NISGTC

The National Information, Security & Geospatial Technologies Consortium

Authoring Organization: Del Mar College
Written by: Richard Smith

Development was funded by the Department of Labor (DOL) Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant No. TC-22525-11-60-A-48; The National Information Security, Geospatial Technologies Consortium (NISGTC) is an entity of Collin College of Texas, Bellevue College of Washington, Bunker Hill Community College of Massachusetts, Del Mar College of Texas, Moraine Valley Community College of Illinois, Rio Salado College of Arizona, and Salt Lake Community College of Utah.

Lesson 1: Reviewing the Basics of Geospatial Data

INTRODUCTION

In this lesson you will learn about basic geospatial data elements, coordinate systems, and how to distinguish between vector and raster data formats. You will be introduced to types of spatial phenomena, attribute data types, map design and map elements.

LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Identify basic geospatial data elements.

2. Explain the various coordinate systems and their importance.

3. Differentiate vector and raster data formats.

LEARNING SEQUENCE

Required Reading	Read the following: • Online Lesson Material
Resources	View the following: • None
Assignments	Complete the following: • Lab: Reviewing the Basics of Geospatial Data • Quiz: Reviewing the Basics of Geospatial Data

INSTRUCTION

Types of Spatial Phenomena

To simplify reality and store the concept of reality in a geospatial data model, it is important to understand the way in which spatial phenomena are structured. There are two types of spatial phenomena that will be discussed in this lesson: discrete and continuous.

Discrete

A discrete spatial phenomenon is anything that exists that is individually distinguishable. It has well-defined boundaries and it is easy to see where it begins and ends. It does not exist between where we observe it to be. For example, a stream or lake is both easy to individually distinguish on the earth (Figure 1). The same thing goes for the roads. Roads have well-defined boundaries and are each easily distinguishable as individual roads (Figure 2).

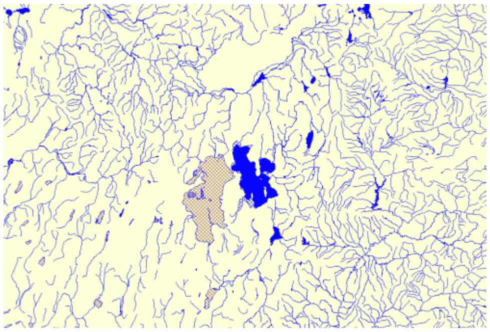


Figure 1 Discrete Spatial Phenomenon Streams and Lakes

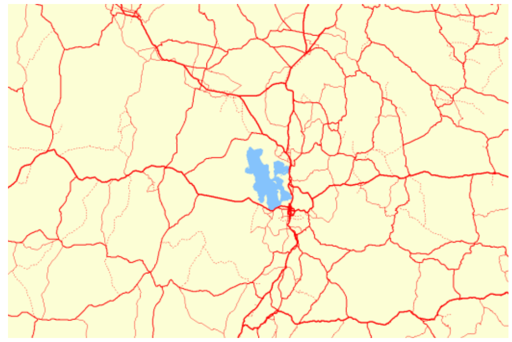


Figure 2 Discrete Spatial Phenomenon Roads

Continuous

A continuous spatial phenomenon is something that exists between our observations. It is data of a continuous nature that cannot be isolated as an individual. It is perhaps best explained through examples. For instance, Figure 3 illustrates how temperature is a continuous phenomenon that gently varies throughout space. A temperature reading at a single location does not represent a well-defined location where it is exactly that temperature. Instead, the temperature reading must be put in the context of the surrounding area as part of the larger "surface" of temperature readings; temperatures should be considered as part of a larger continuous spatial phenomenon. Elevation is another example of a continuous spatial phenomenon. An elevation reading is only a single point in a larger surface (Figure 4).

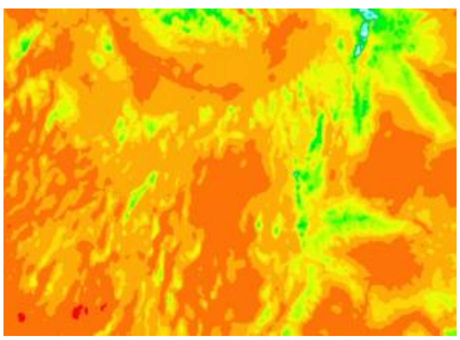


Figure 3 Continuous Spatial Phenomenon

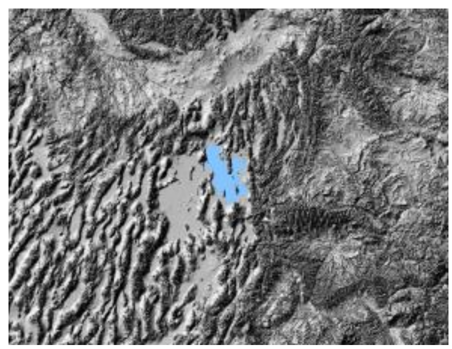


Figure 4 Continuous Spatial Phenomenon Elevation

Attribute Data

Attributes are the non-spatial characteristics that describe spatial entities. Attributes are commonly arranged in tables where a row is equivalent to one entity and a column is equivalent to one attribute or descriptor of that entity. Typically, each row relates to a single object in a spatial data model. It is also typical for each object to have multiple attributes that describe the object. All attributes are often displayed in a table format. Attributes can be stored on a computer using a flat file formator in a database management system.

