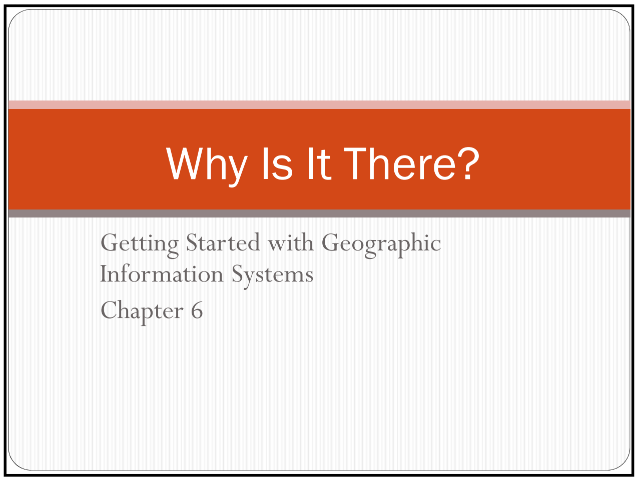
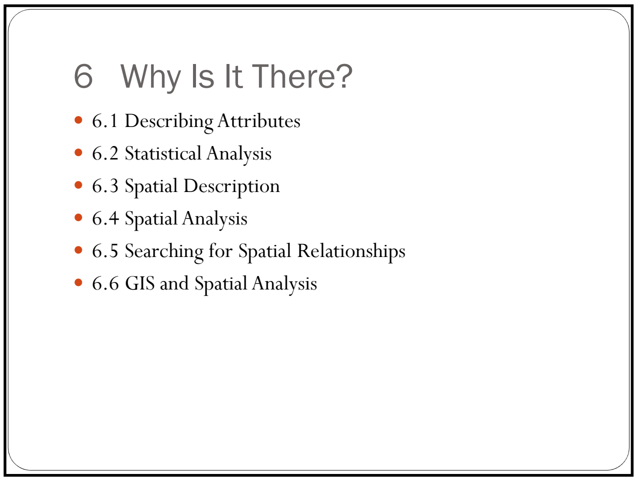


Why Is It There?

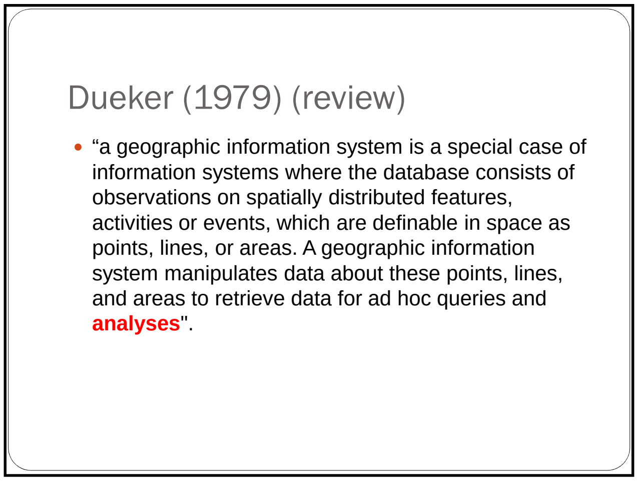
Getting Started with Geographic Information Systems

Chapter 6



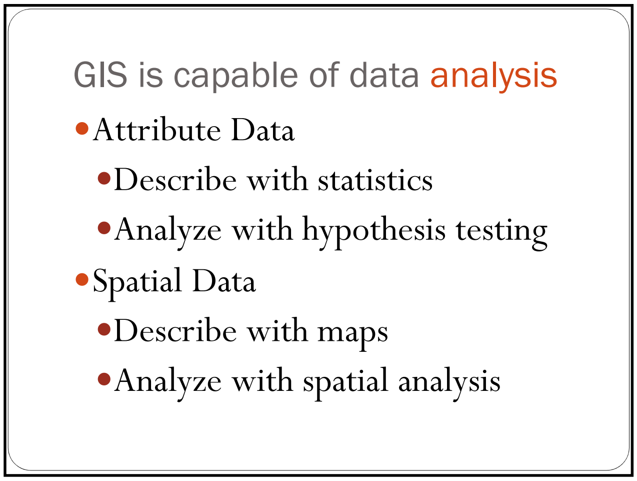
6 Why Is It There?

- 6.1 Describing Attributes
- 6.2 Statistical Analysis
- 6.3 Spatial Description
- 6.4 Spatial Analysis
- 6.5 Searching for Spatial Relationships
- 6.6 GIS and Spatial Analysis



Dueker (1979) (review)

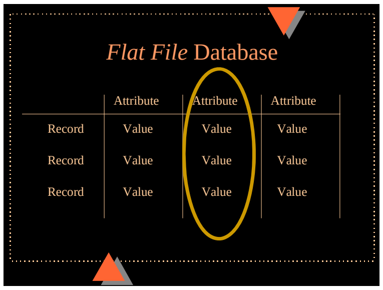
- “a geographic information system is a special case of information systems where the database consists of observations on spatially distributed features, activities or events, which are definable in space as points, lines, or areas. A geographic information system manipulates data about these points, lines, and areas to retrieve data for ad hoc queries and **analyses**".



