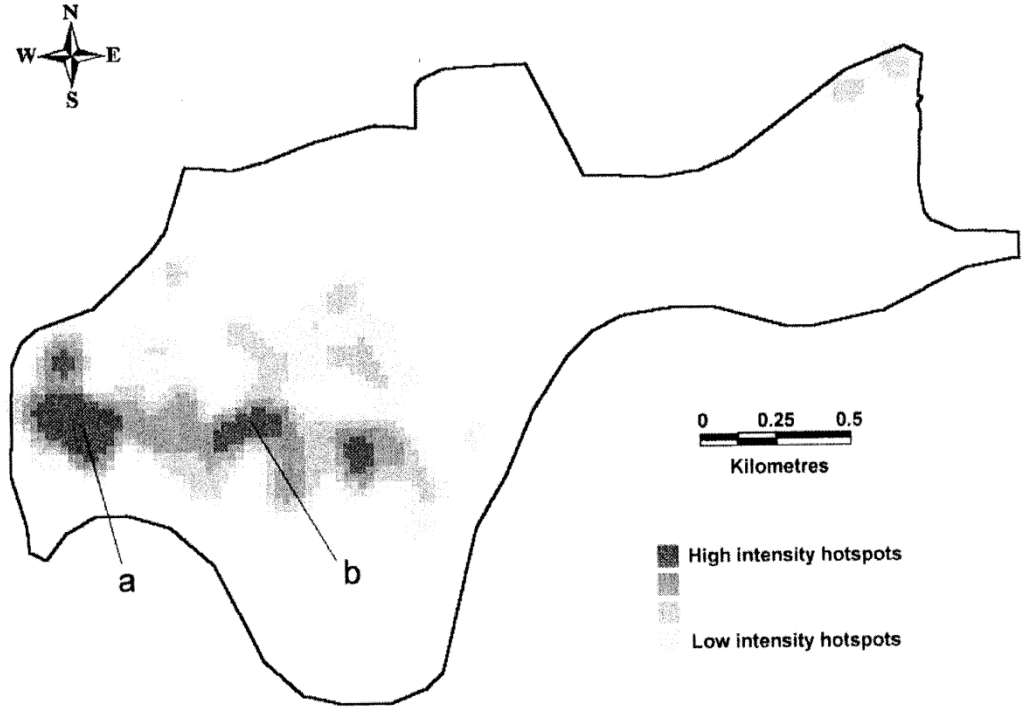


Fig. 1. Raster image of residential burglary intensity for Meadows sub-division. Regions (*a*) and (*b*) indicate two of the highest risk areas

Event Data: Ant Ecology

- Searching for patterns of collocation and dependence in ant nest locations

Appl. Statist. (1983),
32, *No.* 3, *pp.* 293–303

A Bivariate Spatial Point Pattern of Ants' Nests

By R. D. HARKNESS and VALERIE ISHAM

University College London, UK

[Received September 1982. Revised June 1983]

SUMMARY

The nesting behaviour of two species of ant (*Cataglyphis bicolor* and *Messor wasmanni*) is investigated by regarding their nests as a bivariate spatial point pattern. When the nests of each species are considered separately, those of the *Messors* are found to be more regularly spaced than would be the case if they were located at random. There is no strong evidence of such inhibition between the *Cataglyphis* nests. The possibility that the positions of the nests of *Cataglyphis* ants are dependent upon those of the *Messors* is suggested on biological grounds, and is investigated by various methods. No clear evidence for such an association is established.

Keywords: BIVARIATE POINT PATTERN; SPATIAL POINT PROCESS; STATISTICAL ANALYSIS OF SPATIAL DATA: *CATAGLYPHIS* AND *MESSOR* ANTS

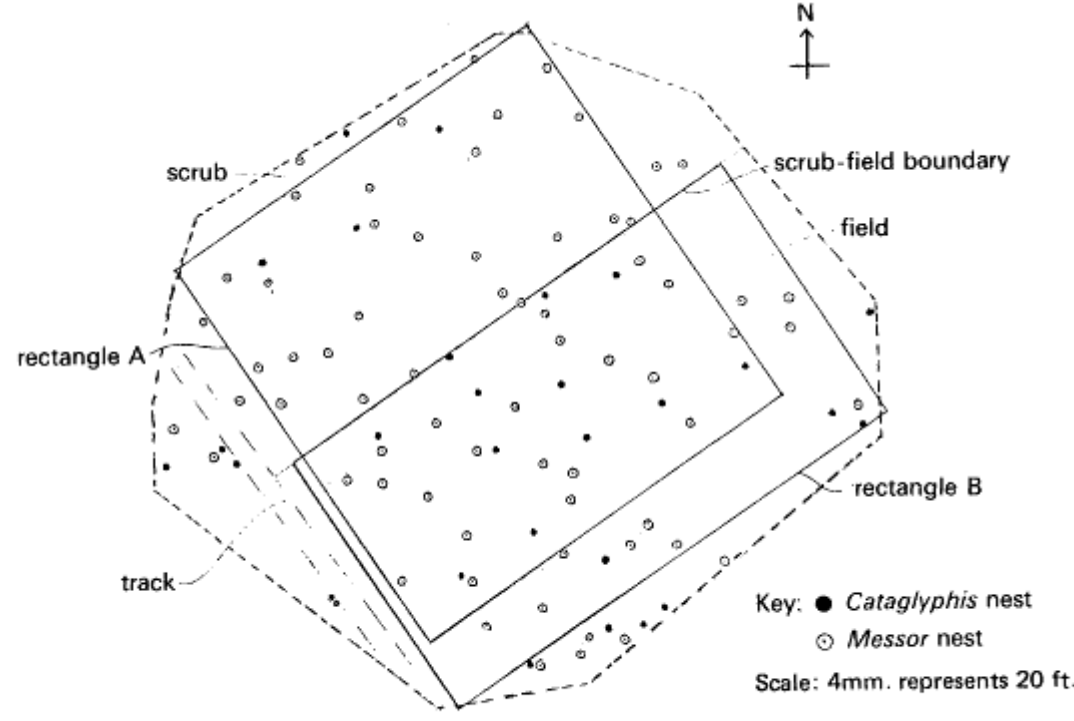


Fig. 1. Sites of nests of *Cataglyphis* and *Messor* ants.

Event Data: Neuroscience/Psychology

- Quantifying neuron location in schizophrenics

Further Evidence of Abnormal Cytoarchitecture of the Entorhinal Cortex in Schizophrenia Using Spatial Point Pattern Analyses

Steven E. Arnold, Delta D. Ruscheinsky, and Li-Ying Han

Previous studies have reported cytoarchitectural abnormalities in superficial laminae of rostral portions of the entorhinal cortex in schizophrenia, including decreased densities of neurons, poorly formed layer II neuron islands, and apparent displacement of layer II-type neurons deep into layer III; however, findings have been controversial, given the qualitative nature of the descriptions and the normal heterogeneity of cytoarchitecture of the region. The x, y coordinates of Nissl-stained neurons were mapped in layers II, III, and V of entorhinal subdivision ER in 8 prospectively accrued patients with schizophrenia and 8 nonneuropsychiatric controls. Indices of neuron dispersion, nearest neighbor distances, and effective radius were determined. An abnormally clustered dispersion of neurons in layer III was present in schizophrenics compared to controls along with a reduced neuron effective radius, whereas the mean nearest-neighbor distance was normal. In layer II, there was a significantly increased effective radius, whereas other indices were normal. No between-group differences were noted in layer V for any variable. These data provide further evidence for subtle aberrant cytoarchitecture in superficial laminae of the entorhinal cortex in schizophrenia and are consistent with neurodevelopmental models of abnormal neuronal pruning, "miswiring," and/or migration in the illness.

Key Words: Schizophrenia, entorhinal, cytoarchitecture, postmortem brain, quadrat, dispersion

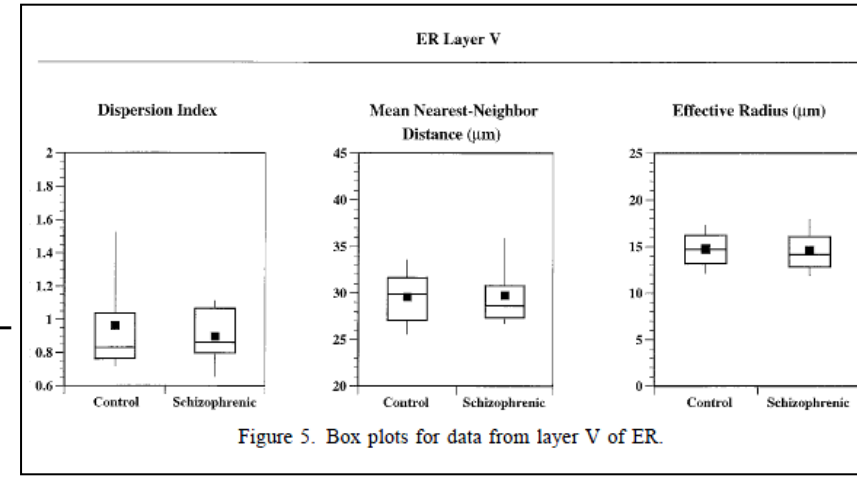


Figure 5. Box plots for data from layer V of ER.

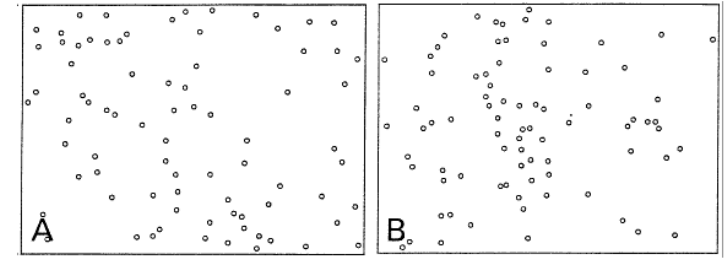


Figure 2. Program-generated point maps of the distribution of neurons in layer III of ER for a normal control (**A**) and a schizophrenic case (**B**). Note the greater degree of clustering that is visibly evident in the field from the schizophrenic patient compared to the control.

Data Types: Event Data

- Focus is on understanding why events occur in some places but not others
  - *Spatial Heterogeneity* (uniform, non-uniform)
  - *Spatial Dependence* (random, dispersed, clustered

- In answering these questions we may also seek to understand:
  - *Where* these processes occur, and
  - *At what scale* they are occurring.

