

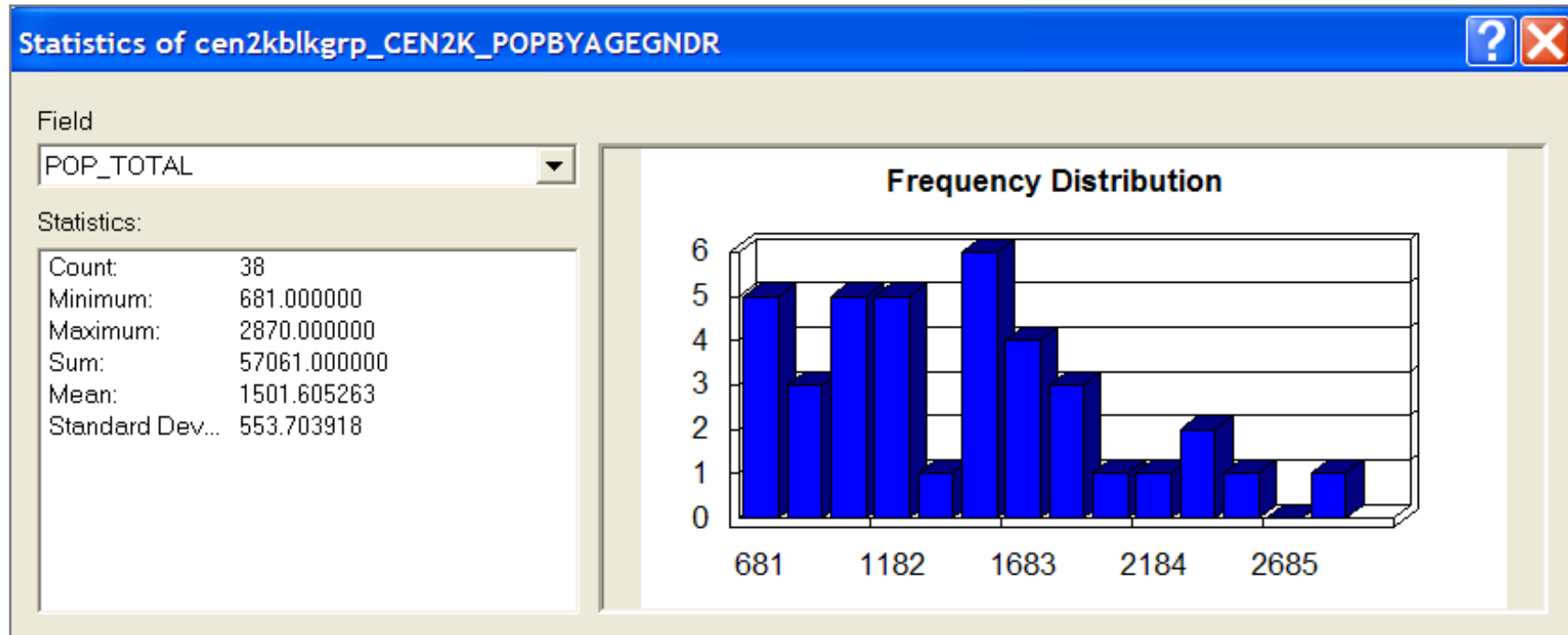
Chapter 6: Why is it There?

- 6.1 Describing Attributes
- 6.2 Statistical Analysis
- 6.3 Spatial Description
- 6.4 Spatial Analysis

GIS is Capable of Data Analysis

- Attribute Data
 - **Describe** with **statistics**
- Spatial Data
 - **Describe** with **maps**
 - **Analyze** with **spatial analysis**


Describe with Statistics



Scales of Measurement

- **Attribute data** can be divided into four types

1. The Nominal Scale
2. The Ordinal Scale
3. The Interval Scale
4. The Ratio Scale

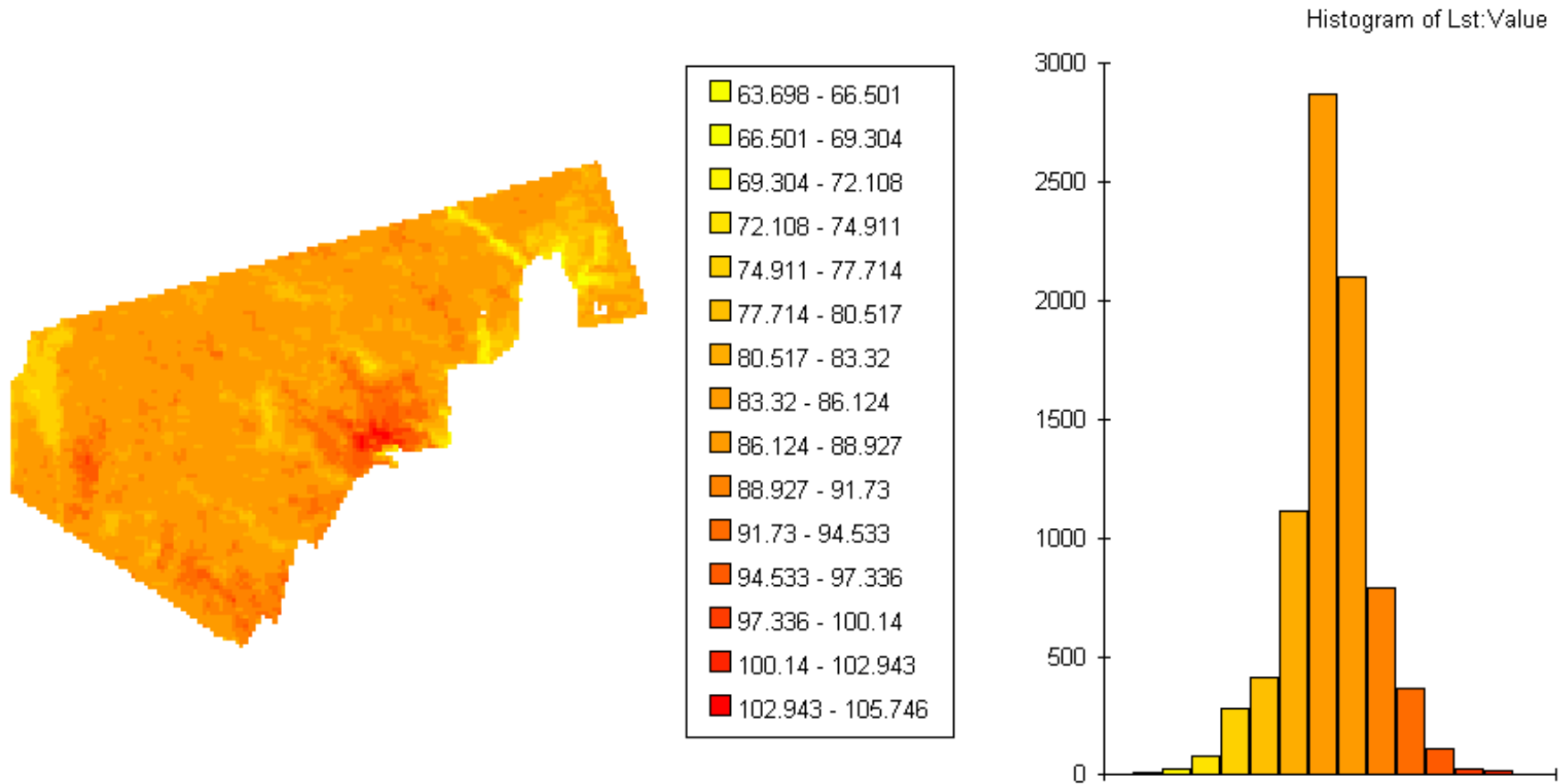


As we progress through these scales, the types of data they describe have increasing information content

Ordinal, Interval, & Ratio Attributes

- The extremes (**min** and **max**) of an attribute are the **highest and lowest values**, and the **range** is the **difference between them** in the units of the attribute.
- The **count** is the **number of records**.
- The **sum** is all the numeric values of an attribute **added together**.
- 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.