GIS is capable of data analysis

- Attribute Data
  - Describe with statistics
  - Analyze with hypothesis testing
- Spatial Data
  - Describe with maps
  - Analyze with spatial analysis

Describing one attribute



Flat File Database

	Attribute	Attribute	Attribute
Record	Value	Value	Value
Record	Value	Value	Value
Record	Value	Value	Value

Attribute Description

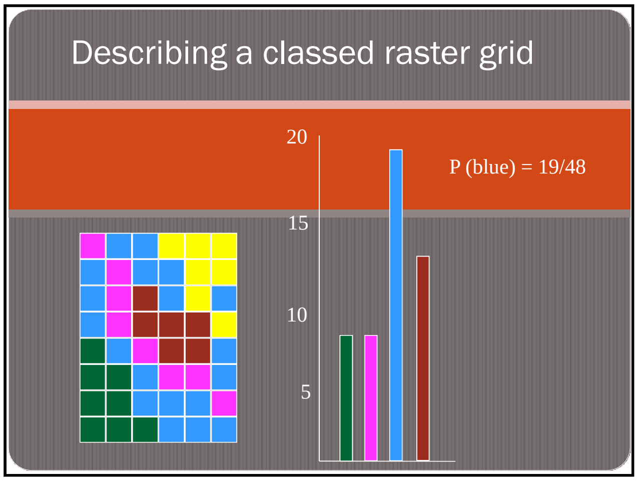
- The extremes of an attribute are the highest and lowest values, and the range is the difference between them in the units of the attribute
- A histogram is a two-dimensional plot of attribute values grouped by magnitude and the frequency of records in that group, shown as a variable-length bar
- For a large number of records with random errors in their measurement, the histogram resembles a bell curve and is symmetrical about the mean

If the records are:

- Text
  - Semantics of text e.g. “Hampton”
  - word frequency e.g. “Creek”, “Kill”
  - address matching
  - Semantics and word matching are of increasing value, e.g. web search, Metacarta
  - Ontology
- Example: Display all places called “State Street”

If the records are:

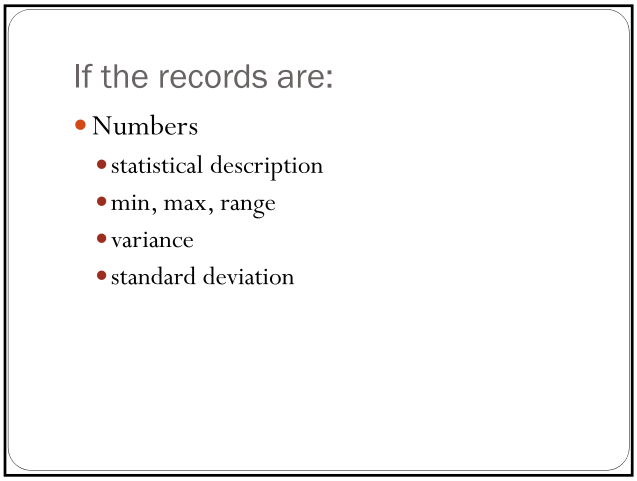
- Classes
  - histogram by class
  - numbers in class
  - contiguity description, e.g. average neighbor (roads, commercial)



Describing a classed raster grid

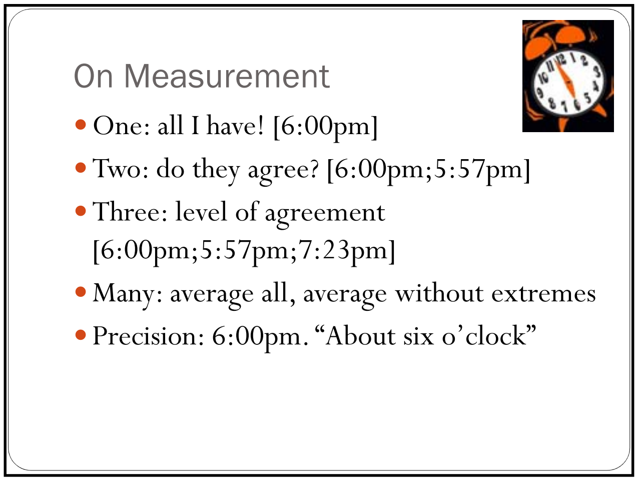
P (blue) = 19/48

🟪	🟦	🟦	🟨	🟨	🟨
🟦	🟪	🟦	🟦	🟨	🟨
🟦	🟪	🟫	🟦	🟨	🟦
🟦	🟪	🟫	🟫	🟫	🟨
🟩	🟦	🟪	🟫	🟫	🟦
🟩	🟩	🟦	🟪	🟪	🟦
🟩	🟩	🟦	🟦	🟦	🟪
🟩	🟩	🟩	🟦	🟦	🟦



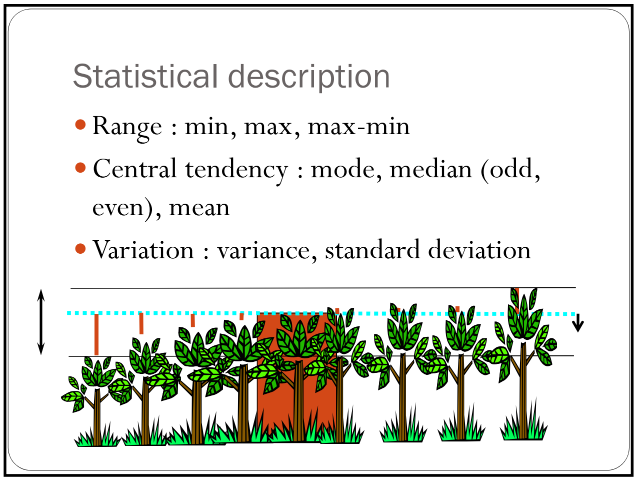
If the records are:

- Numbers
  - statistical description
  - min, max, range
  - variance
  - standard deviation



On Measurement

- One: all I have! [6:00pm]
- Two: do they agree? [6:00pm;5:57pm]
- Three: level of agreement [6:00pm;5:57pm;7:23pm]
- Many: average all, average without extremes
- Precision: 6:00pm. “About six o’clock”