Event Data: Basic Assumptions

- Are events equally likely everywhere in our study area?
  - Spatial Heterogeneity
- Are the locations of other events likely to influence the location of other events?
  - Spatial Dependence
- Purpose of analysis is to:
  - A)Identify the presence/absence of a visible pattern
  - B)Characterize the process that generates that pattern
  - C)Quantify extent/importance/variability in process

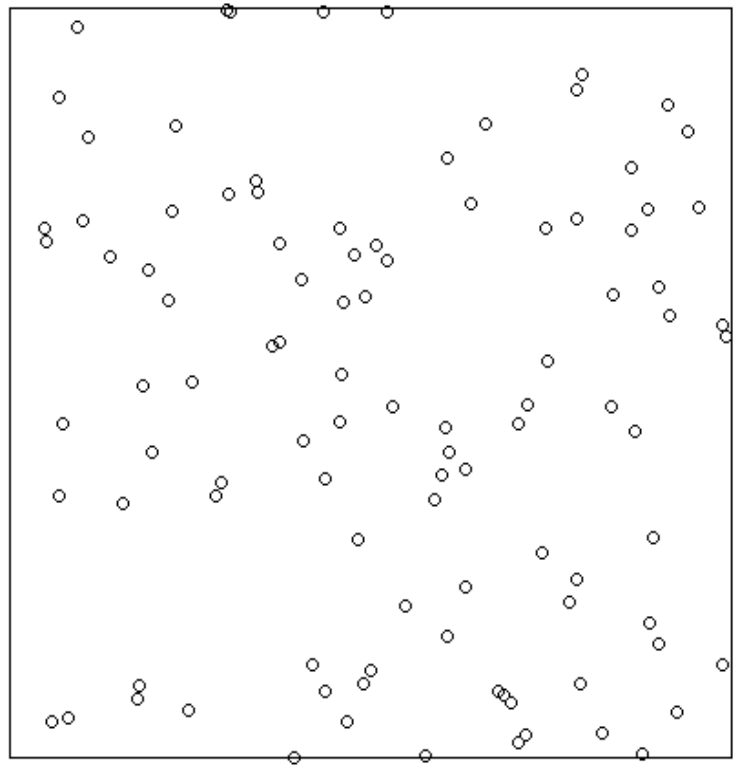
Event Data: Problems with our Data

- Can points occur *anywhere* and in any number?

- Cross sectional repeated observation can introduce dependence
- In some cases points will be centroids or street locations
- How accurate are your coordinates?
- What is your study area?

Complete Spatial Randomness (CSR)

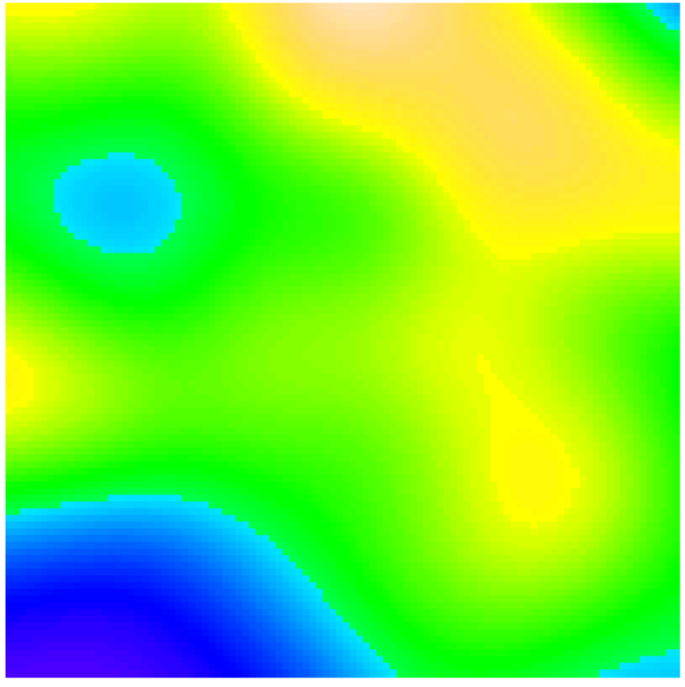
- Null hypothesis
- A homogenous Poisson point process
  - Each point has a uniform probability of getting any value of x and y
  - Points are neither attractive or repulsive
  - Points selected with replacement--they can be stacked

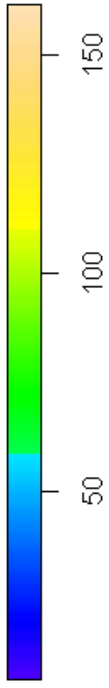


Spatial Heterogeneity

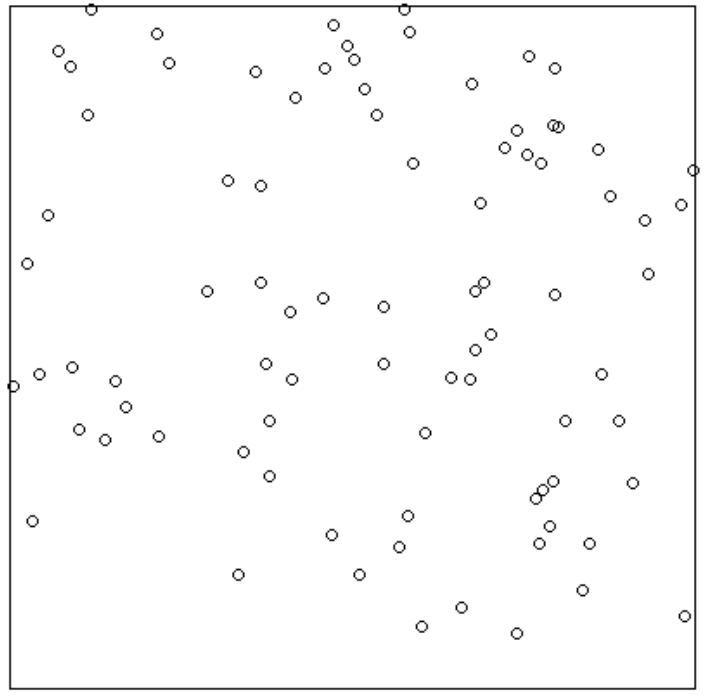
- What if we were looking at mountain goat sightings and had an expectation that they would be more likely at higher elevations….

Given a map of elevation like this….



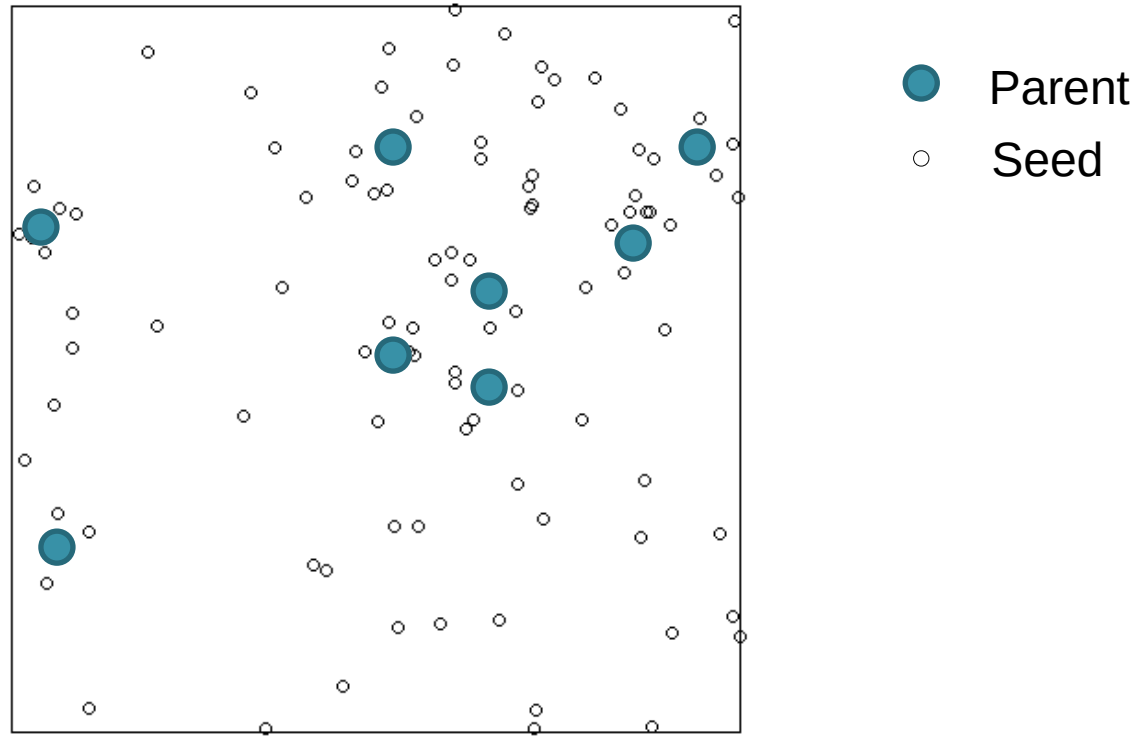


We might expect goat sightings like this



Spatial Dependence

- What if a plant randomly scatters its seeds within a 10 meter distance of itself, with increasing frequency closer in.



To Reiterate…

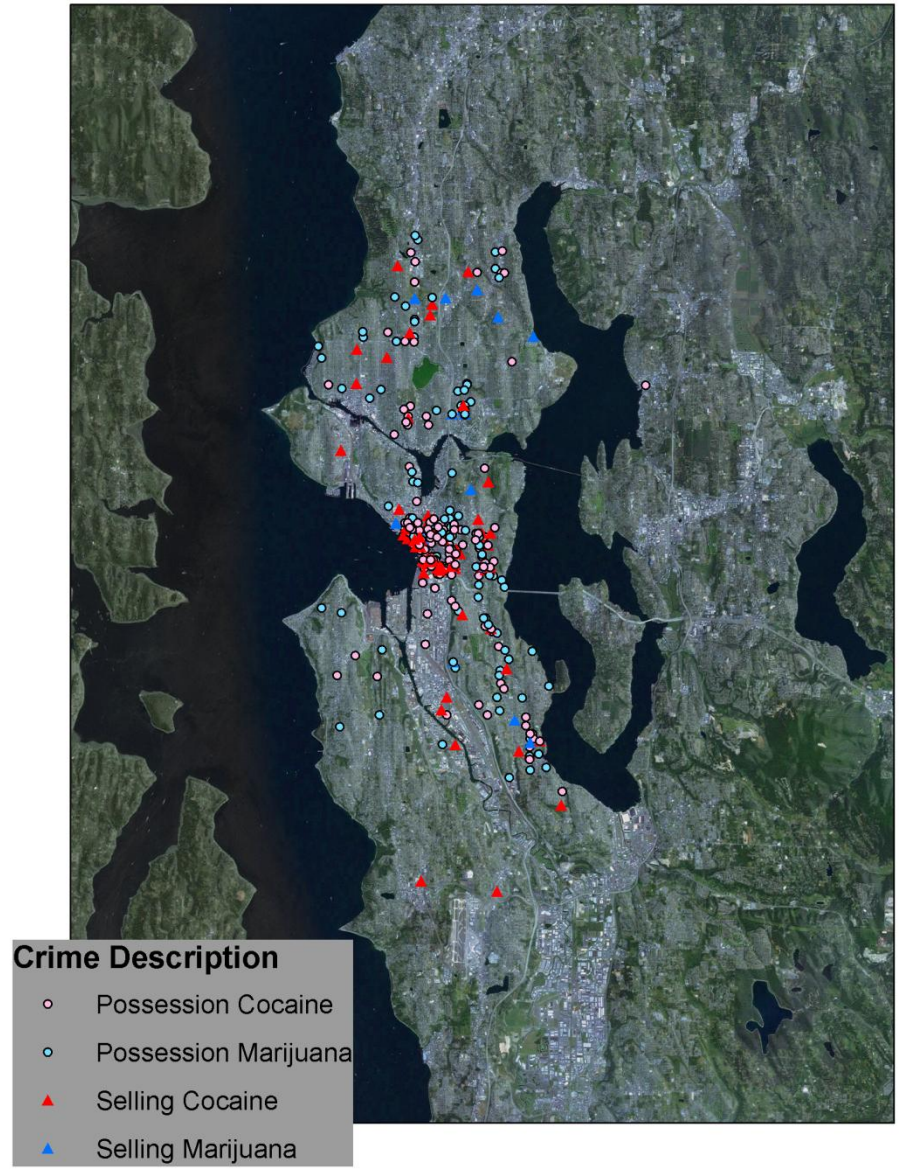
- Purpose of analysis is to:
  - A)Identify the presence of some pattern
  - B)Characterize the process that generated that pattern
  - C)Quantify
    - Extent → K function, Density Surface
    - Importance → Nearest Neighbor, Model fitting
    - Variability → Density Surface