LST Distribution Example

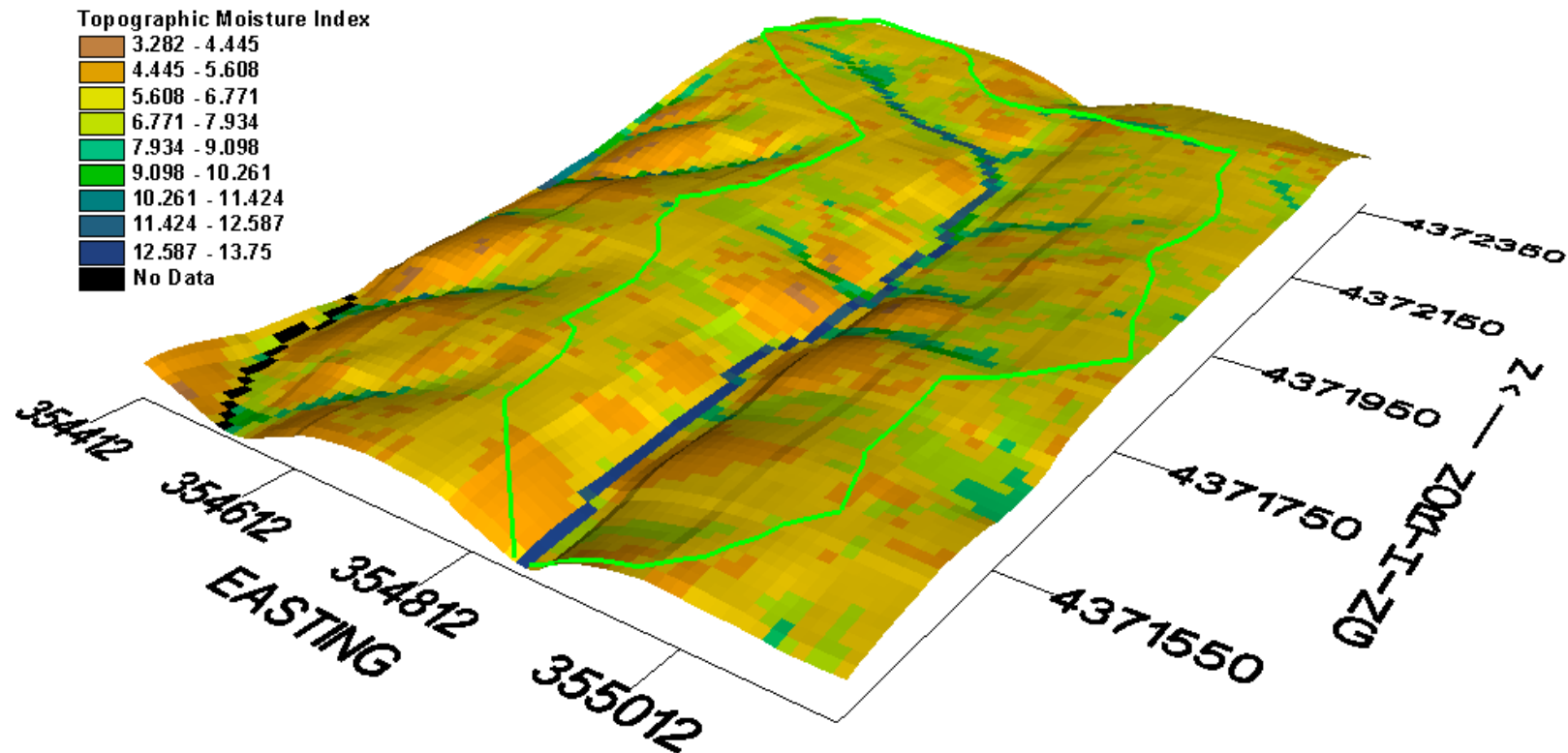


Mean = 85.4994 degrees F

Standard Deviation = 4.3092 degrees F

Ordinal, Interval, & Ratio Attributes

- Pond Branch is a 37.55 hectare watershed, which is equivalent to 375,500 m² (1 hectare = 10,000 m²)
- Using 11.25m x 11.25m pixels (126.5625 m²), there are ~ 2966 pixels from which we can draw TMI values



Ordinal, Interval, & Ratio Attributes

- It would clearly be **impractical** to try and get a sense of the distribution of TMI values in Pond Branch by looking at a table of 2966 values
- We need a data reduction approach by which we can **reduce** the number of values to a **manageable amount**, which in turn lends itself to some sort of graphical display
- For ordinal, interval, and ratio scale data, we can make use of **histograms** for this purpose, and building a histogram involves following a multi-step procedure ...

Building a Histogram

1. Developing an ungrouped frequency table

- That is, we **build a table** that **counts the number of occurrences** of each variable value from lowest to highest:

TMI Value

Ungrouped Freq.

4.16

2

4.17

4

4.18

0

...

...

13.71

1

•We could attempt to construct a bar chart from this table, but it would have too many bars to really be useful

Building a Histogram

2. Construct a grouped frequency table

- This table has **classes** of values (in a sense we are reducing our data back to the ordinal scale for display purposes)
- The decision on **how to perform the grouping** is a subjective one, but there are some common guidelines:
- Use class intervals with **simple bounds** and a **common width** (i.e. categories have same range)
- Adjacent intervals **should not overlap** (each value should fit into one class)

Building a Histogram

3. Select an appropriate number of classes

- There are formulae available to make this decision **objectively**, but in reality it is a somewhat subjective decision
- If you have **more observations**, you usually need **more classes**, because when you put observations together in a class, you are considering them to have the **same value for display purposes** → there is a **trade-off** here between **simplicity** and **loss of information**

Building a Histogram

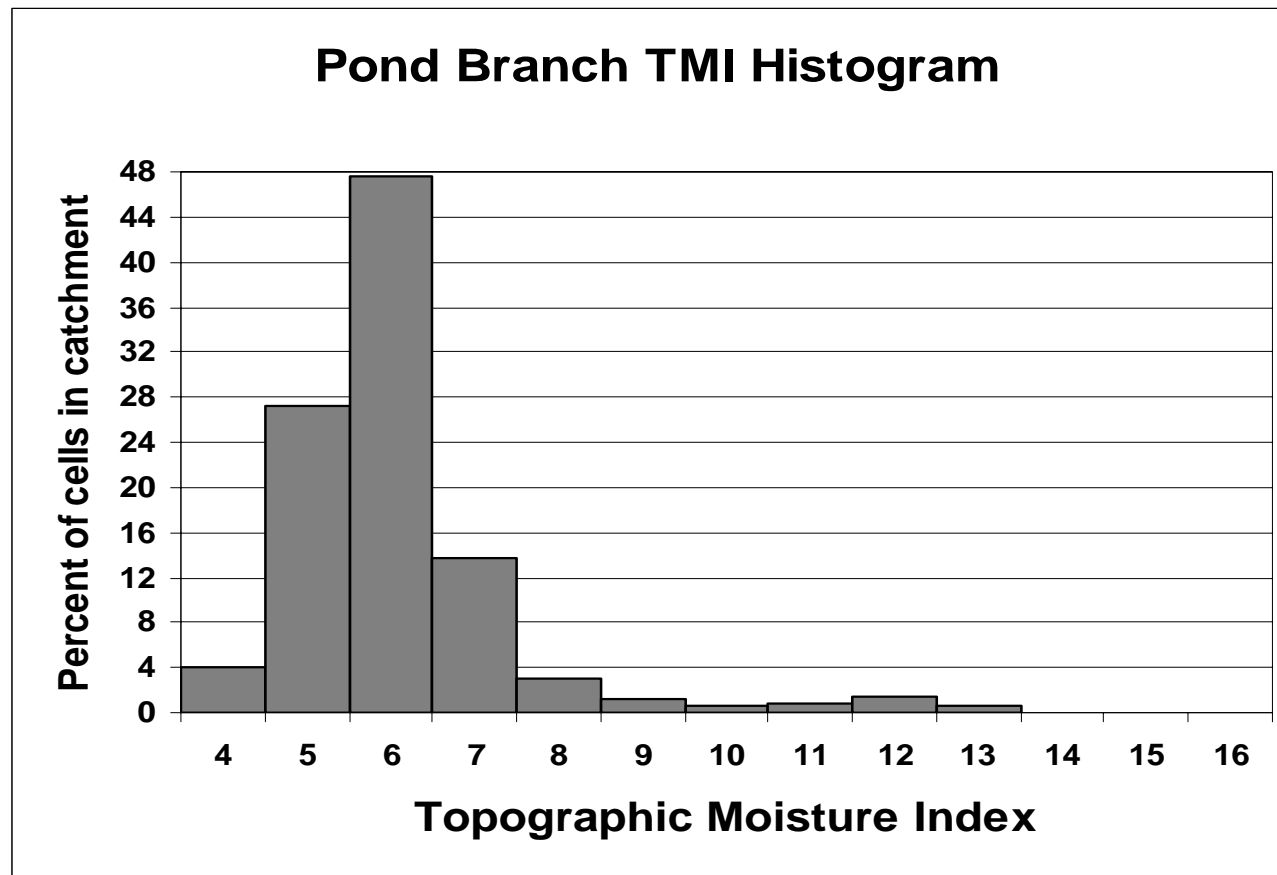
3. Select an appropriate number of classes cont.