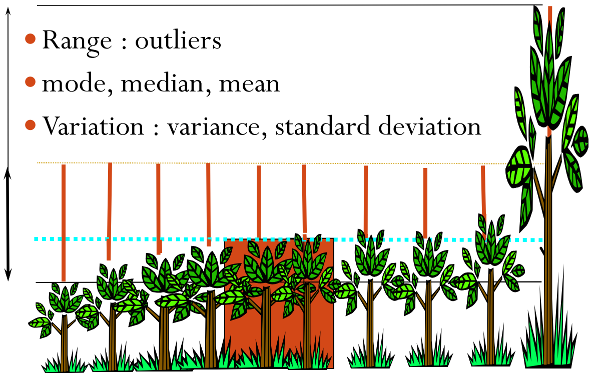


Statistical description

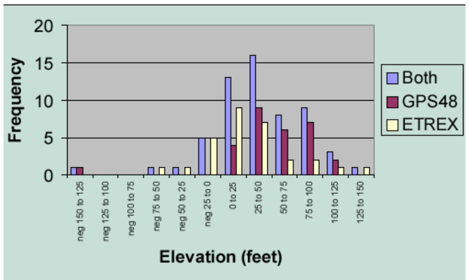
- Range : min, max, max-min
- Central tendency : mode, median (odd, even), mean
- Variation : variance, standard deviation

Statistical description

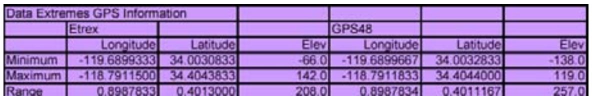
- Range : outliers
- mode, median, mean
- Variation : variance, standard deviation



Elevation (GPS example)



GPS Example Data: Elevation



Data Extremes GPS Information						
	Etrex			GPS48		
	Longitude	Latitude	Elev	Longitude	Latitude	Elev
Minimum	-119.6899333	34.0030833	-66.0	-119.6899667	34.0032833	-138.0
Maximum	-118.7911500	34.4043833	142.0	-118.7911833	34.4044000	119.0
Range	0.8987833	0.4013000	208.0	0.8987834	0.4011167	257.0

Mean

- Statistical average
- Sum of the values for one attribute divided by the number of records

$$\bar{X} = \sum_{i=1}^{n} X_i / n$$

Computing the Mean

Sum of attribute values across all records, divided by the number of records.

Add all attribute values down a column, / by # records

A representative value, and for measurements with normally distributed error, converges on the true reading.

A value lacking sufficient data for computation is called a missing value. Does not get included in sum or n.

Variance

The total variance is the sum of each record with its mean subtracted and then multiplied by itself

The standard deviation is the square root of the variance divided by the number of records less one

For two values, there is only one variance

Standard Deviation

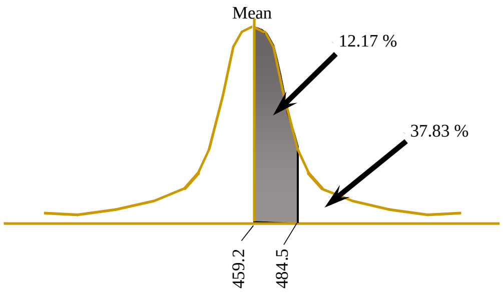
- Average difference from the mean
- Sum of the mean subtracted from the value for each record, squared, divided by the number of records-1, square rooted.

$$\text{st.dev.} = \sqrt{\frac{\Sigma\ (X_i - \overline{X}\ )^2}{n - 1}}$$

GPS Example Data: Elevation Standard deviation

- Same units as the values of the records, in this case meters.
- Average amount readings differ from the average
- Can be above of below the mean
- Elevation is the mean (459.2 meters)
- plus or minus the expected error of 82.92 meters
- Elevation is most likely to lie between 376.28 meters and 542.12 meters.
- These limits are called the error band or margin of error.

The Bell Curve



Samples and populations

- A sample is a set of measurements taken from a larger group or population.
- Sample means and variances can serve as estimates for their populations.
- Easier to measure with samples, then draw conclusions about entire population.

Testing Means

Mean elevation of 459.2 meters

standard deviation 82.92 meters

what is the chance of a GPS reading of 484.5 meters?

484.5 is 25.3 meters above the mean

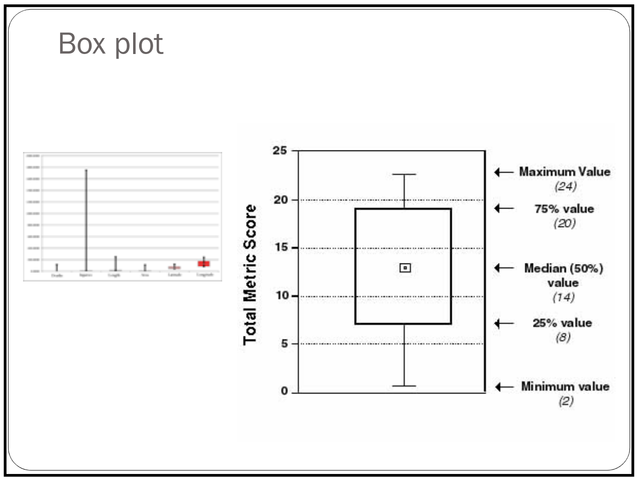
0.31 standard deviations (Z-score)