Step 1: Understand Your Data (and your study area)

- Map it.
- Descriptive Statistics
  - Geographic Mean
  - Standard Distance
  - Density
- Density Surface
  - Quadrats
  - Other Density Measures

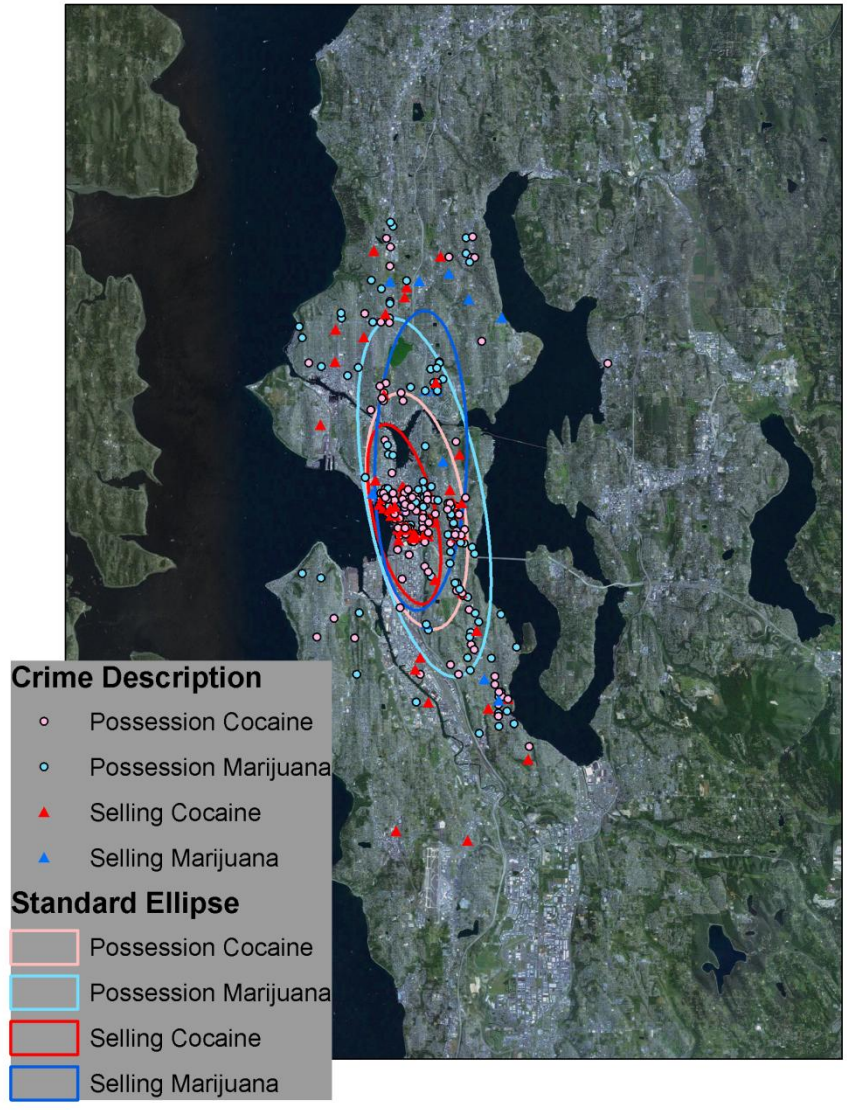
Mapping Event Data

- Where points are actually located.
- Possibly divided by type
- Background conditions underneath for comprehension



Geographic Mean and Standard Distance

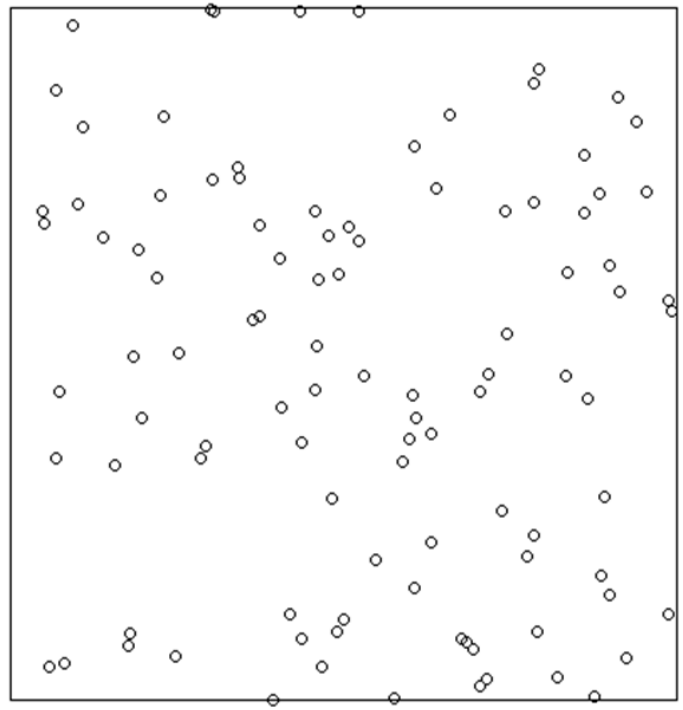
- Center, possibly weighted by event characteristic.
- Minimum area ellipse containing 66% of all cases in study



Describing Point Event Data

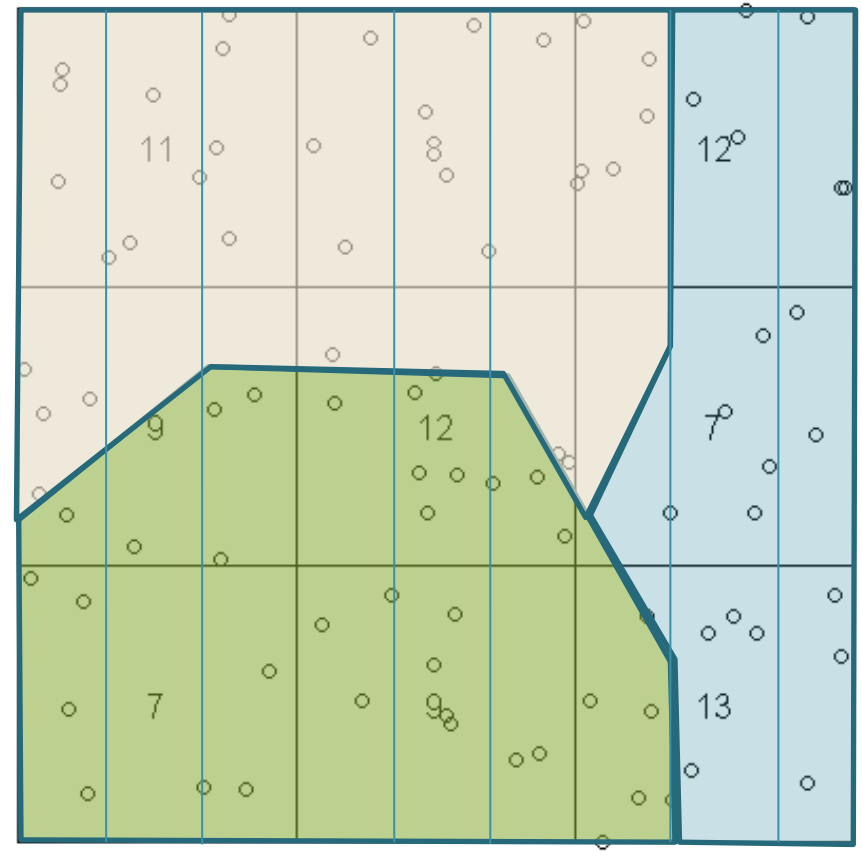
- *Global Point Density*
  - Equivalent to taking the mean
  - **Points per unit distance squared**
  - A *global* measure

  - 100 points in a 10m x 10m area = 1pt/m^2



Decomposing Density—How good a fit is the global measure?