<u>Class</u>	<u>Frequency</u>
4.00 - 4.99	120
5.00 - 5.99	807
6.00 - 6.99	1411
7.00 - 7.99	407
8.00 - 8.99	87
9.00 - 9.99	33
10.00 - 10.99	17
11.00 - 11.99	22
12.00 - 12.99	43
13.00 - 13.99	19

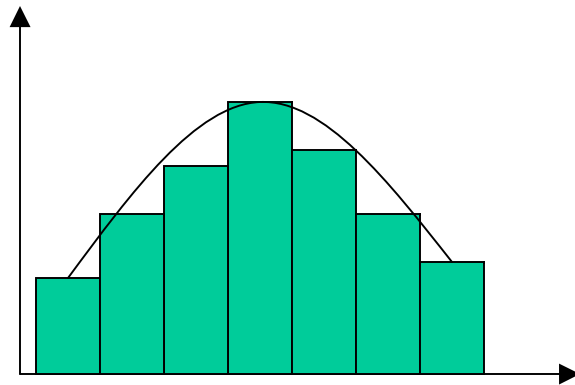
Building a Histogram

4. Plot the frequencies of each class

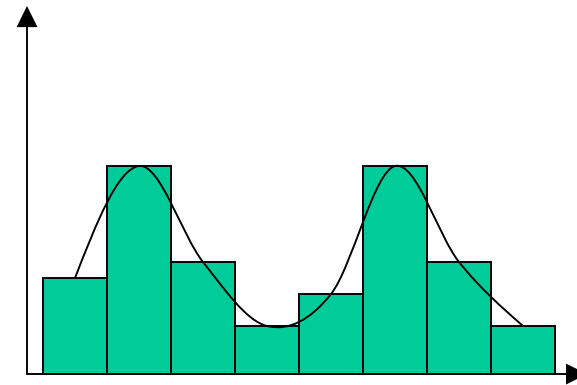
- All that remains is to create the plot:



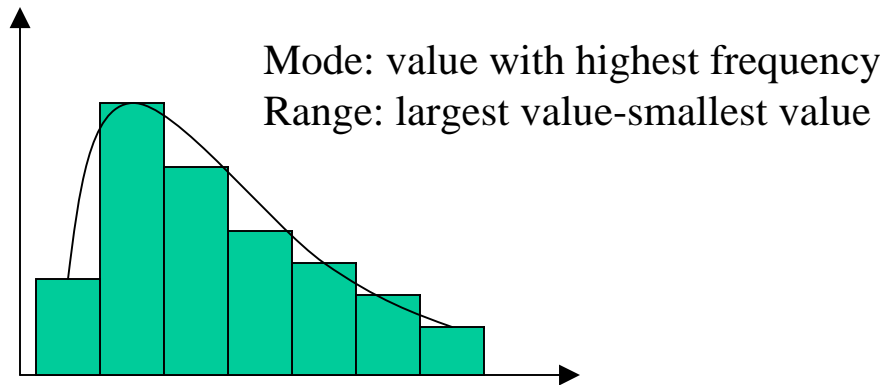
Shapes of Histograms



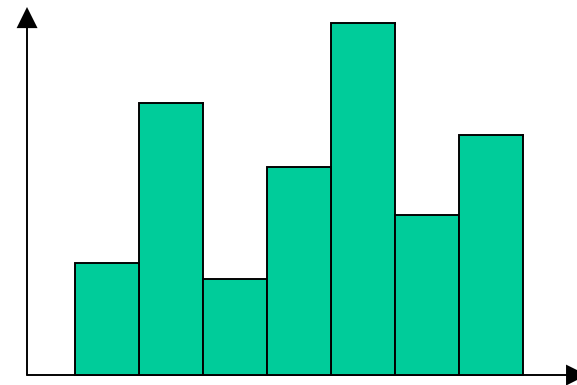
Bell Shaped



Bimodal



Skewed



Random

- Developing a histogram from attribute data is one level of data reduction; we can describe bell shaped distributions using parameters that provide a more concise summary

Measures of Central Tendency

- Think of this from the following point of view:
We have some distribution in which we want to **locate the center**, and we need to choose an appropriate measure of central tendency. We can choose from:
 1. Mode
 2. Median
 3. Mean
- Each of these measures is appropriate to different distributions / under different circumstances

Measures of Central Tendency - Mode

- 1. Mode** – This is the most frequently occurring value in the distribution
 - In the event that **multiple values** tie for the **highest frequency**, we have a **problem** ...
 - A potential **solution** in this situation involves constructing **frequency classes** and identify the **most frequently occurring** class
 - This is the only measure of central tendency that can be used with **nominal data**
 - The mode allows the distribution's peak to be located quickly

Measures of Central Tendency - Median

- 2. Median** – This is the value of a variable such that **half** of the observations **are above** and **half are below** this value i.e. this value divides the distribution into two groups of equal size
- Note: When the distribution has an **even number** of observations, finding the median requires **averaging two numbers**
 - The **key advantage** of the median is that its value is **unaffected** by extreme values at the end of a distribution (which potentially are **outliers**)

Measures of Central Tendency - Mean

3. Mean – a.k.a. average, the most commonly used measure of central tendency

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n}$$

Sample mean

- When we compute a mean using these basic formulae, we are assuming that each observation is equally significant

Measures of Central Tendency - Mean

3. Mean cont. – We can also calculate a **weighted mean** using some weighting factor:

$$\bar{x} = \frac{\sum_{i=1}^{i=n} w_i x_i}{\sum_{i=1}^{i=n} w_i}$$

Weighted mean

e.g. What is the average income of all people in cities A, B, and C:

<u>City</u>	<u>Avg. Income</u>	<u>Population</u>
A	\$23,000	100,000
B	\$20,000	50,000
C	\$25,000	150,000

Here, population is the weighting factor w_i and the average income is the variable of interest x_i

Measures of Central Tendency - Mean

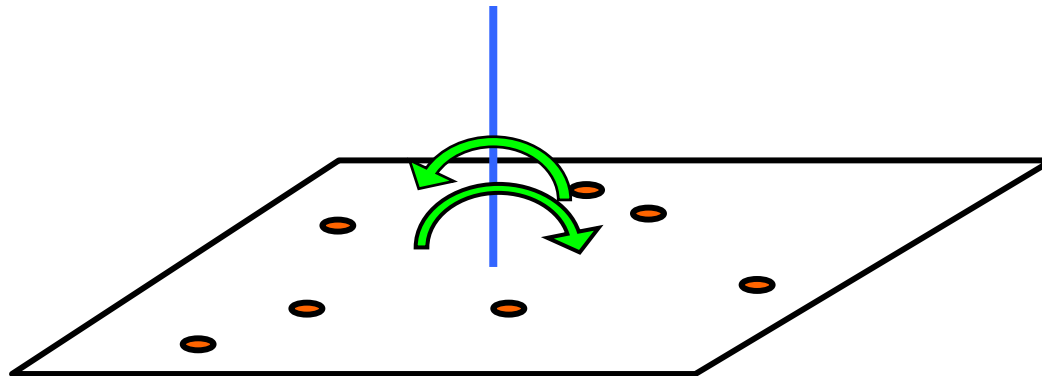
3. Mean cont. – A standard geographic application of the mean is to locate the center (a.k.a. **centroid**) of a **spatial distribution** by assigning to each member of the spatial distribution a gridded coordinate and calculating the mean value in each coordinate direction → **Bivariate mean** or **mean center**

For a set of (x,y) coordinates, the **mean center** (\bar{x}, \bar{y}) is computed using:

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} \quad \bar{y} = \frac{\sum_{i=1}^{i=n} y_i}{n}$$

The Centroid

- The **mean center** can be found for a set of points by taking an **average of coordinates**, and this is also known as a **centroid**
- The centroid can also be thought of as the **balance point** of a set of points, as it **minimizes** the sum of the distances squared



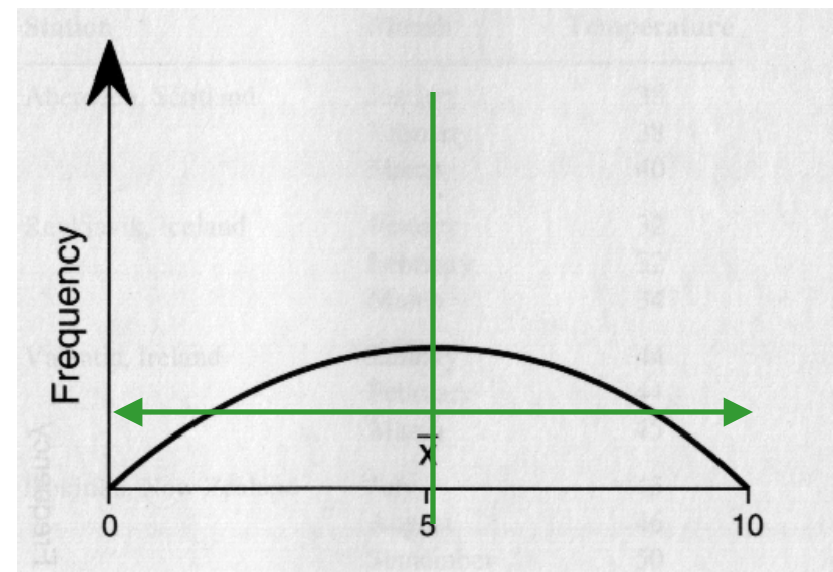
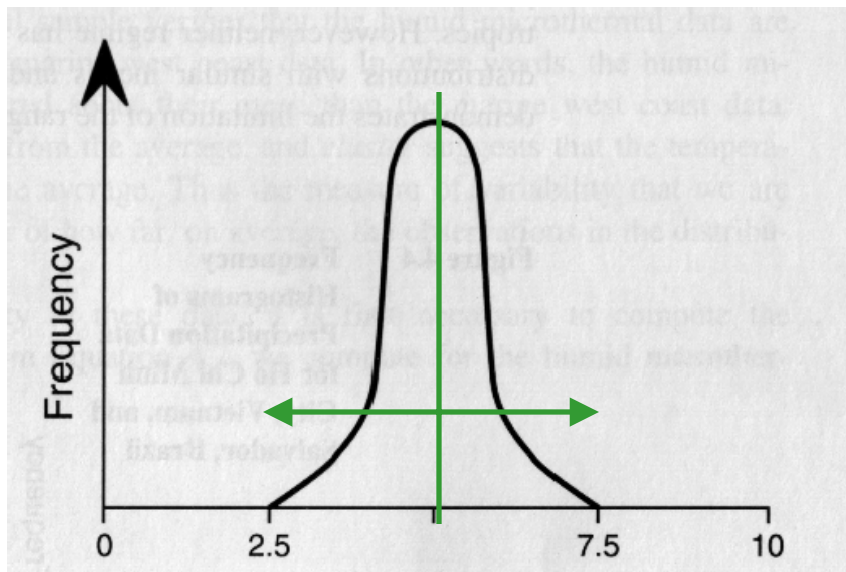
Measures of Central Tendency - Mean