0.1217 of the curve lies between the mean and this value

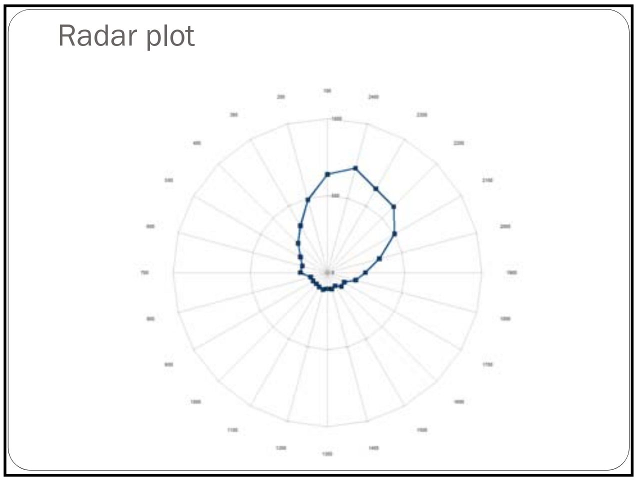
0.3783 beyond it

Data exploration

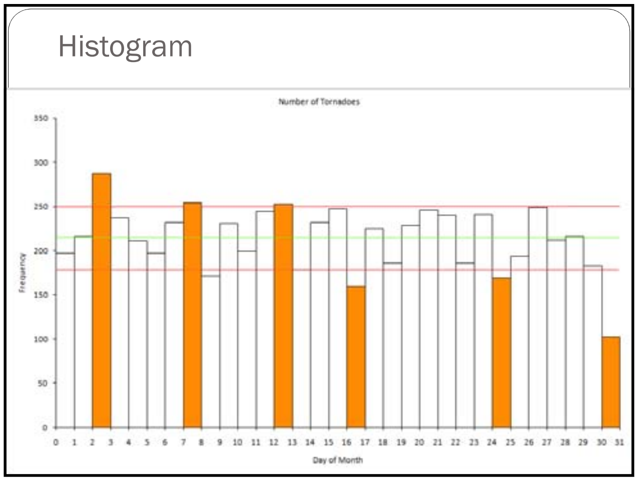
- Basic descriptive graphics can quickly summarize attribute
- Exploratory mapping: isoline, choropleth
- Some cross-variable methods e.g. bivariate choropleth
- Box plot, radar plot, histogram, scatter plot



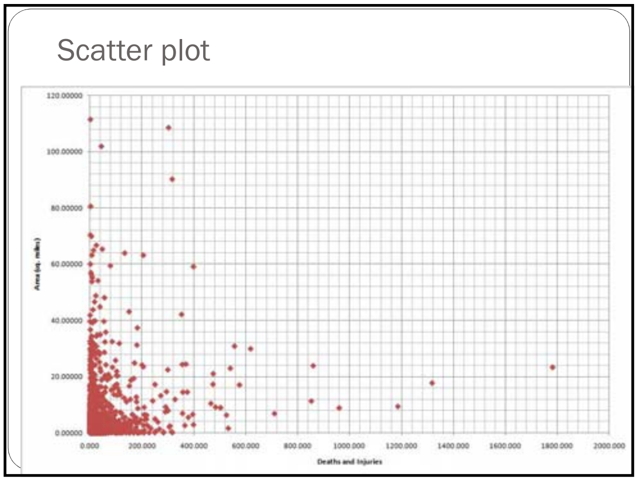
Box plot



Radar plot



Histogram



Scatter plot

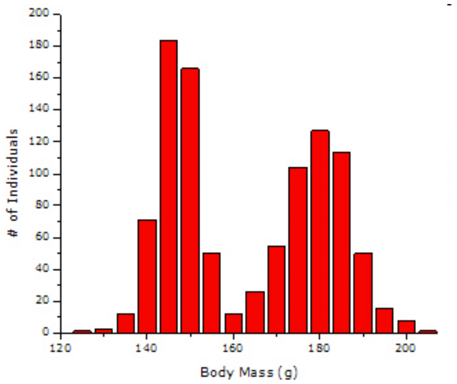
Hypothesis testing

- Set up NULL hypothesis (e.g. Values or Means are the same) as H_0
- Set up ALTERNATIVE hypothesis. H_1
- Test hypothesis. Try to reject NULL.
- If null hypothesis is rejected alternative is accepted with a calculable level of confidence.

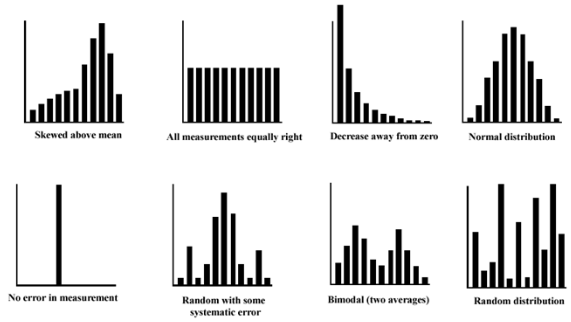
Testing the Mean

- Mathematical version of the normal distribution can be used to compute probabilities associated with measurements with known means and standard deviations.
- A test of means can establish whether two samples from a population are different from each other, or whether the different measures they have are the result of random variation.

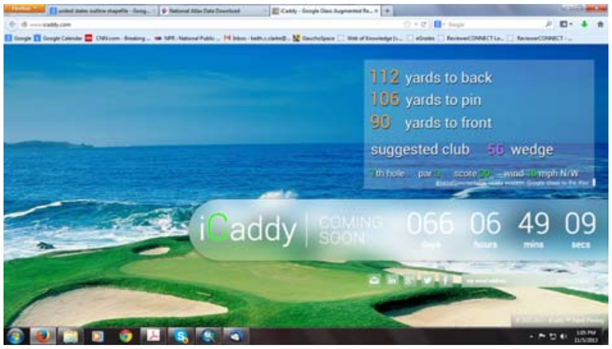
One distribution or two?



Alternative attribute histograms



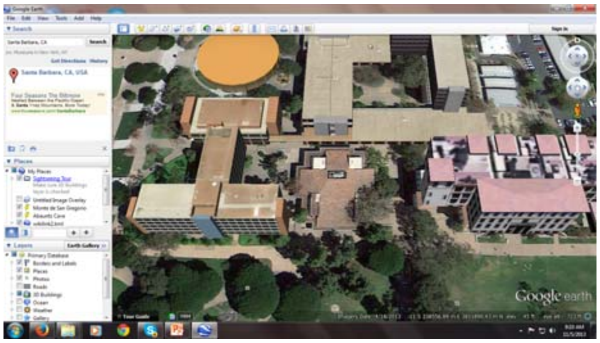
G176A in Action: www.icaddy.com



Accuracy

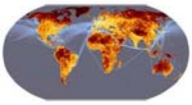
- Determined by testing measurements against an independent source of higher fidelity and reliability.
- Must pay attention to units and significant digits.
- Can be expressed as a number using statistics (e.g. expected error).
- Accuracy measures imply accuracy users.