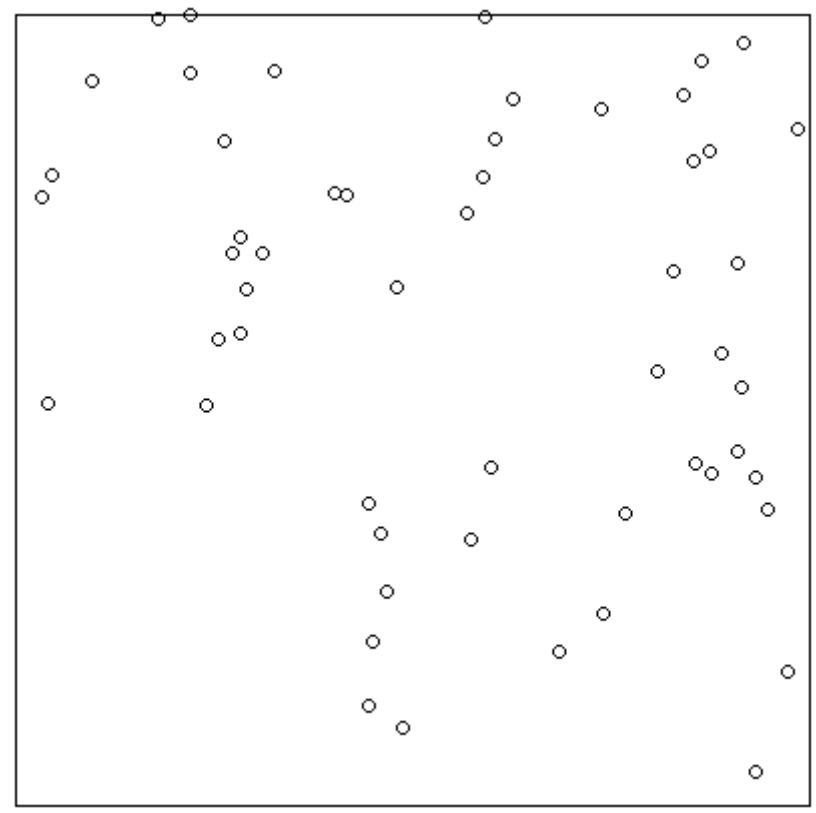
- Same Study Area
- Density = 1 per m^2

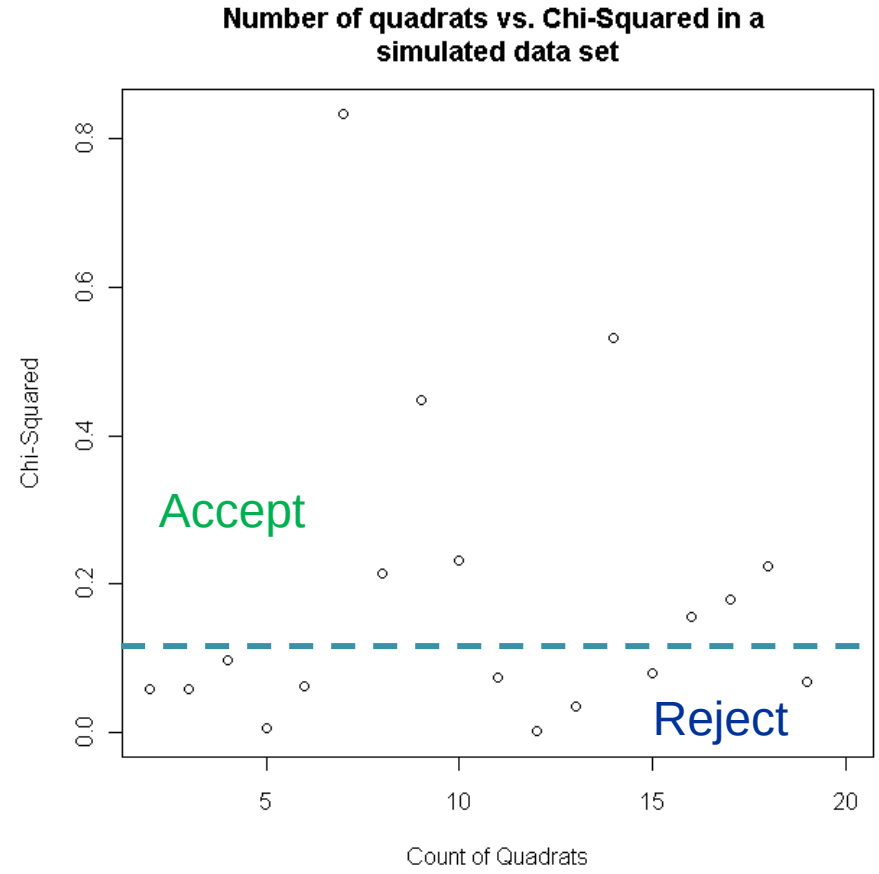


- Define a 3 by 3 matrix (3.33m x 3.33m Quadrats)
- Count points per quadrat
- Compare to expected

$$\frac{100\ points}{100\ m^2} = \frac{\frac{100 points}{9 quadrats} = 11.1\ points}{\frac{100\ m^2}{9\ quadrats} 11.1\ square\ meters}$$

How useful are quadrats and associated Chi-Square measures?



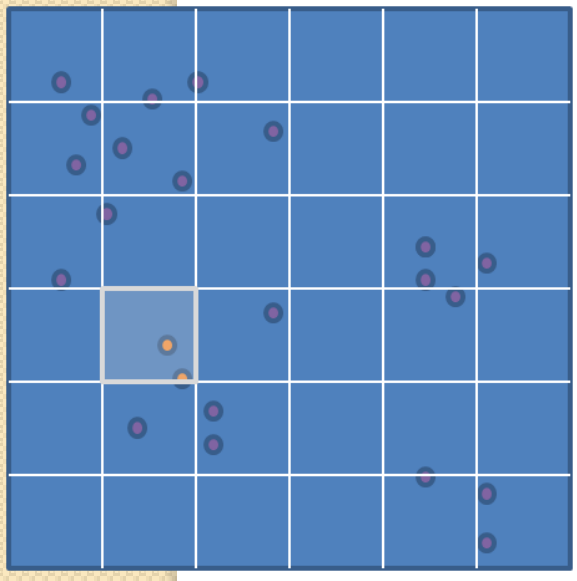


Given simulated data where spatial heterogeneity is present (we generate the points non-randomly)…

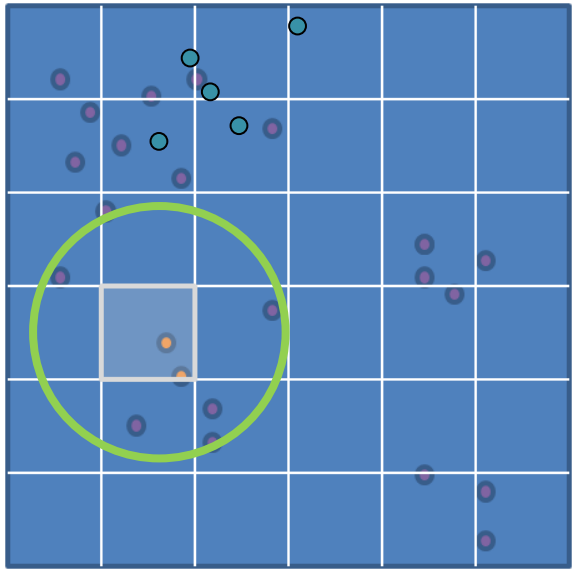
Whether we accept or reject a difference from CSR depends on how we draw quadrats—Type II error 8 out of 18 times!

Density Surfaces: Quadrat, Point, & Kernel Density

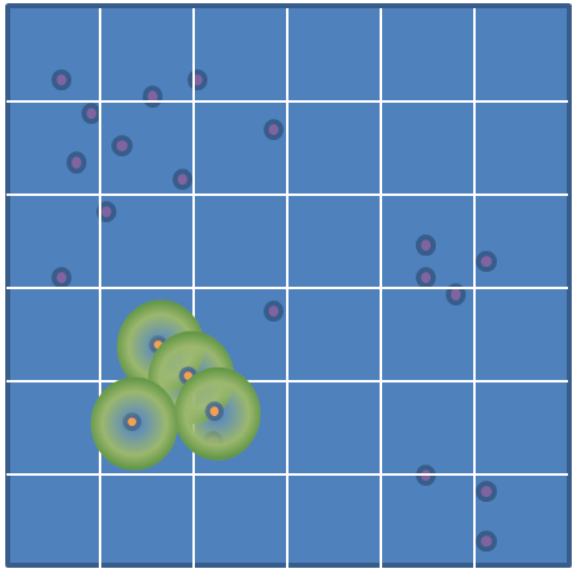
Quadrat Density (clustering)



Point Density



Kernel Density



Lab Part I: Descriptive Statistics and Density Surfaces with Seattle Crime Data

- Subsetting our Data
- Standard Ellipses
- Quadrats
- Point Density
- Kernel Density

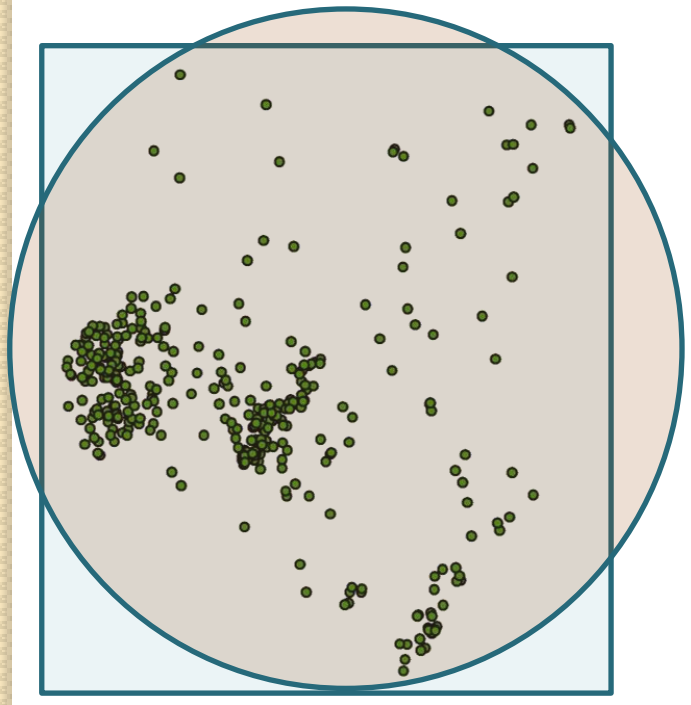
Step 2: Quantify Clustering and Spatial Extent

- Nearest Neighbor Analysis
- K function

Nearest Neighbor Analysis

- Tests for and describes the structure of spatial relationships among events.
  - Average distance from each feature to its nearest neighboring feature
  - Randomness (variation) of the pattern gauged against expected index for study area size
  - Simple, but fast

Nearest Neighbor Analysis



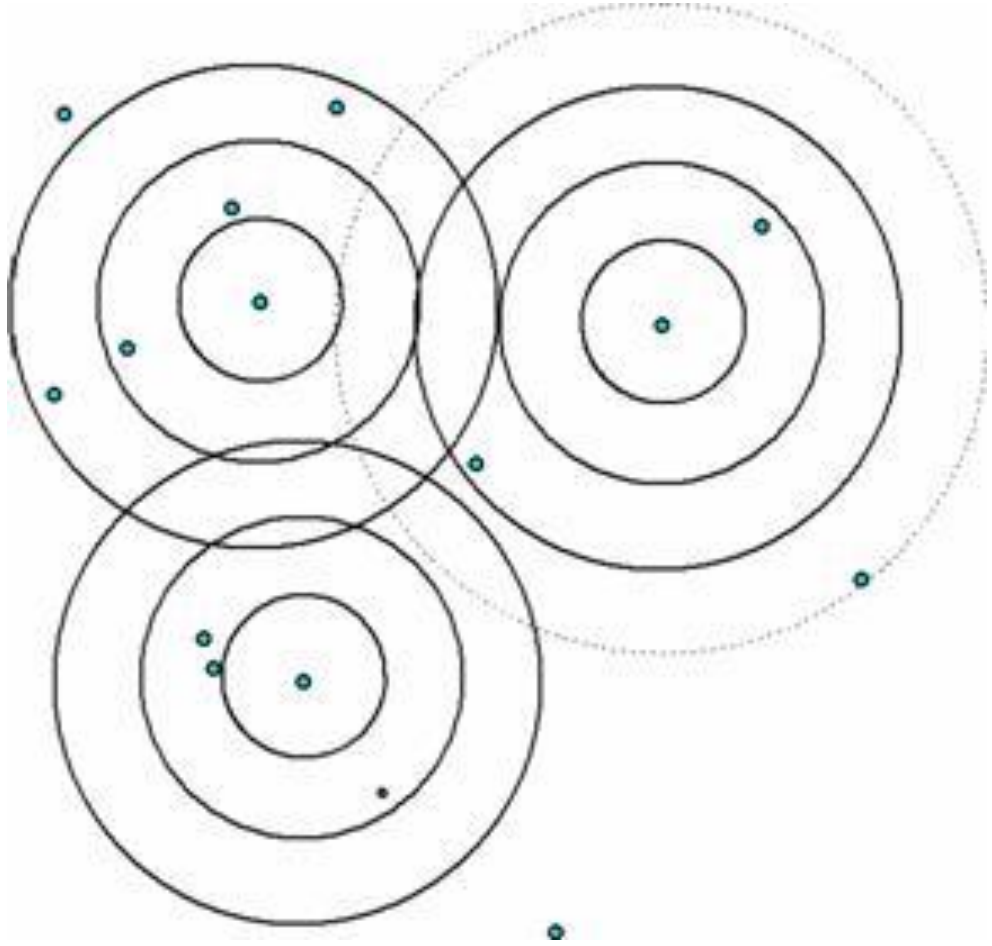
- Depends heavily on how we define the study area:
  - Minimum Bounding Rectangle
  - Standard Distance Circle
- Boundary effects will also be relevant