3. Mean cont. – NOTE: Mean centers are very sensitive to outliers! For example:

- Suppose we calculated the **geographic center** of the **United States** when considering **all** states.
- Consider the **influence** that **Hawaii and Alaska** would have on the result if we chose to include them in the calculation.
- Clearly we would get a **very different result** with and without these two states

Why Do We Need Measures of Dispersion at all?

- Measures of central tendency **tell us nothing** about the variability / dispersion / deviation / range of values about the central value. Consider the following two unimodal symmetric distributions:



Source: Earickson, RJ, and Harlin, JM. 1994. Geographic Measurement and Quantitative Analysis. USA: Macmillan College Publishing Co., p. 91.

Measures of Dispersion - Range

1. **Range** – this is the most **simply formulated** of all measures of dispersion
 - Given a set of measurements $x_1, x_2, x_3, \dots, x_{n-1}, x_n$, the range is defined as the **difference** between the largest and smallest values:

$$\text{Range} = x_{max} - x_{min}$$

- This is another descriptive measure that is **vulnerable** to the influence of **outliers** in a data set, which result in a range that is not really descriptive of most of the data

Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. **Variance etc.** – As an alternative to taking the absolute values of the statistical distances, we can square each deviation before taking their sum, which yields the **sum of squares**:

$$\text{Sum of Squares} = \sum_{i=1}^{i=n} (x_i - \bar{x})^2$$

- The sum of squares expresses the **total square variation about the mean**, and using this value we can calculate variances and standard deviations for both populations and samples

Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. Variance etc. cont. – Variance is formulated as the sum of squares divided by the population size or the sample size minus one:

$$S^2 = \frac{\sum_{i=1}^{i=N} (x_i - \bar{x})^2}{n - 1}$$

Sample variance

Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. **Variance etc. cont.** – Standard deviation is calculated by taking the square root of variance:

$$S = \sqrt{\frac{\sum_{i=1}^{i=N} (x_i - \bar{x})^2}{n - 1}}$$

Sample standard deviation

- Why do we **prefer** standard deviation over variance as a measure of dispersion? Magnitude of values and units match means

Measures of Dispersion – Variance, Standard Deviation, Z-scores

2. **Variance etc. cont.** – Just as the mean can be applied to spatial distributions through the bivariate mean center and weighted mean center formulae (computed by considering the (x,y) coordinates of a set of spatial objects), standard deviation can be applied to examining the dispersion of a spatial distribution. This is called **standard distance** (SD):

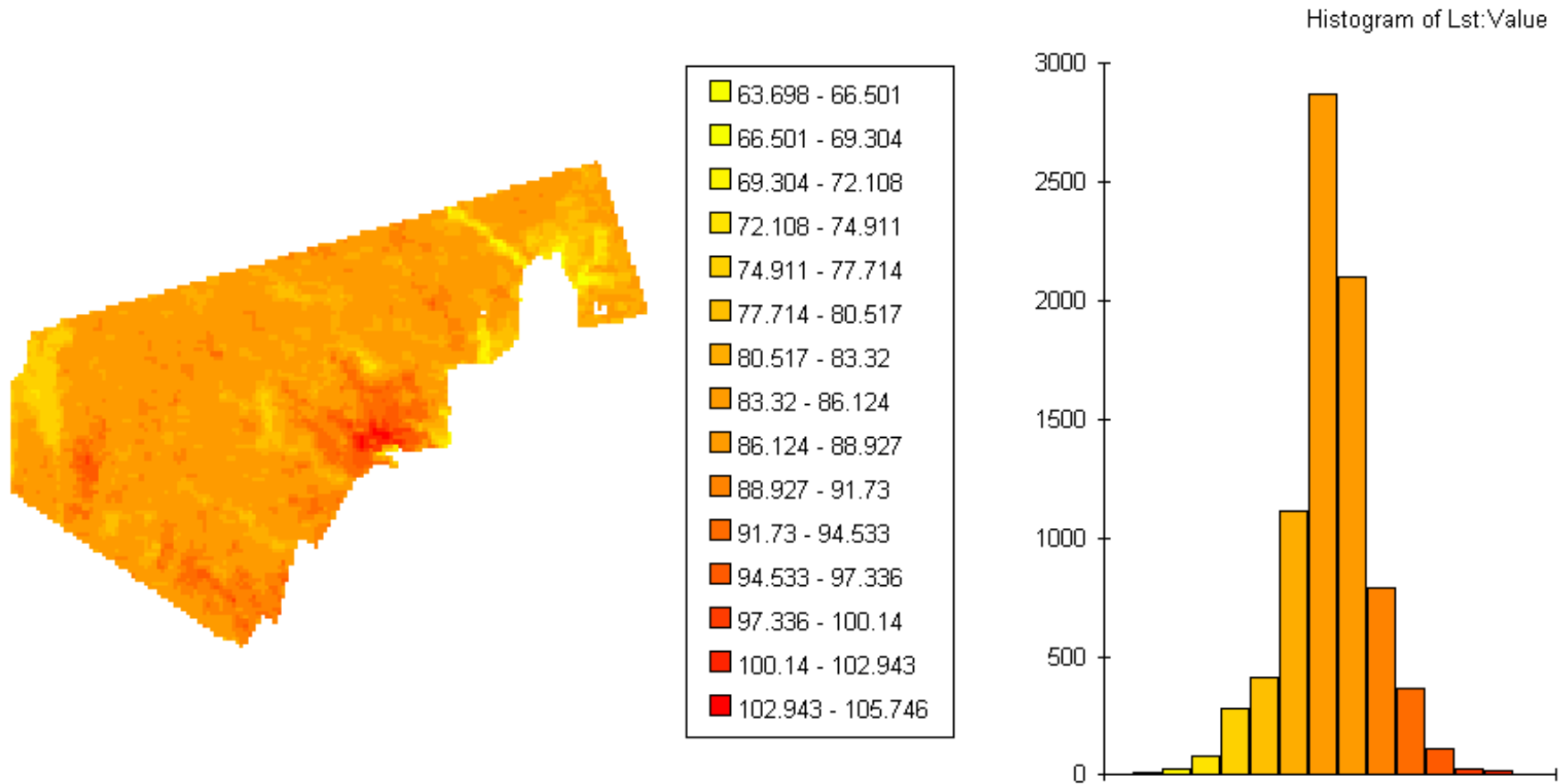
$$SD = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}{n - 1} + \frac{\sum_{i=1}^{i=n} (y_i - \bar{y})^2}{n - 1}}$$

Measures of Dispersion – Variance, Standard Deviation, Z-scores

- 2. Variance etc. cont.** – Sometimes, we want to **compare** data from different distributions, which in turn have different means and variances
- In these circumstances, it's convenient to have a **standardized** measure of dispersion that can be calculated for an individual observation. The **z-score** (a.k.a. standard normal variate, standard normal deviate, or just the standard score) is calculated by subtracting the sample mean from the observation, and then dividing that difference by the sample standard deviation:

$$\text{Z-score} = \frac{x - \bar{x}}{S}$$

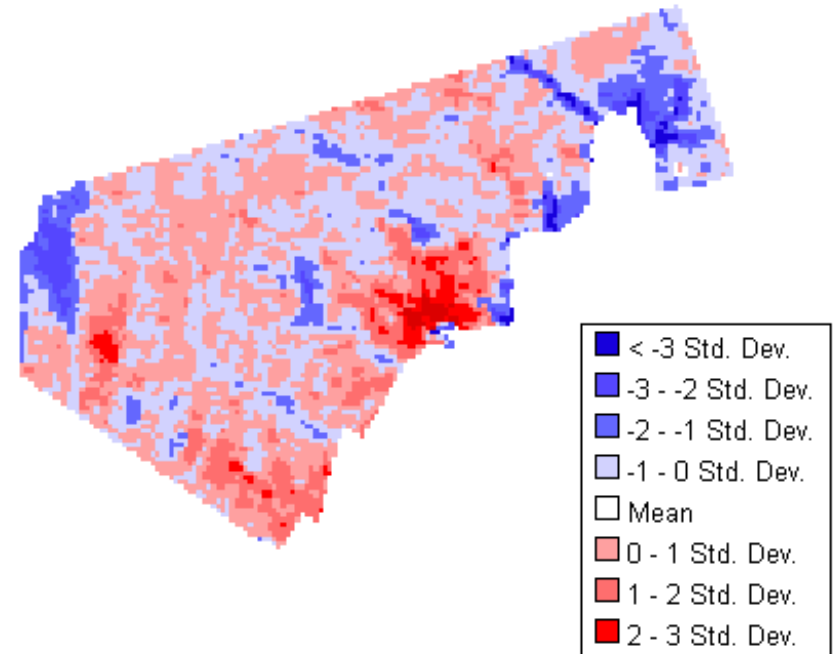
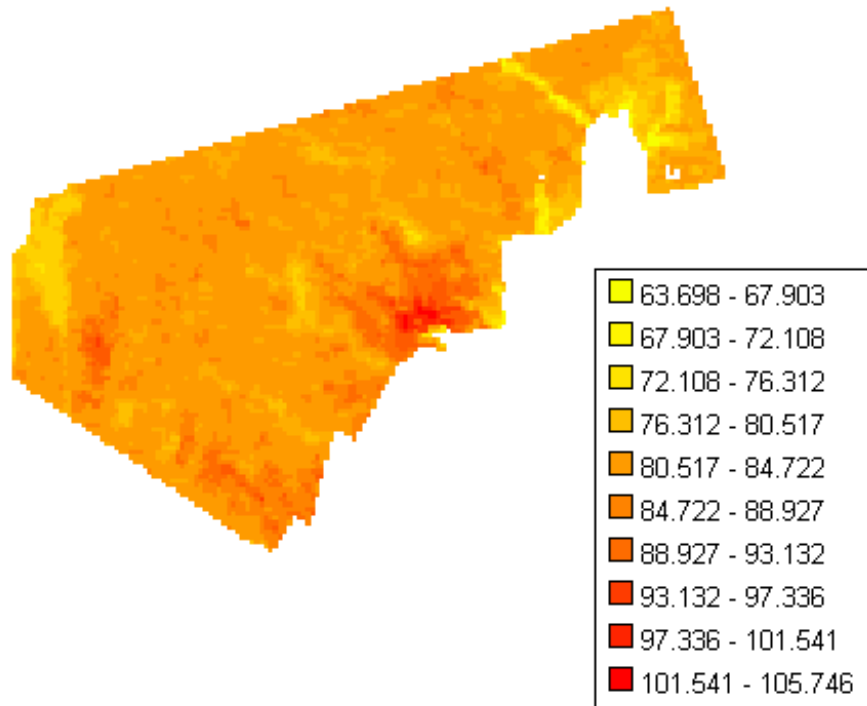
LST Distribution Example



Mean = 85.4994 degrees F

Standard Deviation = 4.3092 degrees F

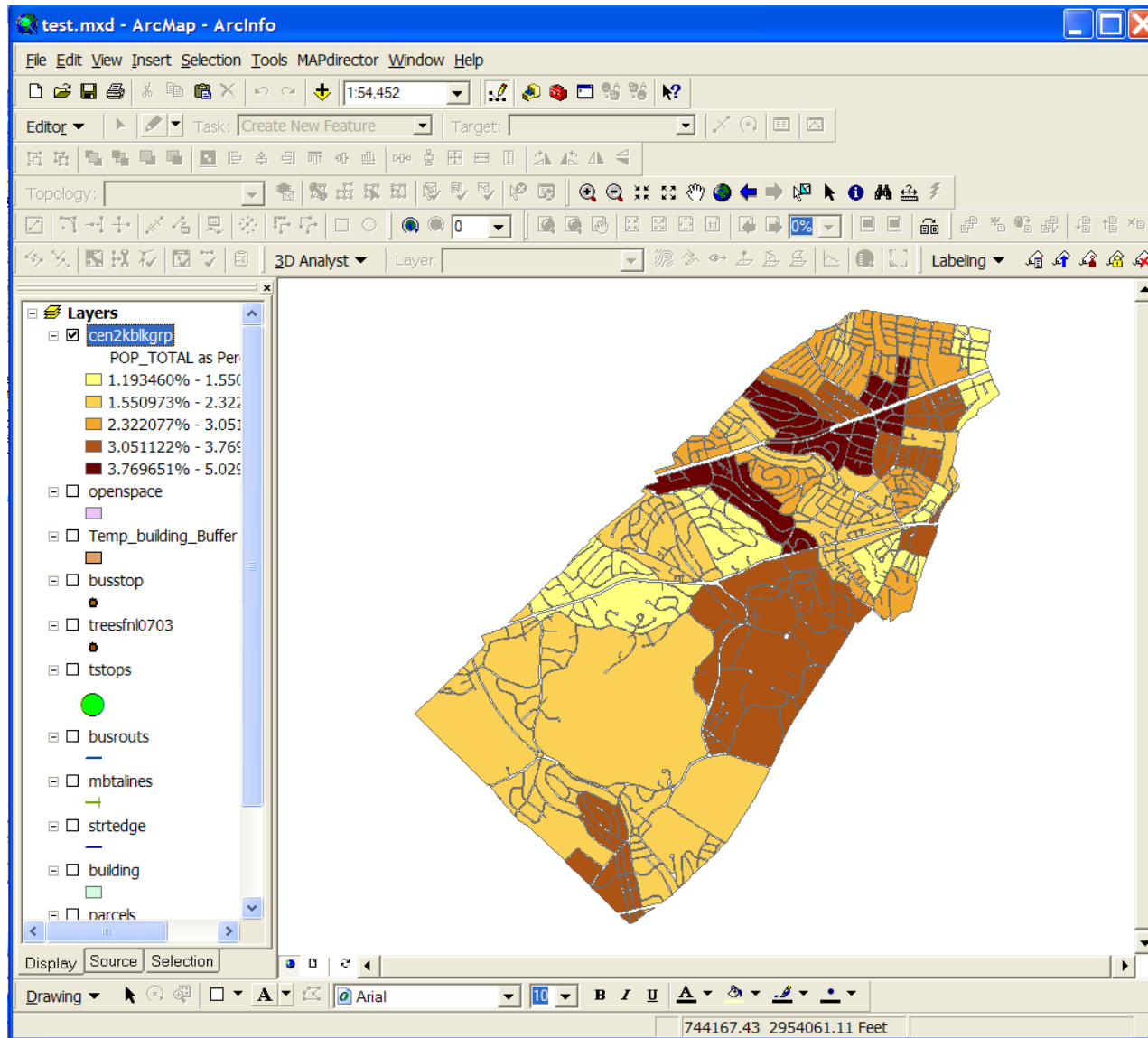
LST Z-Scores



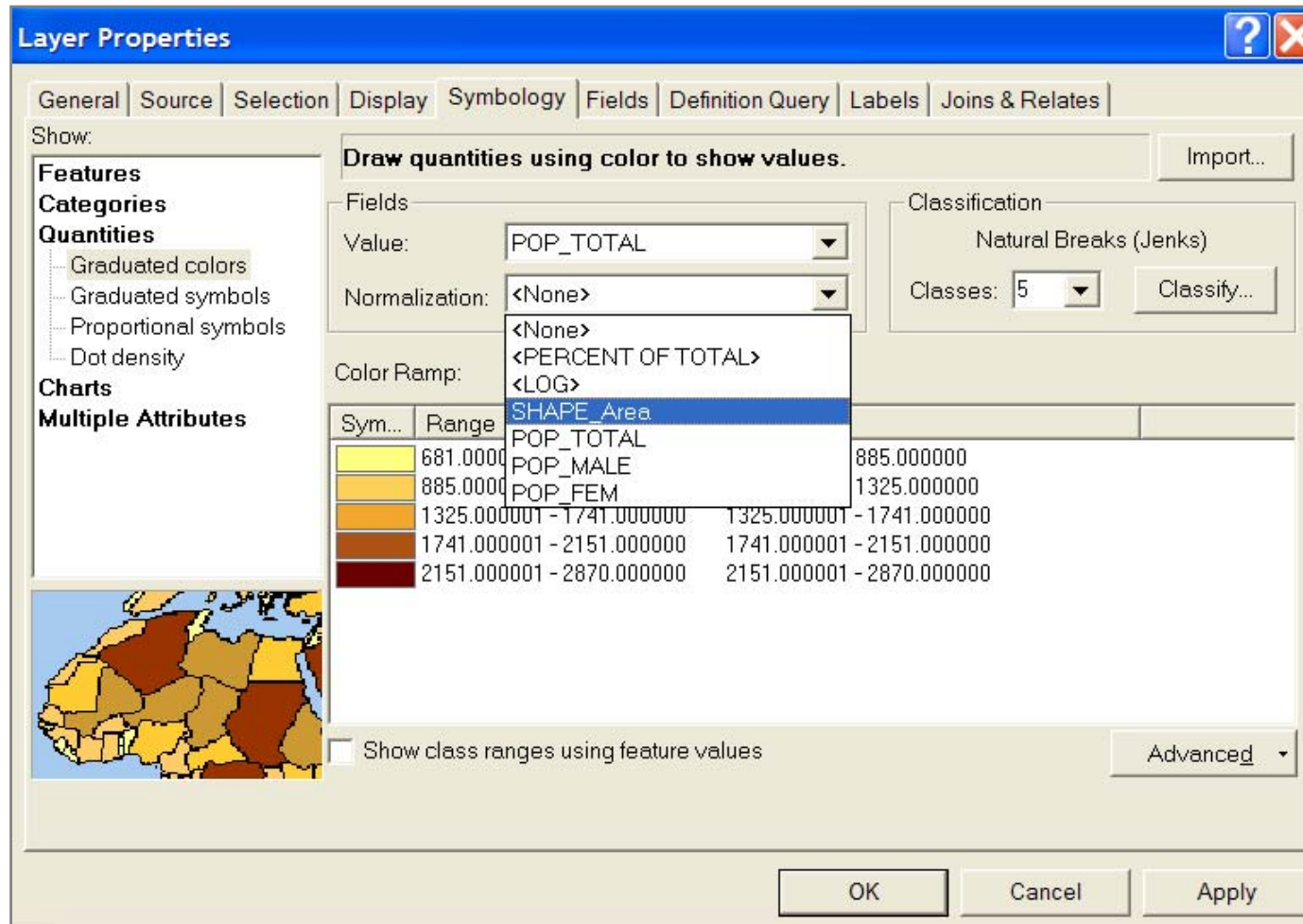
Mean = 85.4994 degrees F
Std. Dev. = 4.3092 degrees F