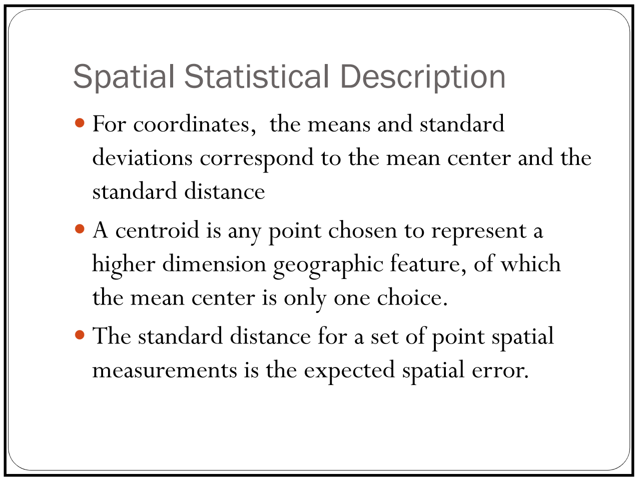
How accurate? 0.01mm?



The difference is the map

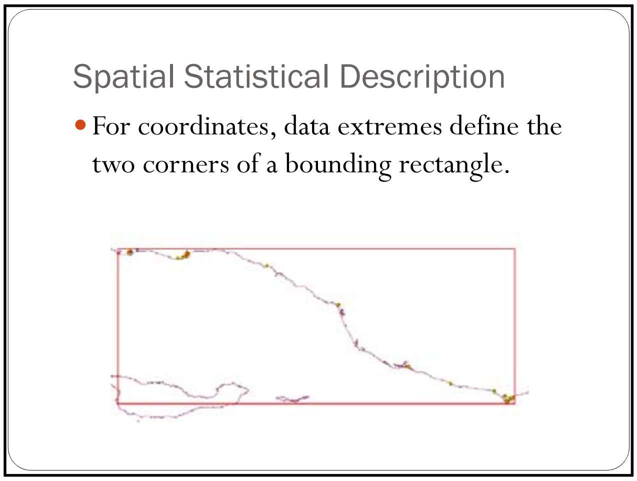
- GIS data description answers the question: Where?
- GIS data analysis answers the question: Why is it there?
- GIS data description is different from statistics because the results can be placed onto a map for visual analysis.





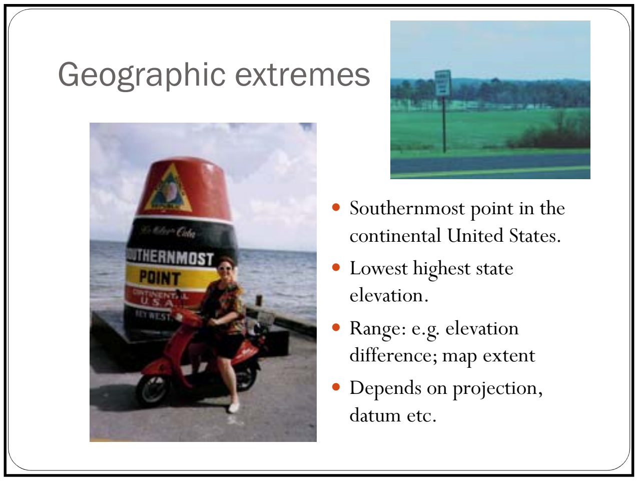
Spatial Statistical Description

- For coordinates, the means and standard deviations correspond to the mean center and the standard distance
- A centroid is any point chosen to represent a higher dimension geographic feature, of which the mean center is only one choice.
- The standard distance for a set of point spatial measurements is the expected spatial error.



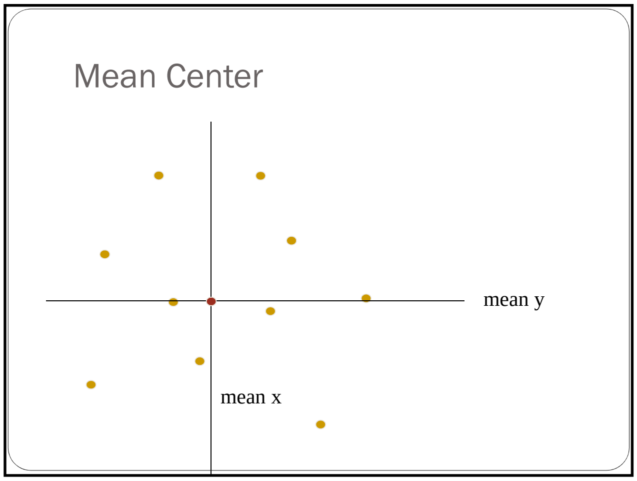
Spatial Statistical Description

- For coordinates, data extremes define the two corners of a bounding rectangle.



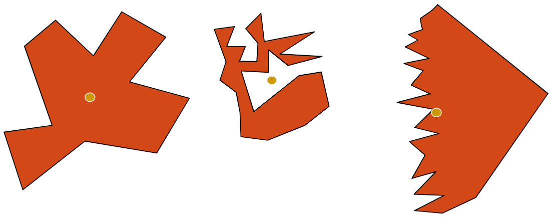
Geographic extremes

- Southernmost point in the continental United States.
- Lowest highest state elevation.
- Range: e.g. elevation difference; map extent
- Depends on projection, datum etc.

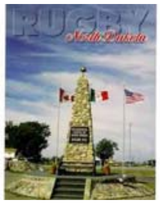


Mean Center

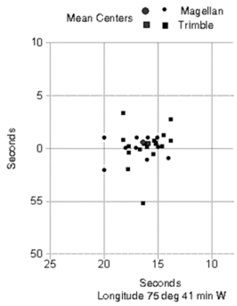
Centroid: mean center of a feature

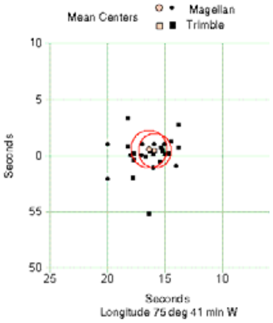


Mean center?