Ripley's K-function

- Similar to multiple point densities, but provides a global statistic, rather than a cell by cell reference value
- Estimates the scale(s) at which clustering occurs

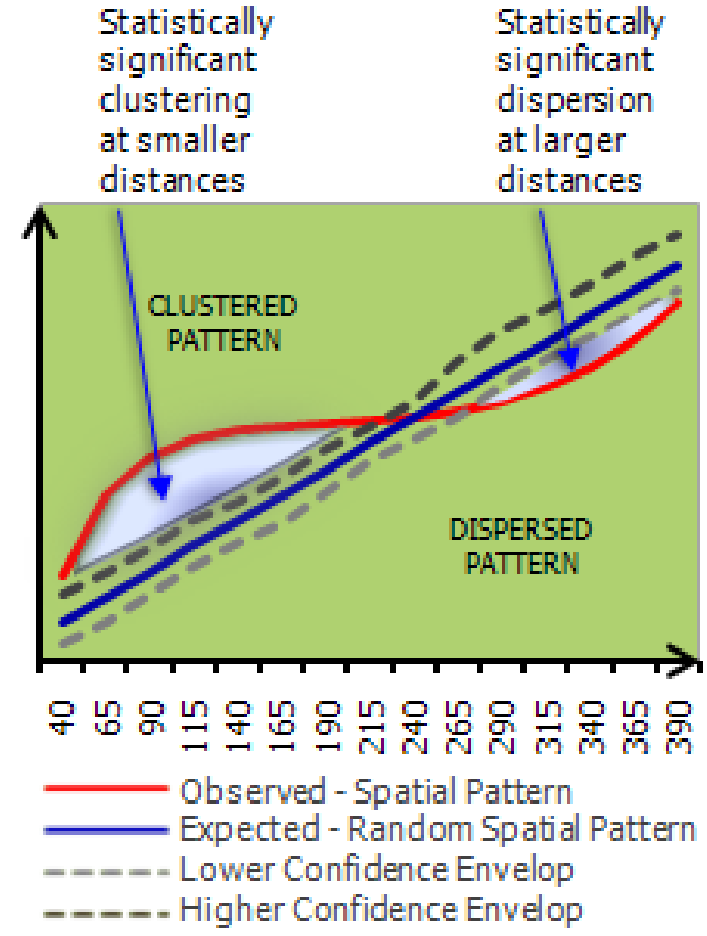
Ripley's K-function

- For each point i and circle radius d count the points j that are within the area of the circle and divide by n, the total number of points



Ripley's K-function

- ArcGIS has options for adjusting assumptions about study area and boundary effects
- Method can take an extraordinary amount of time to process for large numbers of points



Lab Part II:

- Average Nearest Neighbor
- Ripley's K function

Step 3: Going Further

- We still don't know *where* clusters are occurring
- We still haven't worked through a process to compare theoretical explanations for why we observed the patterns we did

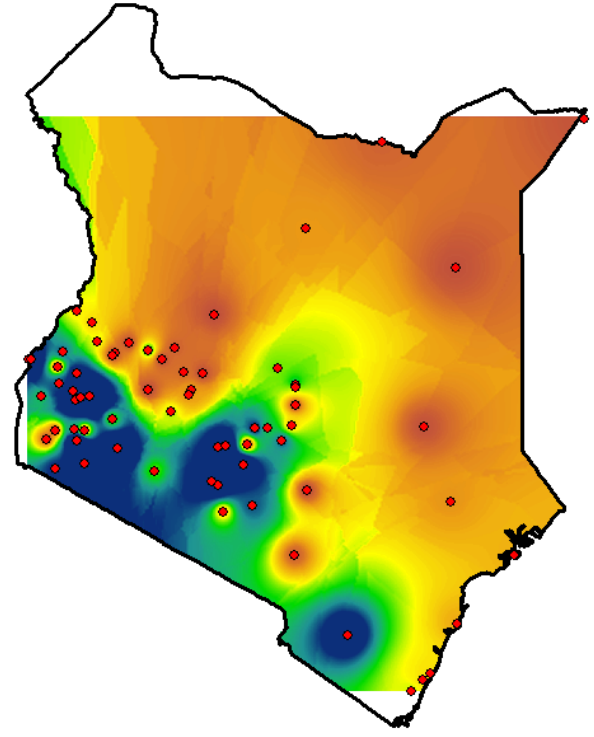
SAMPLE POINTS & INTERPOLATION

Outline

- Point Sample Data
- What is Interpolation?
- Methods of converting samples to surfaces
- Simple: Nearest Neighbor/Moving Average,
- Deterministic: IDW/Trend Surface
- Geostatistics: Kriging

Data Types: Point Sample Data

- X,Y coordinate pairs with measured value (attribute) at every point location
- Samples of larger continuous pattern
- Randomly or Regularly Spaced
- By their very nature, point samples are understood to be an incomplete subset of overall pattern/process

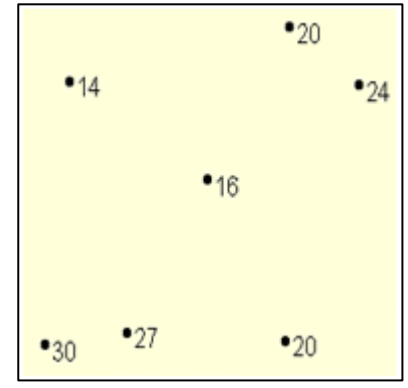


What is Interpolation?

- *Surface Interpolation* creates a continuous (or prediction) surface from sampled point values
- Assumption: spatially distributed objects are spatially correlated (things closer together tend of have similar characteristics)
- E.g. rainfall on one side of street is good predictor of rainfall on the other side
  - Other Side of Town? Other side of County?

Interpolating a Rainfall Surface

- Input: point dataset of known rainfall-level values
- Output: raster interpolated from these points.
- The unknown values are predicted with a mathematical formula that uses the values of nearby known points.



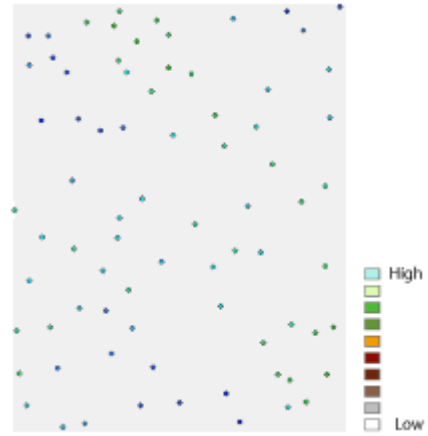
Input rainfall point data

13	14	16	20	23
14	14	16	19	24
18	16	16	18	22
24	22	19	19	21
30	27	23	20	20

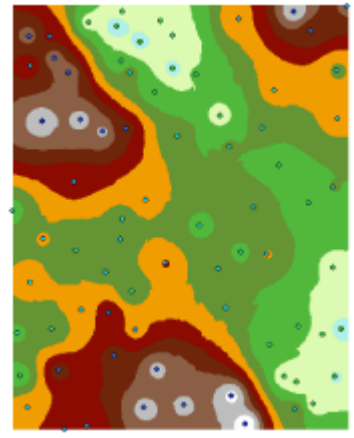
Interpolated rainfall surface

Interpolating an Elevation Surface

- A typical use for point interpolation is to create an elevation surface from a set of sample measurements.
- Each symbol in the point layer represents a location where the elevation has been measured. By interpolating, the values for each cell between these input points will be predicted.



Input Elevation Point Data

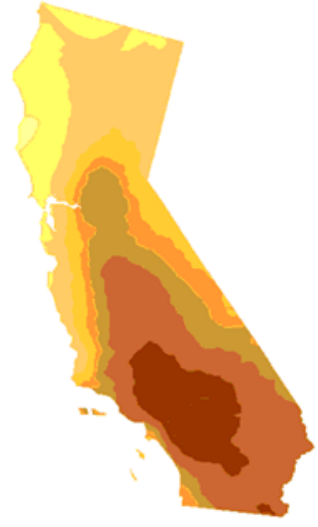


Interpolated elevation surface

Interpolating a Concentration Surface

- Study the correlation of the ozone concentration on lung disease in California
- Ozone measurements at locations monitoring stations used to generate interpolated surface, providing predictions for any location in California

Point locations of ozone monitoring stations



Interpolation prediction surface

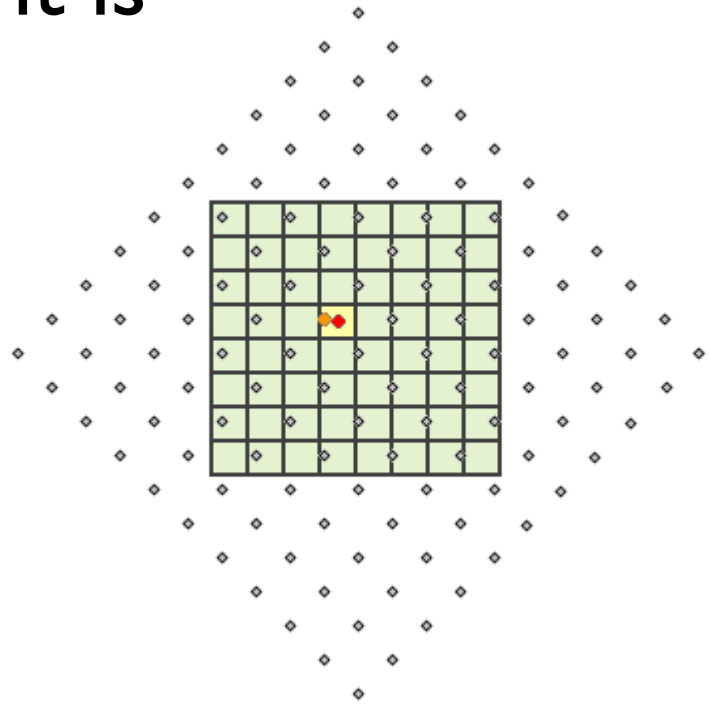
Interpolation Techniques

There are a variety of methodologies used to convert from sample points to a surface

- Simple (not in ArcGIS)
  - *Nearest Neighbor & Moving Average*
- Deterministic (mathematical formulas determine smoothness of surface)
  - *Inverse Distance Weighting (IDW) and Trend Surface*
- Geostatistical (statistical models include autocorrelation)
  - *Kriging*

Nearest Neighbor Interpolation

- Simply assign the value of a cell to that of the nearest known sample point.
- Most appropriate when sample points are somewhat regular, or when it is simply a matter of filling in some missing data.
- Creates a stepped surface