$$\text{Z-score} = \frac{x - \mu}{\sigma}$$

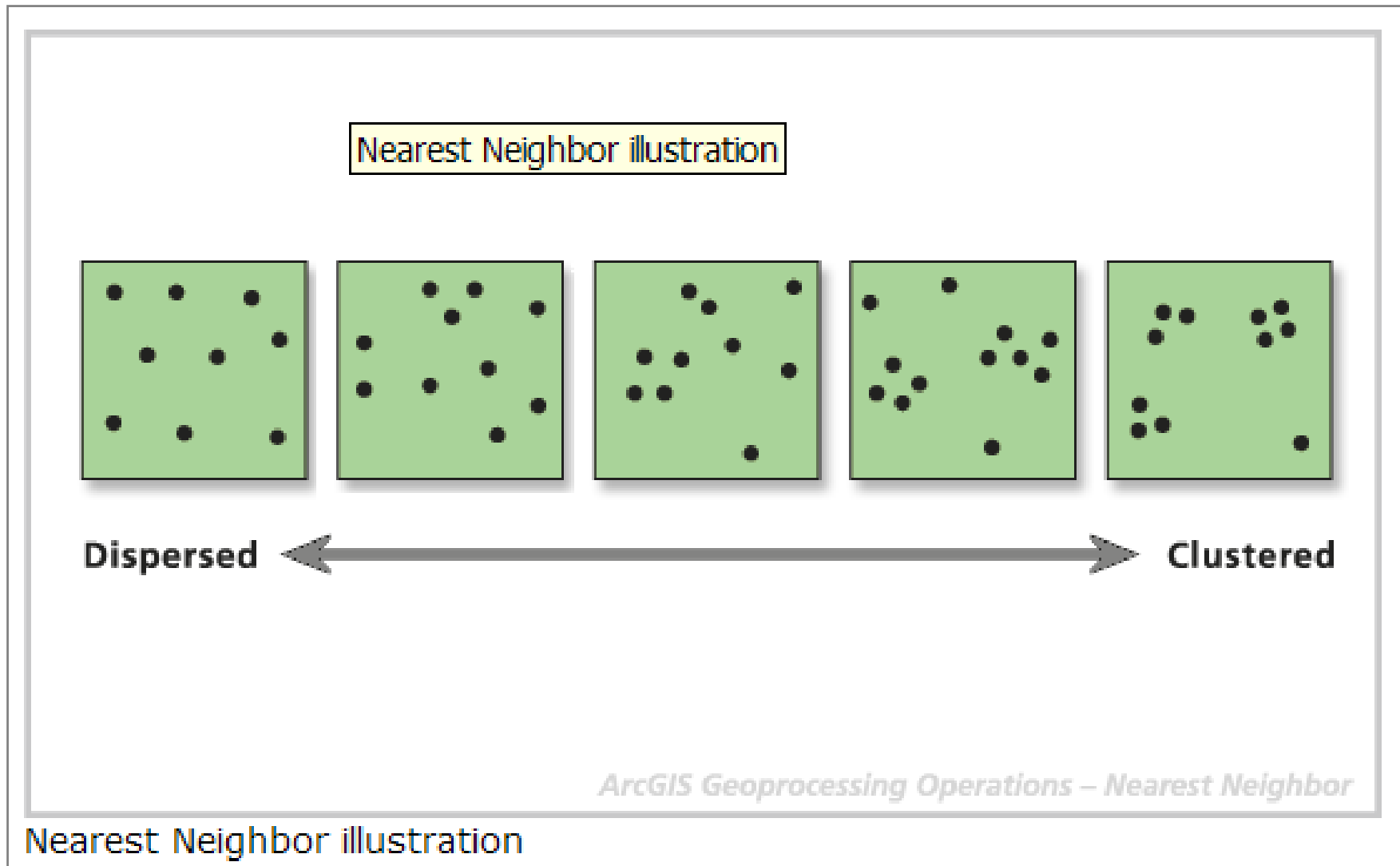
Describe With Maps



Normalization



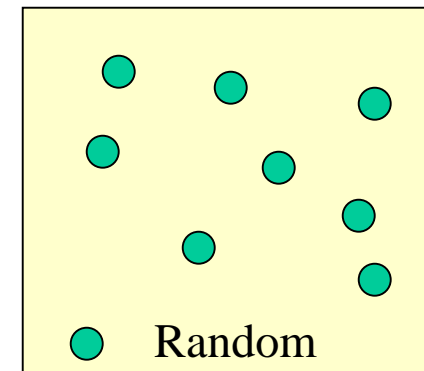
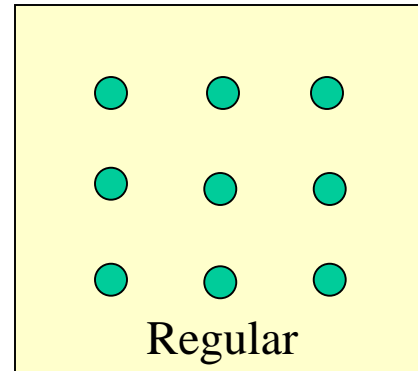
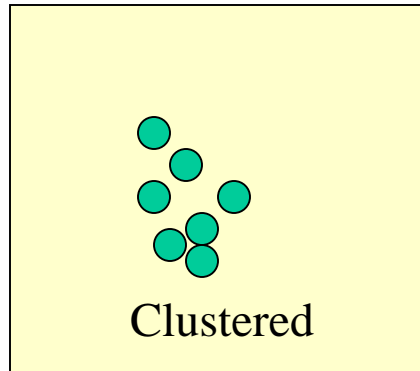
Spatial Statistics Tools – Average Nearest Neighbor



Average Nearest Neighbor

- The **average nearest neighbor** distance tool **measures the distance** between each feature centroid and its nearest neighbor's centroid location.
- It then **averages all** of these nearest neighbor distances. If the average distance is **less than the average** for a hypothetical random distribution, the distribution of the features being analyzed are **considered clustered**.
- If the average distance is **greater** than for a hypothetical random distribution, the features are **considered dispersed**.

Average Nearest Neighbor



- Point patterns can be characterized by the **distance between neighboring points**. If we define d_i as the distance between a point and its nearest neighbor, the **average distance between neighboring points** can be written as:

$$D_A = \frac{\sum_{i=1}^n d_i}{n}$$

Average Nearest Neighbor

- The **index** is expressed as the **ratio** of the **observed distance divided by the expected distance** (expected distance is based on a hypothetical random distribution with the same number of features, covering the same total area).
- Hence if the **index is less than 1**, the pattern exhibits **clustering**; if the index is **greater than 1**, the trend is toward **dispersion**.