Comparing spatial means





GIS and Spatial Analysis

- Descriptions of geographic properties such as shape, pattern, and distribution are often verbal
- Quantitative measure can be devised, although few are computed by GIS.
- GIS statistical computations are most often done using retrieval options such as buffer and spread.
- Also by manipulating attributes with arithmetic commands (map algebra).

Example: Intervisibility

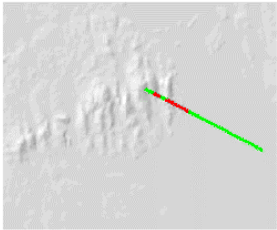


Figure 1a. Line-of-Sight Profile

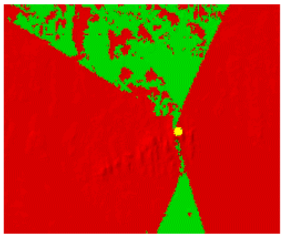
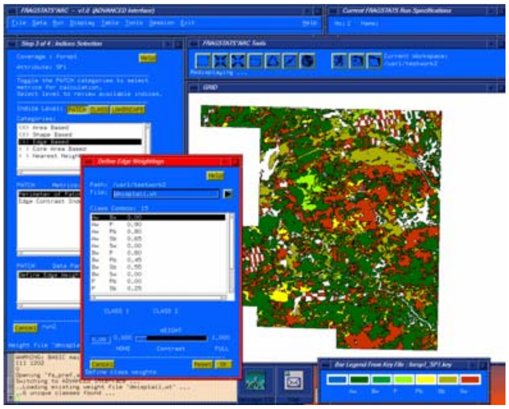


Figure 1b. Masked Area Plot

Source: Mineter, Dowers, Gittings, Caldwell ESRI Proceedings

Example: Landscape Metrics


FRAGSTATS*ARC – v1.0 (ADVANCED Interface)
File Data Run Display Table Tools Session Exit
Current FRAGSTATS Run Specifications
Step 3 of 4 : Indices Selection
FRAGSTATS*ARC Tools
GRID
Define Edge Weightings
Bar Legend From Key File

Example NNS

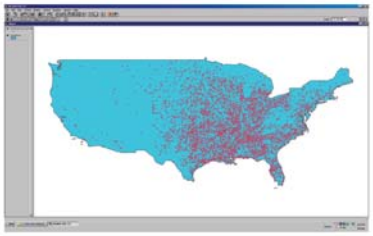
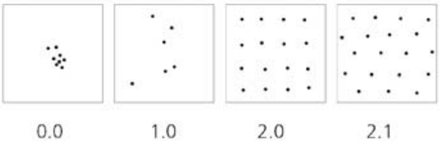


Figure 6.15: ArcView 3.2 with the nearest neighbor script loaded. The script requires a polygon and a point layer, and computes R and other values. For the tornadoes leading to deaths or injuries shown, R=0.722.



A textbook example

- Lower 48 United States
- 1996 Data from the U.S. Census on gender
- Gender Ratio = # females per 100 males
- Range is 96.4 to 114.4
- What does the spatial distribution look like?

Gender Ratio by State: 1996

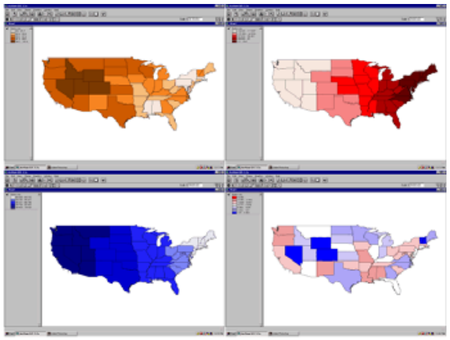
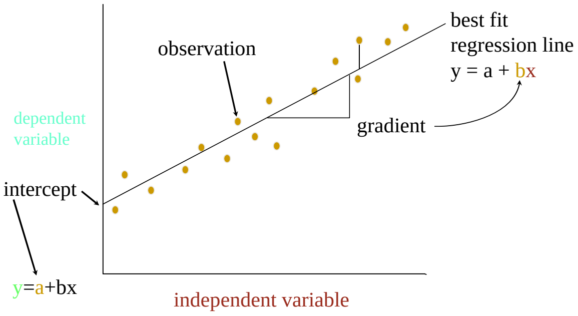


Figure 6.8 Data for the Gender Ratios example plotted on the USA lower-48 states base map. Upper left: the gender ratio data (2000 Census, Male to Female ratio). Upper right: Longitude of state capitals. Lower Left: The linear model. Lower Right: residuals from the model.

Searching for Spatial Pattern

- A linear relationship is a predictable straight-line link between the values of a dependent and an independent variable. (y = a + bx) It is a simple model of the relationship.
- A linear relation can be tested for goodness of fit with least squares methods. The coefficient of determination r-squared is a measure of the degree of fit, and the amount of variance explained.