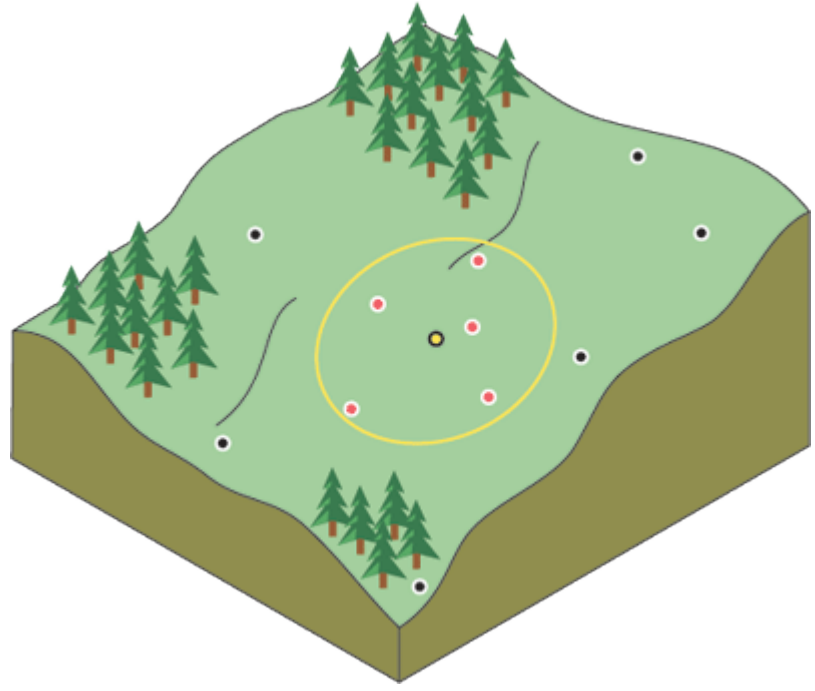
Moving Average Interpolation

- User defines a circle of a given radius
- For each grid cell we center the circle and average the values of any points that fall within the circle
- Zero values if no sample points fall within the circle

Similar technique was used to create a density surface: calculate sum, rather than average, of point values and divide result by the area of the circle

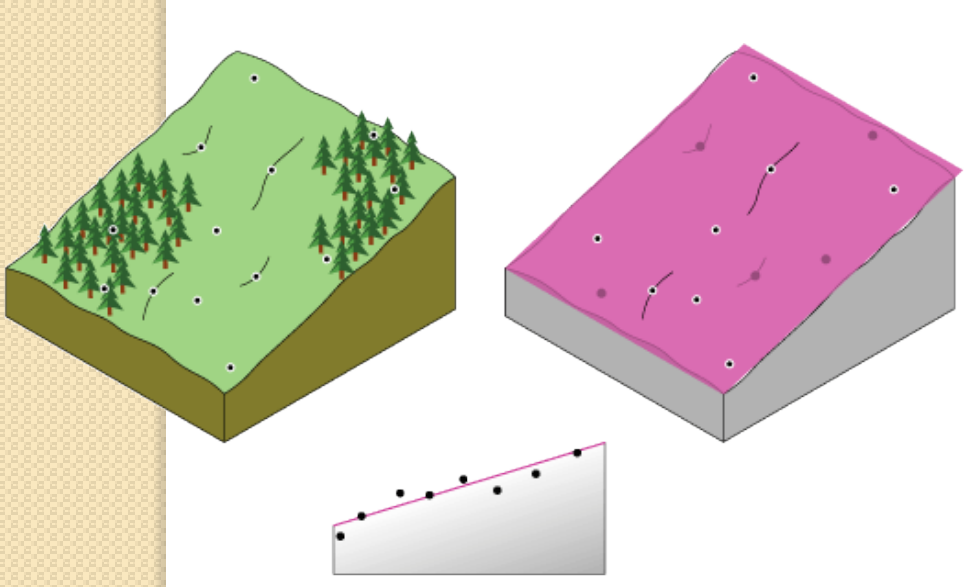
Inverse Distance Weighting

- Grid surface based on the sample values weighted by their distance from the cell in question
- Closer values weighted higher than distant values, thus the name
- Parameters:
  - Rate of decay (Power)
  - Points used (Variable or Fixed radius)
  - Barriers/Breaklines

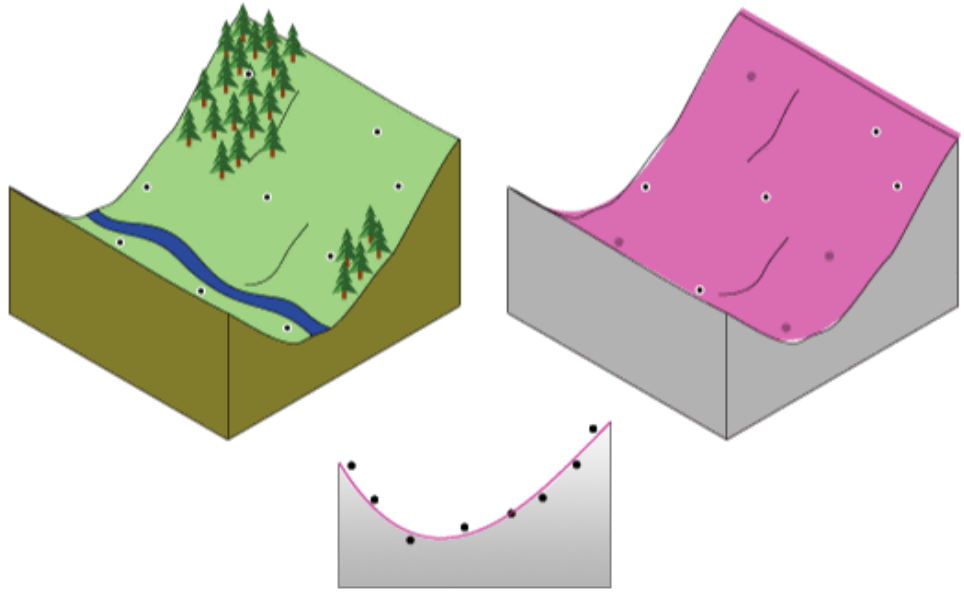


Trend Surface Analysis

- Technique for fitting smooth surface (global polynomial function) to sample points – often exploratory



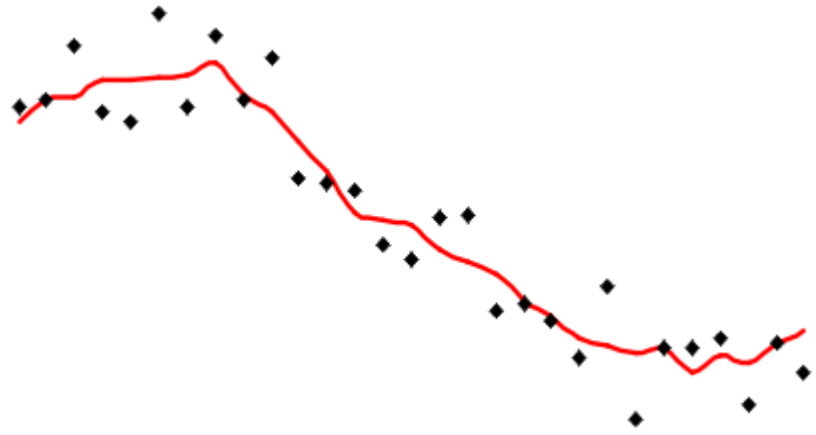
First-order polynomial (linear) Trend Surface



Second-order polynomial (quadratic) Trend Surface

Limitations

- All models cannot predict a value beyond the range of the sample data
- The most extreme values in any map produced from sample data, will be values already in the sample data, and not values at unmeasured locations.
- Red line (interpolated) is less extreme than any of the sample values represented by the point symbols.
- Strong assumption that kriging attempts to address



average of the nearest 5 observations to interpolate values

Lab Part III:

- Inverse Distance Weighted Interpolation
- Trend Surface Analysis

Geostatistics

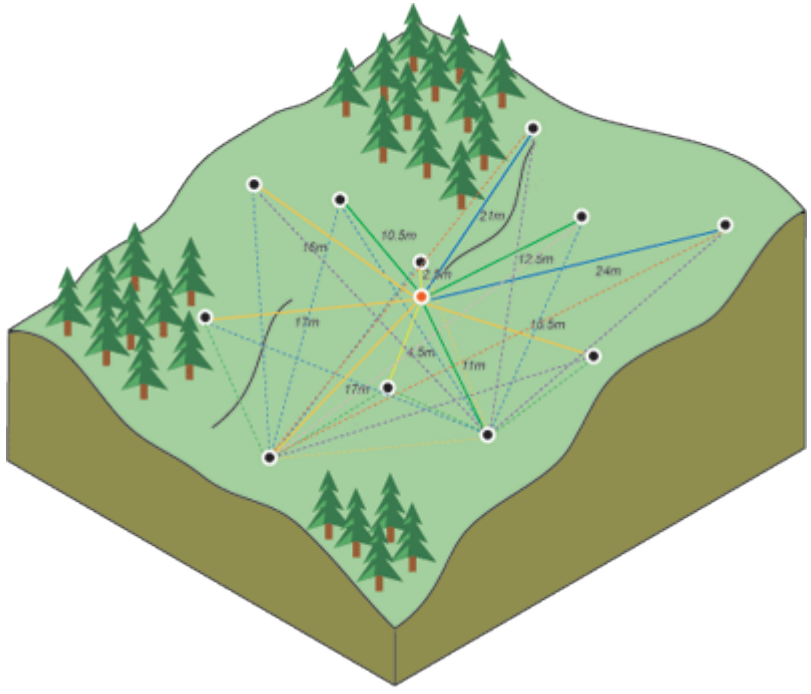
- Grew from
  - Meteorologists: interpolate weather characteristics from sparse data
  - Mining Engineers: estimate quantities of minerals in bodies of rock from drill cores
- Based on statistical methods that include measure of spatially correlated variation or autocorrelation (statistical relationships among points). Produce both prediction surface and measure of certainty/accuracy.