The Nearest Neighbor Index

- We can calculate the **expected distance** (D_E) between randomly distributed points using:

$$D_E = \frac{1}{2} \sqrt{\frac{A}{n}} \quad \text{where } A \text{ is the area and } n \text{ is the \# of points}$$

- We can determine the degree to which a set of points is randomly distributed by **comparing** the **actual distance** between the points (D_A) with the **expected distance** (D_E), taking the **ratio** between the two, known as the **nearest neighbor index** (NNI):

$$NNI = \frac{D_A}{D_E}$$

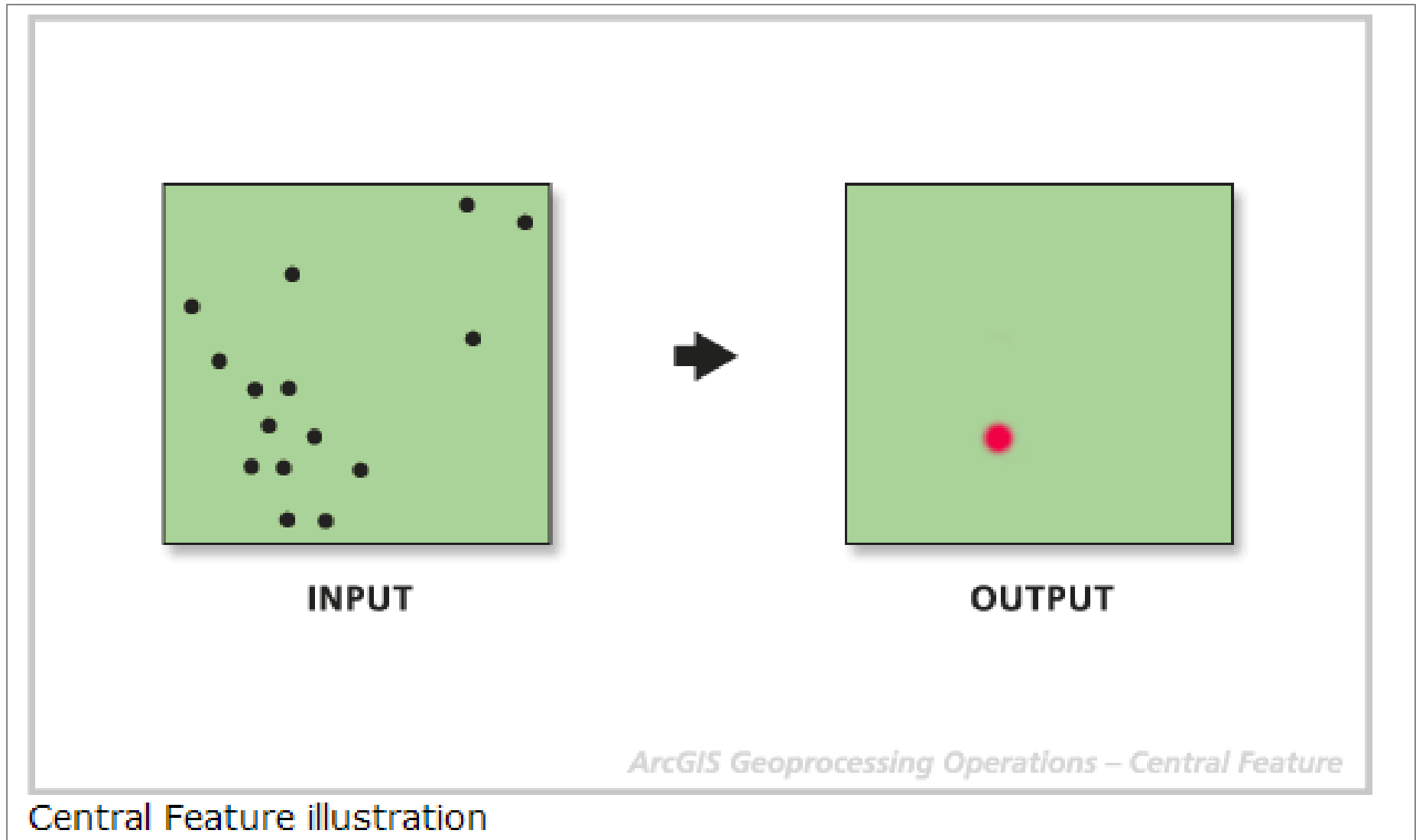
- **Random** points: $D_A \sim D_E$, $\therefore NNI \sim 1$
- **Clustered** points: $D_A \sim 0$, $\therefore NNI \sim 0$
- **Dispersed** points: D_A larger up to max. $NNI = 2.1491$

Average Nearest Neighbor

Possible applications

- Evaluate **competition or territory**. Quantify and compare **patterns in distributions** for a variety of plant or animal species.
- Compare an **observed distribution** to a **control distribution**. For example, in a timber analysis, you may want to compare the pattern of harvested areas to the pattern of harvestable areas in order to determine if cut areas are more clustered than you would expect given the distribution of harvestable timber overall.
- This stat is most appropriate when **the study area is fixed**: comparing average nearest neighbor distances for different types of retail stores within a particular county, or comparing a single type of retail for a fixed study area over time.

Spatial Statistics Tools – Central Feature



Central Feature

- The Central Feature tool identifies the **most centrally located feature** in a point, line, or polygon feature class.
- **Distances** from **each feature's** geometric centroid to **every other feature's** geometric centroid in the dataset are **calculated and summed**.
- Then the feature associated with the **shortest accumulative distance to all other features** is selected and copied to a newly created output feature class.

Central Feature

Potential Applications

- For example, if you wanted to **build a performing arts center**, you could **calculate the central feature** weighted by population to identify the town accessible to the most people in the region and make it a top candidate. The Central Feature tool is useful for **finding the center** when there is **travel between the features and the center**.

Optimization

- Spatial analysis can be used to solve many **problems of design**, such as “where is the best place to build a new x”
- The decision as to where to build a new facility is often approached from the point of view of **maximizing access**, or **minimizing travel time** from a certain catchment or service area,
 - e.g. if we identify a developing area where the nearest hospital is an unacceptably long drive away, we may know we want to locate a hospital in that area ... but **where should we put it** to best serve the residents in the area and minimize overall travel time for the area?
- To do, we can identify the **point of minimum aggregate travel** (MAT)

Applications of the MAT

- We previously looked at how to calculate the **centroid**, which **minimizes** the **sum of distances squared**, but **not** the **sum of distances** from each point
 - The center with that property is called the **point of minimum aggregate travel** (MAT)
 - The MAT must be found by **iteration** rather than by calculation (i.e. testing successive points by accumulating travel time until the minimum is found)
- Because it **minimizes distance**, the MAT is a useful point at which to locate any central service
 - e.g., a school, hospital, store, fire station
 - finding the MAT is a **simple instance** of using spatial analysis for **optimization**

Optimizing Point Locations

- The MAT is a **simple case** of optimization: We are trying to locate **one service location** with the goal of minimizing the total distance traveled from nearby locations to the service location
- We can come up with **more complicated optimization problems** when we consider **several points at once**:
- For example, the operator of a chain of convenience stores or fire stations might want to optimize service for many locations at once:
 - Where are the **best locations** to **add new services**?
 - Which **existing** service locations **should be dropped**?

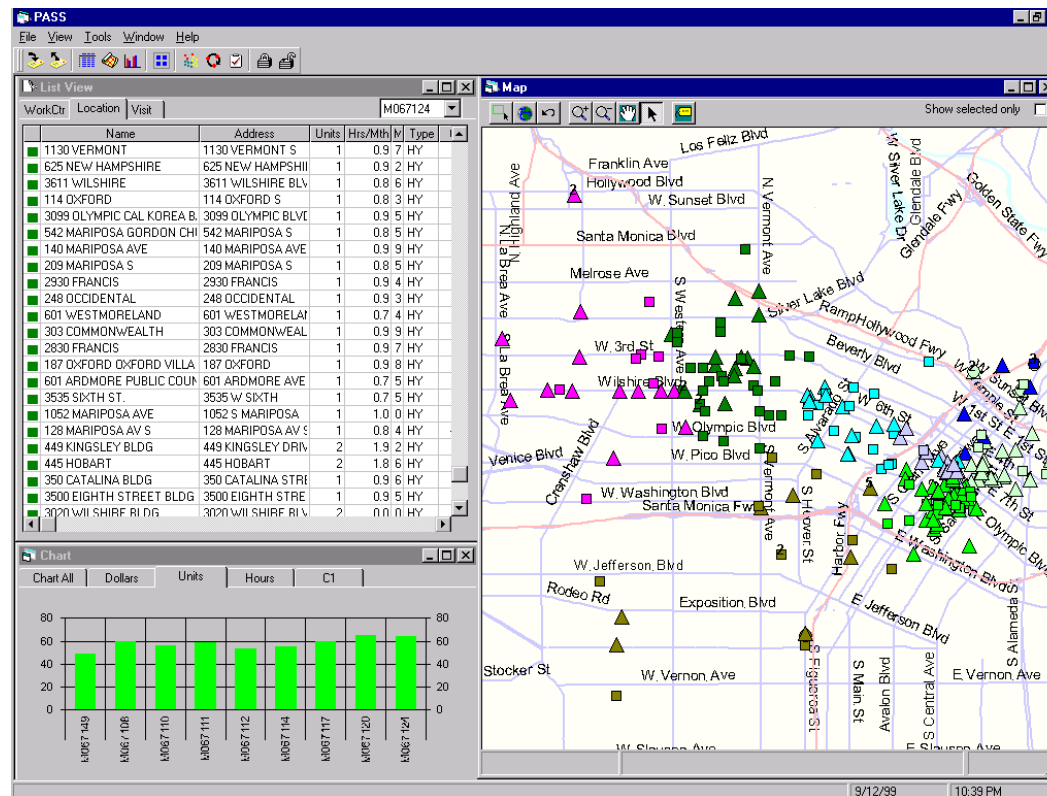
Location-Allocation Problems

- This class of problems is known as **location-allocation problems**, and solving them usually involves **designing locations** for services, and **allocating demand** to them to achieve specified goals
- Those goals might include:
 - minimizing total distance traveled
 - minimizing the largest distance traveled by any customer
 - maximizing profits
 - minimizing a combination of travel distance and facility operating cost