Simple linear relationship



observation

best fit
regression line
y = a + bx

dependent
variable

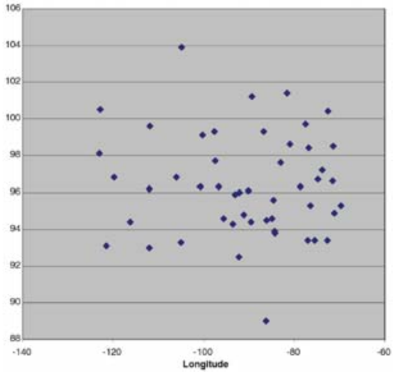
gradient

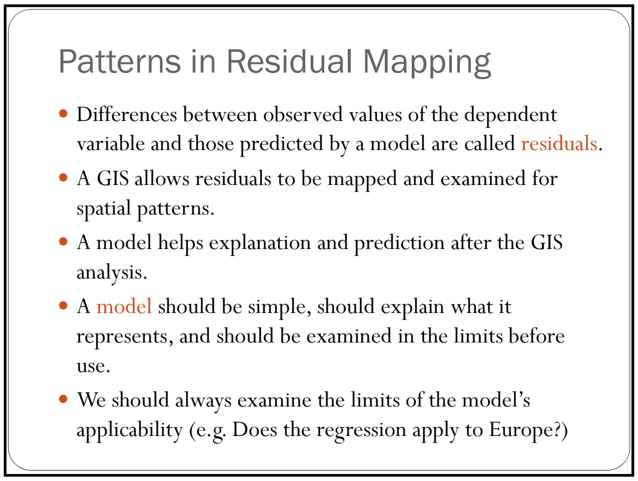
intercept

y=a+bx

independent variable

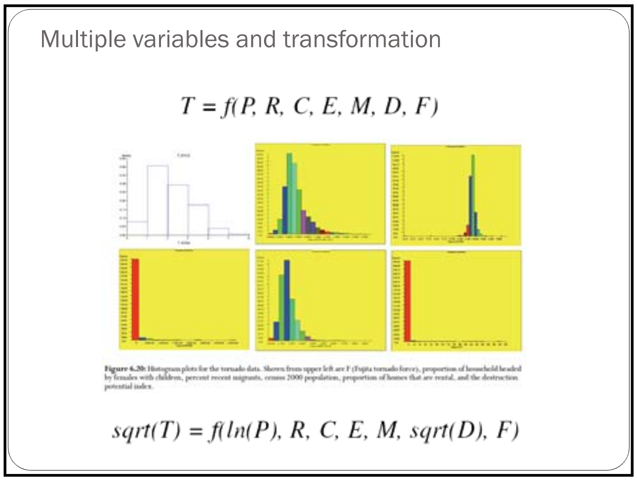
Testing the relationship





Patterns in Residual Mapping

- Differences between observed values of the dependent variable and those predicted by a model are called residuals.
- A GIS allows residuals to be mapped and examined for spatial patterns.
- A model helps explanation and prediction after the GIS analysis.
- A model should be simple, should explain what it represents, and should be examined in the limits before use.
- We should always examine the limits of the model's applicability (e.g. Does the regression apply to Europe?)

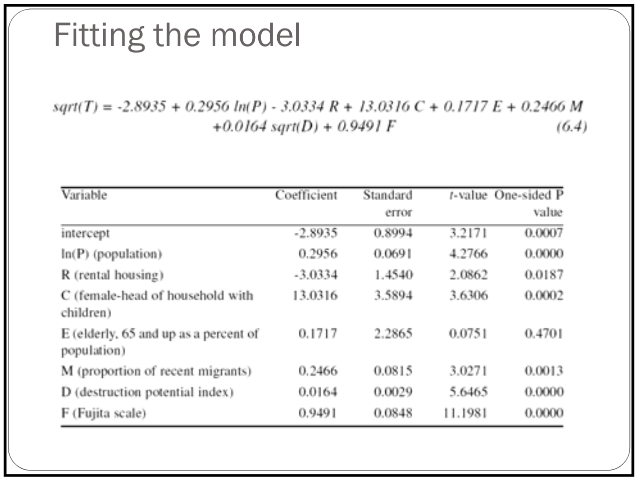


Multiple variables and transformation

$$T = f(P, R, C, E, M, D, F)$$

Figure 6.20: Histogram plots for the tornado data. Shown from upper left are F (Fujita tornado force), proportion of household headed by females with children, percent recent migrants, census 2000 population, proportion of homes that are rental, and the destruction potential index.

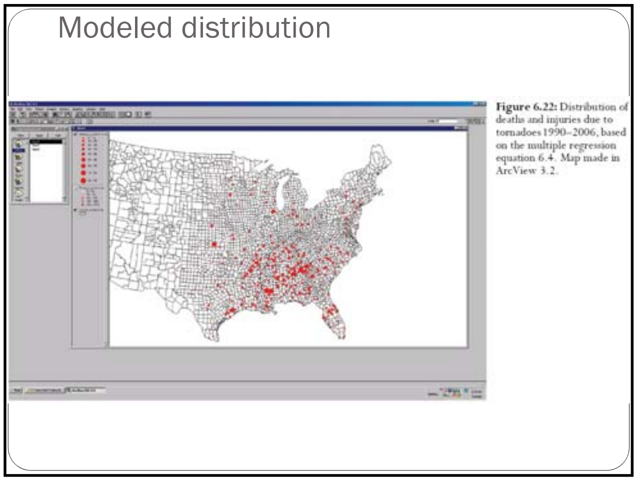
$$sqrt(T) = f(ln(P), R, C, E, M, sqrt(D), F)$$



Fitting the model

$$sqrt(T) = -2.8935 + 0.2956\ ln(P) - 3.0334\ R + 13.0316\ C + 0.1717\ E + 0.2466\ M + 0.0164\ sqrt(D) + 0.9491\ F \qquad (6.4)$$

Variable	Coefficient	Standard error	*t*-value	One-sided P value
intercept	-2.8935	0.8994	3.2171	0.0007
ln(P) (population)	0.2956	0.0691	4.2766	0.0000
R (rental housing)	-3.0334	1.4540	2.0862	0.0187
C (female-head of household with children)	13.0316	3.5894	3.6306	0.0002
E (elderly, 65 and up as a percent of population)	0.1717	2.2865	0.0751	0.4701
M (proportion of recent migrants)	0.2466	0.0815	3.0271	0.0013
D (destruction potential index)	0.0164	0.0029	5.6465	0.0000
F (Fujita scale)	0.9491	0.0848	11.1981	0.0000



Modeled distribution

Figure 6.22: Distribution of deaths and injuries due to tornadoes 1990–2006, based on the multiple regression equation 6.4. Map made in ArcView 3.2.