Geostatistics - Kriging

- Kriging is the primary geostatistical interpolation method implemented in ArcGIS

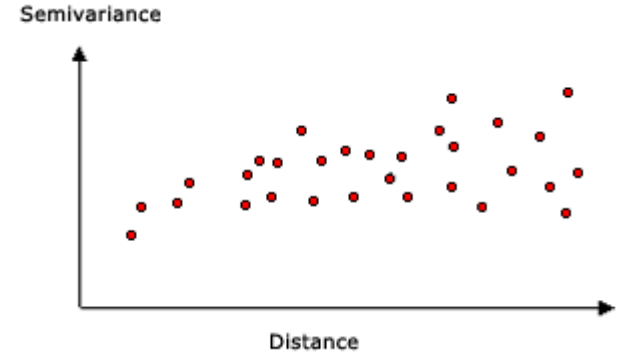
- Multistep process:
  - Exploratory statistical analysis of data
  - Variogram modeling*
  - Creating the surface*
  - (optional) Exploring a variance surface

Uncover dependency rules – Semivariogram model

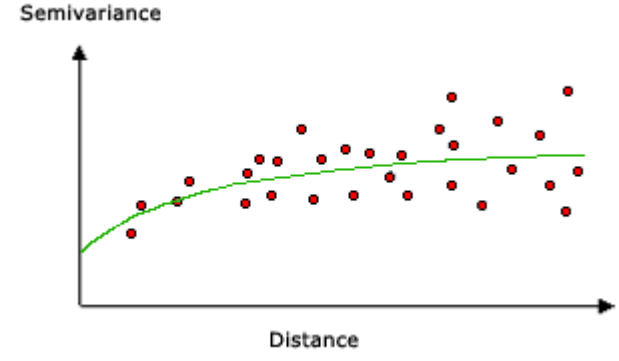


1) Calculated difference squared b/w all paired point locations

$Semivariogram(distance_h) = 0.5 * average\{(value_i - value_j)^2\}$



2) Group & average by distance



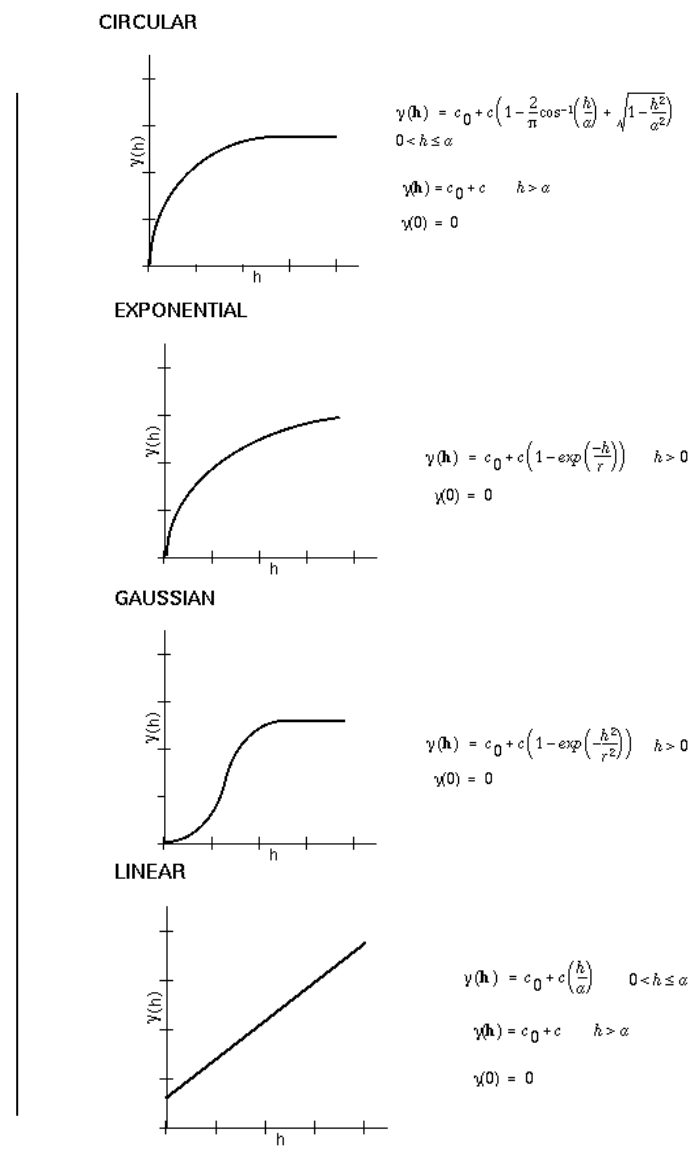
3) Fit semivariogram model function

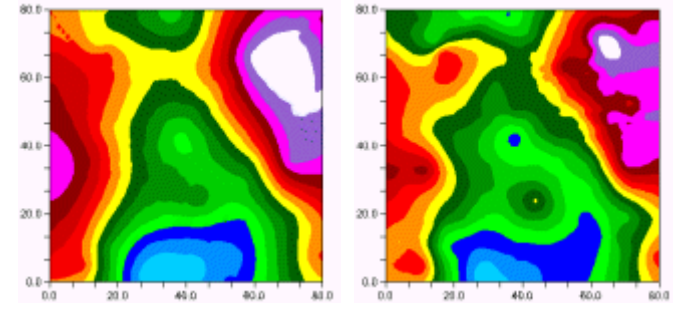
Making a prediction

- IDW uses a simple algorithm based on distance, but kriging weights come from the semivariogram that was developed by looking at the spatial nature of the data.
- Predictions are made for each location, or cell centers, in the study area based on the semivariogram and the spatial arrangement of measured values that are nearby.

Kriging Comments

- Complex method, but simple stats in individual parts (do your homework first)
- Exploratory, Predictive, with Measure of Certainty (what's not to love)
- Part of ArcGIS Interpolation toolbox, ALSO entire Geostatistics Extension
- Co-Kriging (includes co-variant)





Plutonium

Plutonium w/ Carbon

Geostatistics - Readings

- Royle, A. G., F. L. Clausen, and P. Frederiksen. (1981) "Practical Universal Kriging and Automatic Contouring." *Geoprocessing* 1: 377–394.
- Oliver, M. A. (1990) "Kriging: A Method of Interpolation for Geographical Information Systems." *International Journal of Geographic Information Systems* 4: 313–332.
- Burrough, P.A. (2001) "GIS and Geostatistics: Essential Partners for Spatial Analysis." *Environmental and Ecological Statistics*, 8:361-377.

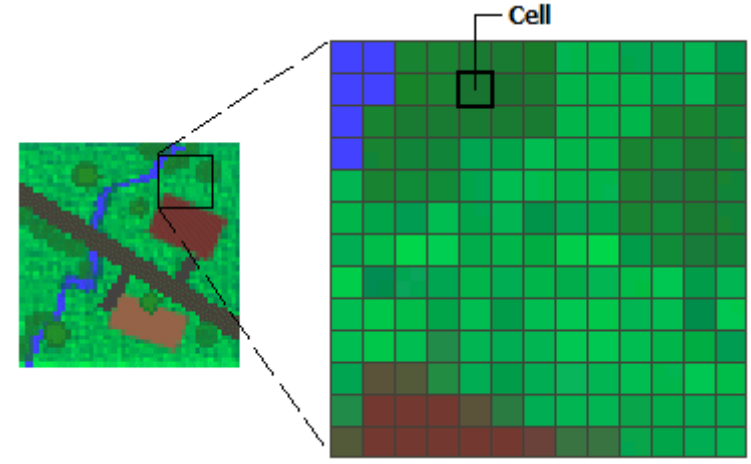
RASTERS BASICS AND MAP ALGEBRA

What is raster data?

- Matrix of cells (or pixels) organized into rows and columns (or a grid) where each cell contains a value representing information, such as temperature.

Rasters are digital aerial photographs, imagery from satellites, surfaces (often derived from density or interpolation), or even scanned maps.

- Thematic data (also known as discrete) represents features such as land-use or soils data.
- Continuous data represents phenomena such as temperature, elevation, or spectral data such as satellite images and aerial photographs.

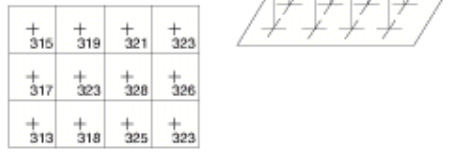


Thematic and continuous rasters may be displayed as base-data layers along with other geographic data on your map but are often used as the source data for further spatial analysis.

Characteristics of Raster Data

Value applies to the center point of the cell

For certain types of data, the cell value represents a measured value at the center point of the cell. An example is a raster of elevation

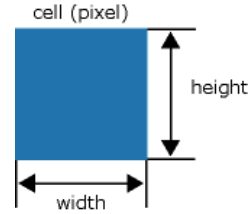


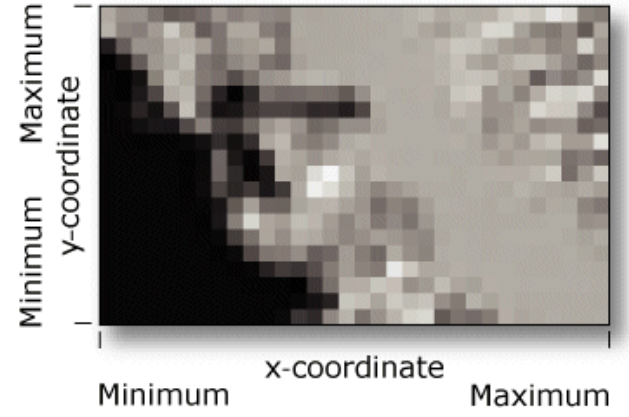
315	319	321	323
317	323	328	326
313	318	325	323

Value applies to the whole area of the cell

For most data, the cell value represents a sampling of a phenomenon, and the value is presumed to represent the whole cell square.

50	45	40	35
35	40	35	25
20	25	30	20

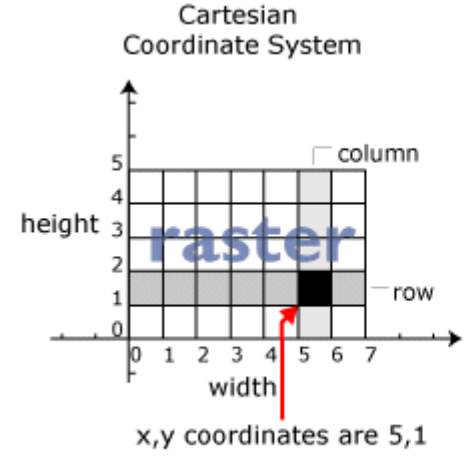


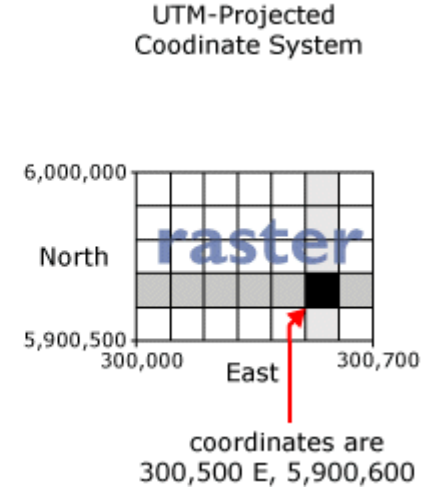


Cell Dimensions vs. Dataset Dimensions

80	74	62	45	45	34	39	56
80	74	74	62	45	34	39	56
74	74	62	62	45	34	39	39
62	62	45	45	34	34	34	39
45	45	45	34	34	30	34	39

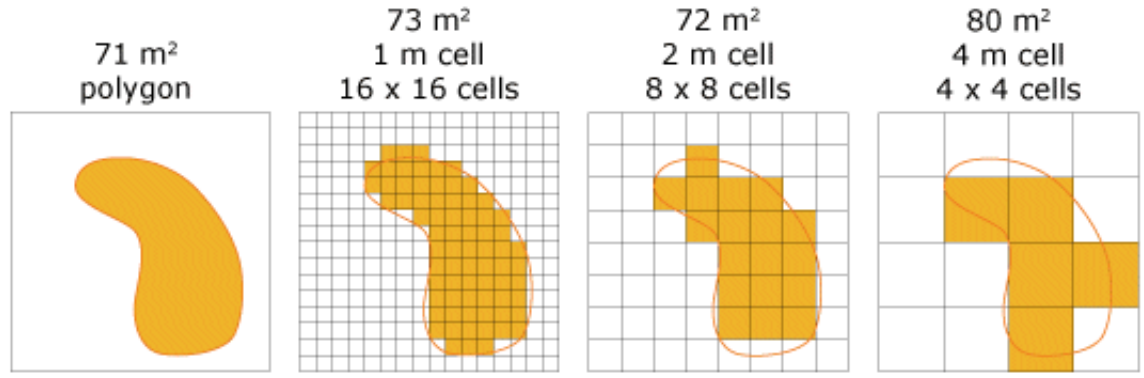
Stored as ordered list of values (80, 74, 62, 45…)