Routing Problems

- Another type of optimization problems is known as **routing problems**:
- Suppose we have a set of locations we need to visit, and the decision we need to make is how to reach them all in a **timely and efficient fashion**
- What we need to do is to search for **optimum routes** among several destinations
- The **‘traveling salesman problem’** is just this sort of problem:
 - Find the **shortest** tour from an origin, through a set of destinations, that returns back to the origin

Routing Problems



- Routing service technicians for Schindler Elevator: Every day this company's service crews must **visit a different set of locations** in Los Angeles. GIS is used to partition the day's workload among the crews and trucks (color coding) and to **optimize the route to minimize time and cost**

Optimum Paths

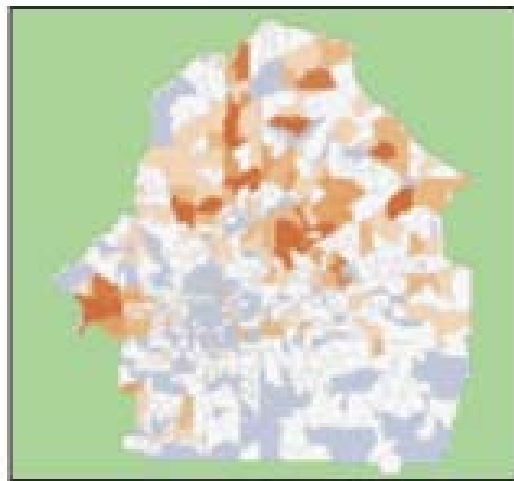
- Another sort of optimization problem is encountered when we have a **known origin and destination**, and we need to find the **best route** between the two, given data that describes the ‘cost’ of taking various paths
- The goal is to find the best path across a **continuous cost surface**
 - The goal is to **minimize total cost**
 - The cost may combine construction, environmental impact, land acquisition, and operating costs
 - This is used to **locate** highways, power lines, pipelines
 - It requires a **raster representation**
- (Finding the optimal route between two locations on a vector road network is a **related** sort of problem)

Least-Cost Path Example

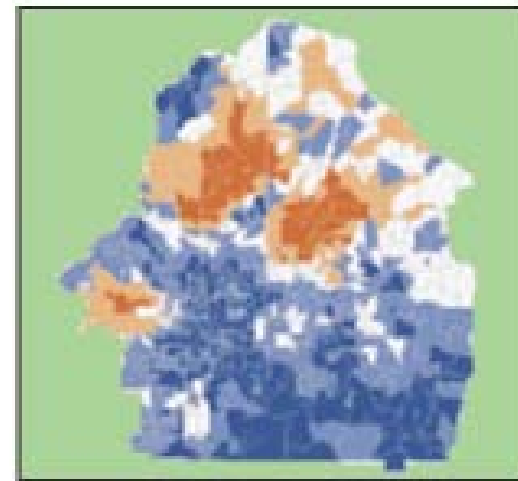


- The figure to the left shows the solution of a **least-cost path problem**:
- The white line represents the **optimum solution**, or path of least total cost, across a friction surface represented as a raster layer
- The area is dominated by a mountain range, and **cost** in this example is determined by **elevation and slope**
- The best route uses a **narrow pass** through the range. The blue line results from solving the same problem using a coarser raster

Spatial Statistics Tools – Hot Spot Analysis



INPUT



Gi* Z SCORES

ArcGIS Geoprocessing Operations – Hot Spot Analysis

Hot Spot Analysis (Getis-Ord G_i^*) illustration

Hot Spot Analysis

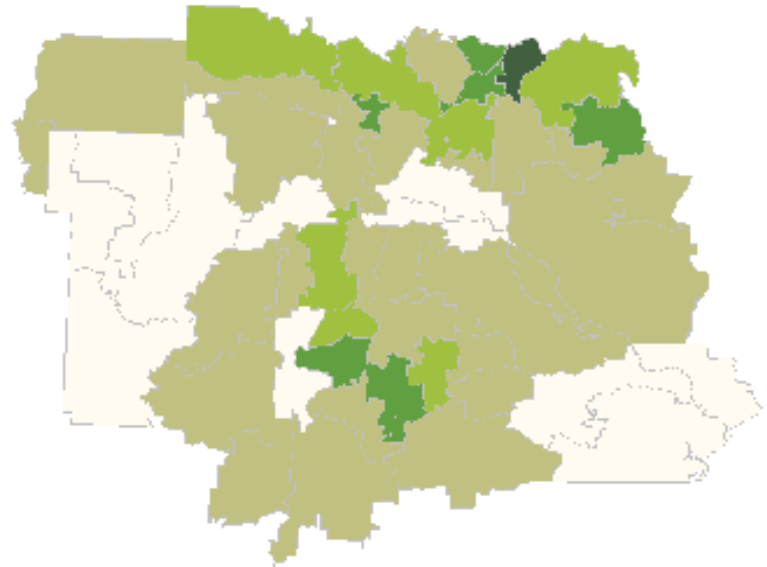
- The hot spot analysis tool **identifies spatial clusters** of statistically significant **high or low attribute values**.
- This tool calculates the **Getis/Ord G_i^* statistic**. The G-statistic tells you whether high values or low values (but not both) **tend to cluster** in a study area. Thus it's used to **identify** whether **hot spots OR cold spots** exist.
- A **high value** for the G-statistic indicates that **high values**-that is, values higher than the mean for the study area-**tend to be found near each other**.
- A **low value** for the G-statistic indicates that **values lower than the mean tend to be found together**.

Hot Spot Analysis

- The G-statistic is useful if you're trying to **identify the presence of clusters** of extremely high or low values.
- For **example**, if you are opening a chain of athletic clothing stores in a region, you could use the G-statistic to see if there are concentrations of people who are very likely to spend money on sporting goods, before you even start to look at specific locations for the stores.

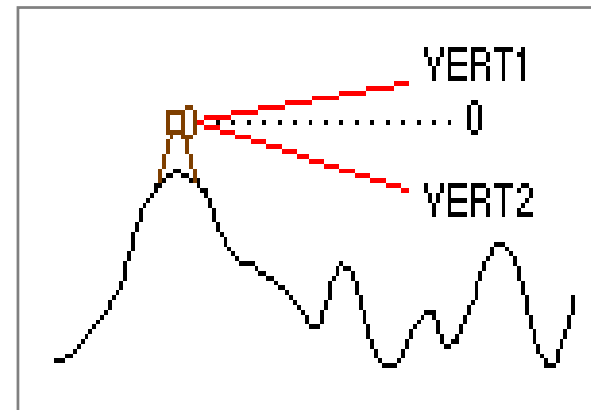
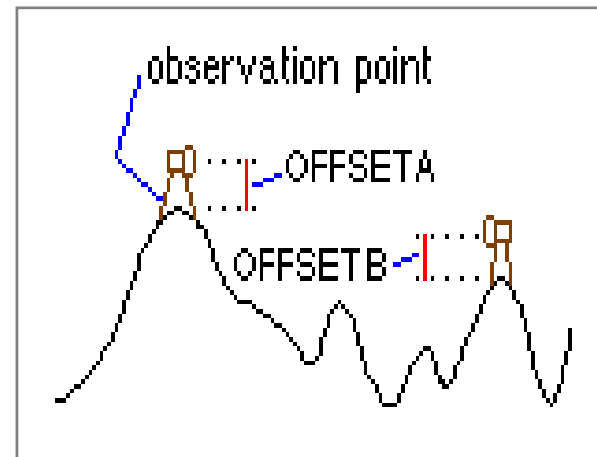
Hot Spot Analysis

- The hot spot analysis tool creates a **new feature class** that duplicates the input feature class then **adds a new results column** for the hot spot (G_i^*) Z score values.



Spatial Analysis Tools – ViewShade Analysis

- **Which areas can be seen** from a fire lookout tower that is 15m high?
- **How frequently** can a proposed disposal site **be seen** from an existing highway?
- **Where** should the next communications repeater tower in a series be located?



Spatial Analysis Tools – Line of Sight

- Uses an **input 3D line** feature class to **determine visibility along its lines.**
- Produces an **output line** feature class that contains **line and target visibility information.**
- If the **target is not visible**, Line of Sight produces an output point feature class that shows the **first obstruction points** along the lines..

Spatial Analysis Tools – Surface Volume

- Calculates the **area and volume** of a functional surface above or below a **given reference plane**.



Chapter 6: Why is it There?

- 6.1 Describing Attributes
- 6.2 Statistical Analysis
- 6.3 Spatial Description
- 6.4 Spatial Analysis

Next Topic:

Making Maps with GIS