Residuals

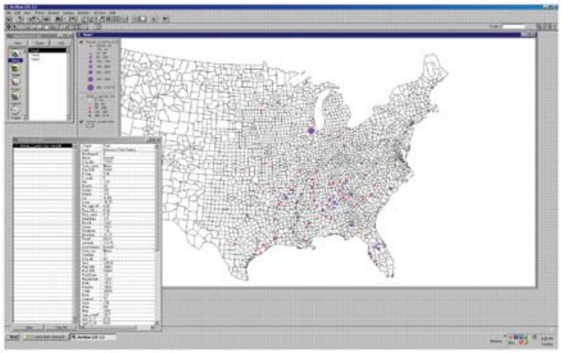
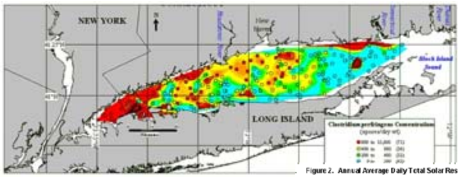


Figure 6.24: Map of regression residuals for the tornado deaths and injuries multiple regression. Highlighted case is for Kendall, Illinois, with 29 deaths and 350 injuries on 8/28/1990, underestimated significantly by the regression model.

What About the Unexplained variance?

- More variables?
- Leave out outliers?
- Different extent?
- More records?
- More spatial dimensions?
- Different methods? e.g. GWR
- More complexity?
- Another model?
- Another approach?

Spatial Interpolation



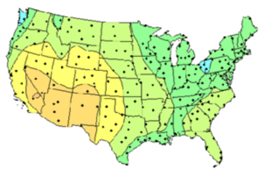
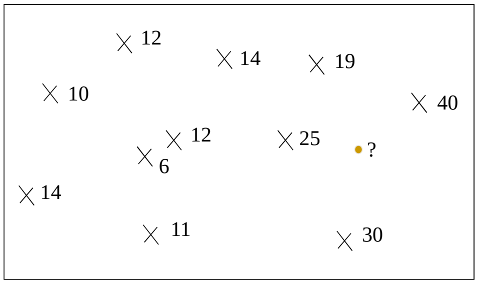


Figure 2. Annual Average Daily Total Solar Resources
Two-Axis Tracking Concentrator

http://www.eia.doe.gov/cneaf/solar.renewables/rea_issues/html/fig2ntrans.gif

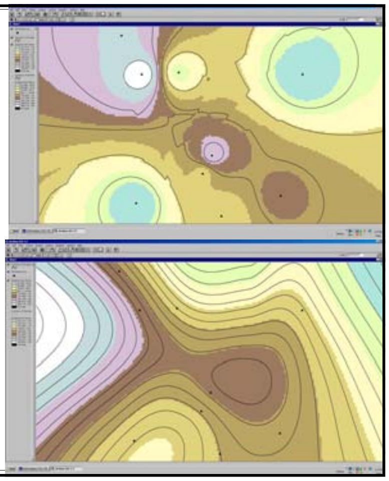
Issues: Spatial Interpolation



meters to water table

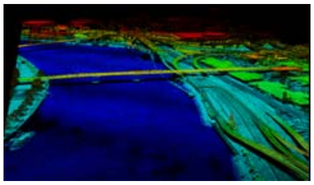
resolution? extent? accuracy? precision?
boundary effects? point spacing? Method?

Interpolation method matters
Top: IDW
Bottom: Spline



Extrapolation

- How far does model extend?
- How fixed in time is the model?
- Data rich vs. data poor areas
- What if there are too many data points?



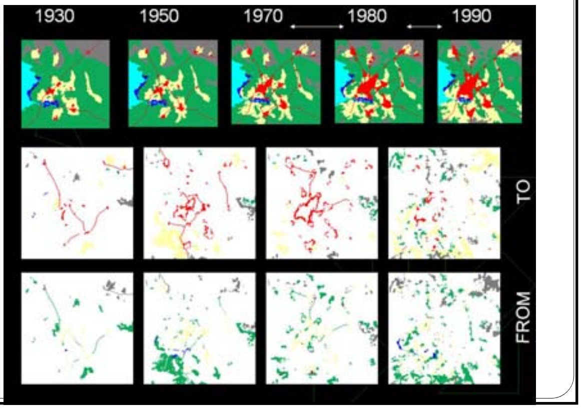
GIS and Spatial Analysis

- Geographic inquiry examines the relationships between geographic features collectively to help describe and understand the real-world phenomena that the map represents.
- Spatial analysis compares maps, investigates variation over space, and predicts future or unknown maps.
- Many GIS systems have to be coaxed to generate a full set of spatial statistics.

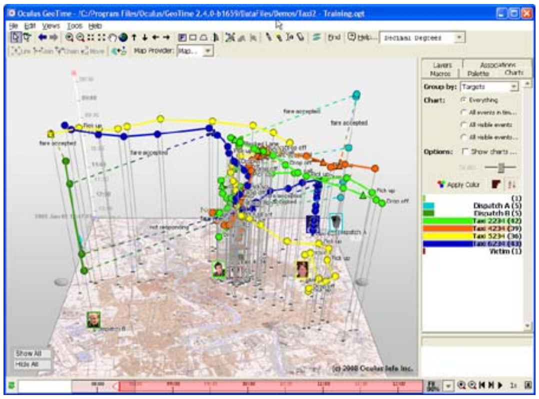
Analytic Tools and GIS

- Tools for searching out spatial relationships and for modeling are only lately being integrated into GIS.
- Statistical and spatial analytical tools are also only now being integrated into GIS, and many people use separate software systems outside the GIS (such as R, MATLAB, SPSS, SAS, GeoDA)
- Real geographic phenomena are dynamic, but GISs (and analyses) have been mostly static. Time-slice and animation methods can help in visualizing and analyzing spatial trends
- GIS places real-world data into an organizational framework that allows numerical description and allows the analyst to model, analyze, and predict with both the map and the attribute data

For example LULCC



Oculus: Geotime



You can lie with…

- Maps
- Statistics
  - Correlation is not causation!
  - Hypothesis vs. Action

Coming next …

- Terrain Analysis