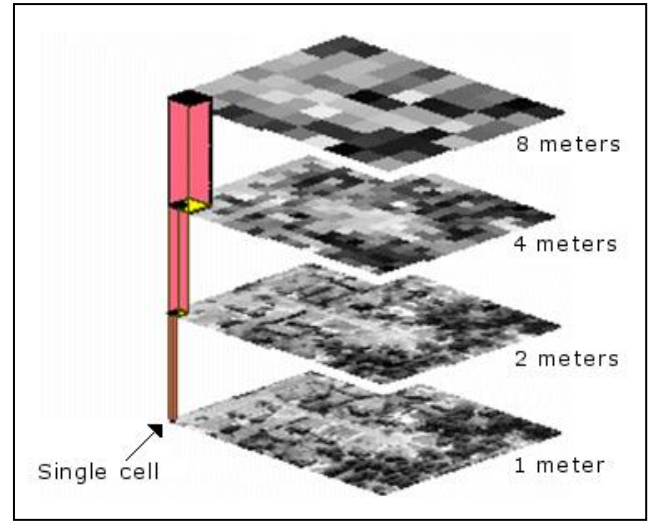
Coordinate System w/ Projection

Raster Level of Detail = Cell Size/Spatial Resolution



- Smaller cell size
- Higher resolution
- Higher feature spatial accuracy
- Slower display
- Slower processing
- Larger file size

- Larger cell size
- Lower resolution
- Lower feature spatial accuracy
- Faster display
- Faster processing
- Smaller file size



Spatial Resolution vs. Scale

Scale 1:50,000
Cell size: 61 cm

Scale 1:2,500
Cell size: 61 cm

Scale 1:20,000
Cell size: 15 m

Scale 1:20,000
Cell size: 15 m

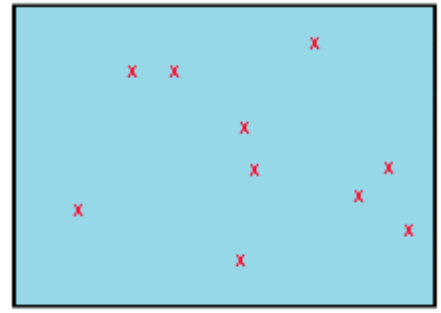
Resolution : Dimensions of cell representing Area on the Ground

Scale : Ratio of distance on the map to corresponding distance on the ground

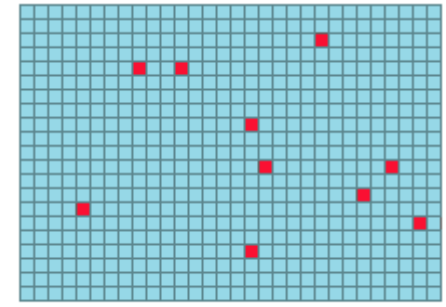
Higher resolution raster (smaller cell size) has greater detail

Smaller scale shows less detail

How vector features are represented in a raster

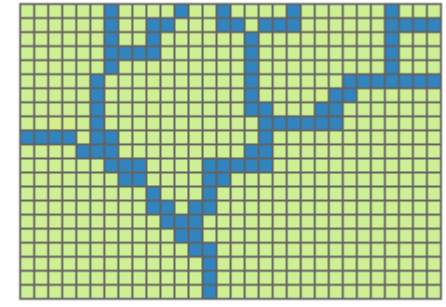


Point features

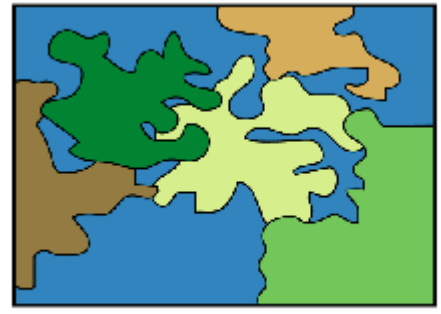


Raster point features

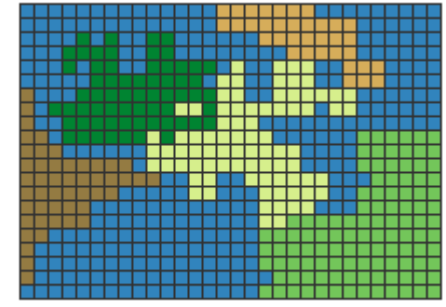
Line features



Raster line features



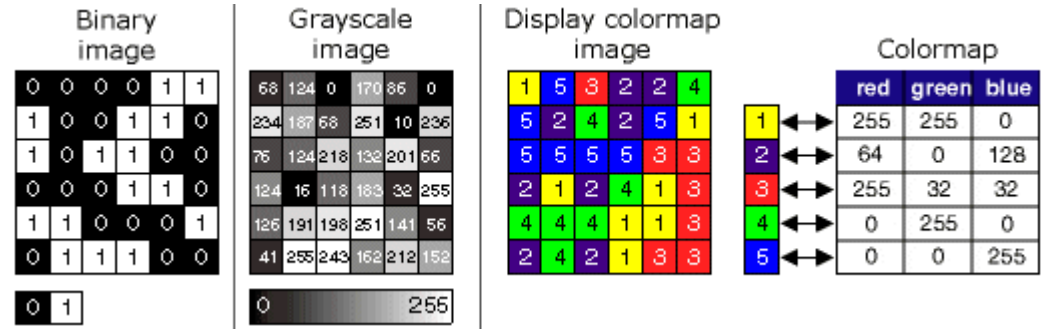
Polygon features



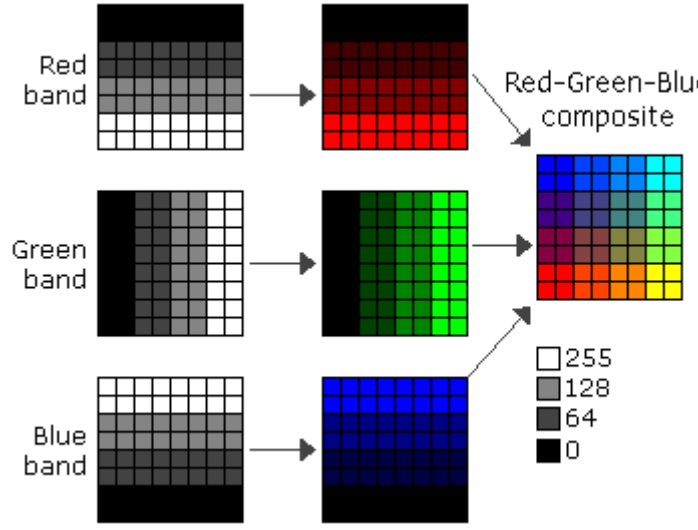
Raster polygon features

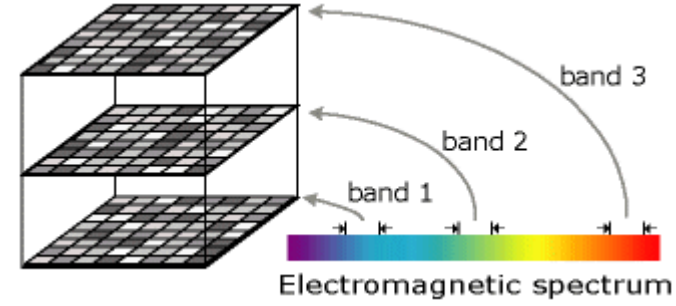
Storing Data: Raster Bands

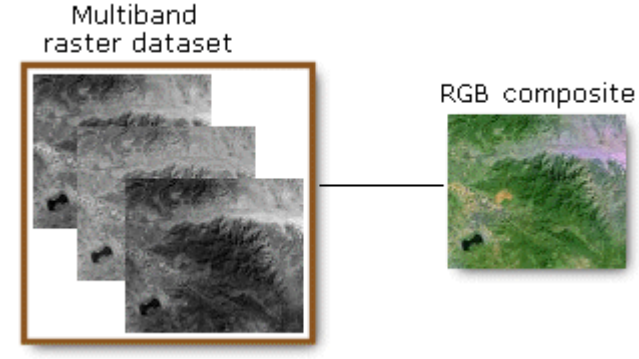
- Single Band



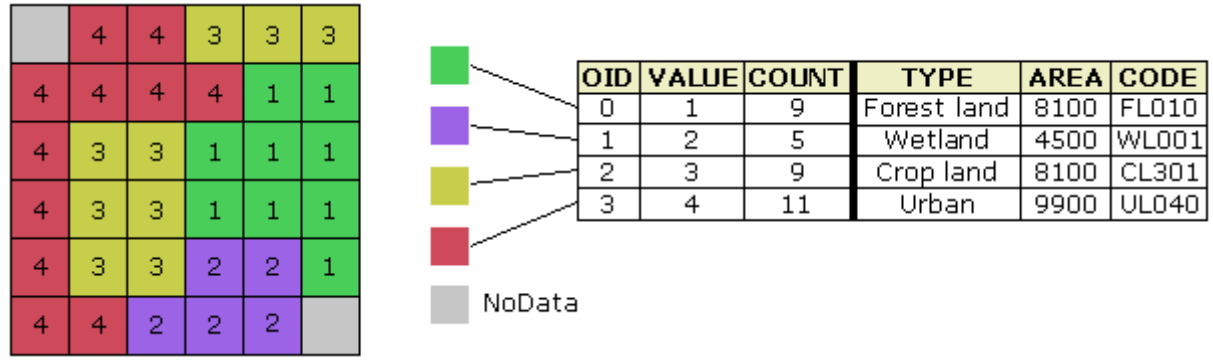
- Multiband







Storing Data: Raster attribute tables



OID	VALUE	COUNT	TYPE	AREA	CODE
0	1	9	Forest land	8100	FL010
1	2	5	Wetland	4500	WL001
2	3	9	Crop land	8100	CL301
3	4	11	Urban	9900	UL040

- Cell values represent or define a class, group, category, or membership
- Maintain original attribute information (left) and store additional fields (right)

Raster Analysis

- **Derive new information**
  Use tools to calculate new data resources, such as distance from roads or population density
- **Identify spatial relationships**
  Explore relationships between layers through weighted overlay and combinations
- **Find suitable locations**
  Combine layers to find areas that are the most suitable for particular objectives
- **Calculate travel cost**
  Create travel cost surfaces to identify optimum corridors, factoring in economic, environmental, and other objectives

Site Location Example: New School in Stowe, VT

- Step 1: Input Datasets
- Step 2: Derive Datasets
- Step 3: Reclassify Datasets
- Step 4: Weight and Combine