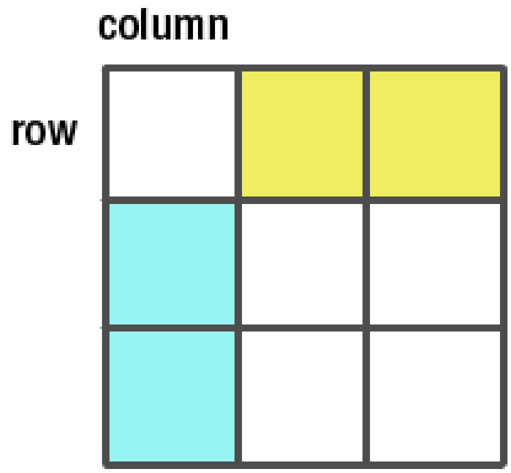


Figure 5 Attributes: Row = 1 Entity, Column = 1 Attribute

For example, consider that a spatial data model that stores the location of fire hydrants. In order for each object to represent a fire hydrant, you would need to store their positions. In addition to positional information, you would also store attributes that describe those fire hydrants. In this example, color, service, date, and flow are being stored as three attributes that describe this particular fire hydrant at this particular position on Earth. The position, color, service, date, and flow will be stored as one row in an attribute table that will contain four columns because there are four descriptors for this fire hydrant.

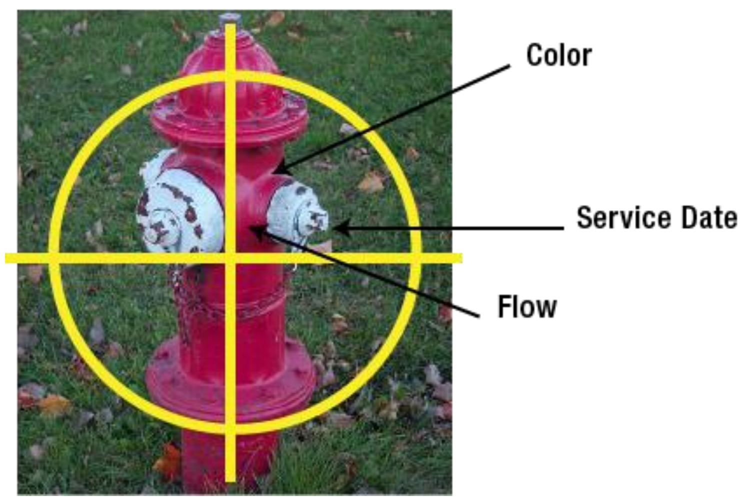


Figure 6 Fire Hydrant: Includes four descriptors – position, color, service date, and flow.

Attribute Data Types

Computers fundamentally "think" differently than humans. While humans see numbers, letters, pictures, and sounds, a computer only sees zeros and ones, which is known as binary data. Therefore, we need a way to translate the numbers, sounds, and videos, as humans know it, to a form in which a computer can understand and store the information. Computer scientists have created data structures that can be used to translate our information into a format which a computer can store in its memory. This data structure is known as a data type. There are four typical data types used in GIS: integer, float/real, text/string, and date. In order to use the computer's memory most efficiently, it is important to specify which data type is going to be used to store information in the computer's memory. It is important to let the computer know which operations are allowed for each data point stored in that memory location using a specific data type.

Vector Data Models

A vector data model defines discrete objects. Examples of discrete objects are fire hydrants, roads, ponds, or a cadastral. A vector data model is broken down into three basic types: points, lines, and polygons. All three types of vector data are composed of coordinates and attributes.

Points

A point uses a single coordinate pair to define its location. Points are considered to have no dimension even though they may have a real world dimension. For the purposes of a GIS, no dimension is assumed. Each point has associated attribute information, and the information is attached to the center of the point. Examples of spatial phenomenon that would be modeled well as points are light poles, manhole covers, and crime locations.

Lines

A line vector type is defined by an ordered set of coordinates. Each line, and curve, is made up of multiple line segments, however, on occasion, curved lines are represented mathematically. There are two words needed to define when discussing lines: a node, and a vertex.

A node is where a line begins or ends. A vertex is where a line changes direction. The smallest possible line will have two nodes, a start node, and an end node. Longer lines will have at least two nodes, and many vertices in between where the line changes direction. Attributes may be attached to the entire line, individual node, or individual vertices, therefore, each line may have multiple rows of attributes in the attribute table.

For example, if a line represents a road, each road segment between two intersections may have its own address information, such as the start address and the end address for that block. An intersection may have an attribute that describes where the intersection has a stop sign or stoplight. The other option is for the entire line to have one row of attributes no matter how complex the line. Examples of spatial phenomenon that are modeled well by lines are roads, pipelines, outlines of objects, and power lines.

Polygons

The last vector data type is the polygon. A polygon is formed by a set of connected lines where the start and end point have the same coordinate. Because the start and end point have the same coordinate, the polygon will close and will have an interior region. Attribute information is attached to the center of the polygon no matter how complex the polygon. Examples of spatial phenomenon modeled well by polygons are lakes, cities, tree stands, and political boundaries.

Raster Data Models

The raster data model best represents continuous objects such as temperature or elevation. A raster can take the form of a regular set of cells, like pixels in a photograph or it can appear as cells arranged in a grid pattern, which is also referred to as a matrix (Figure 6).

Each cell in the raster contains a single value, and the coordinate of each cell of the raster refers to the center of the cell. Therefore, the single value stored in each cell of the raster, applies to the entire cell in the raster matrix. Each cell can be defined by a cell dimension such as the cell width and height. Often, cells in a raster are square, so the cell width and cell height will be the same.

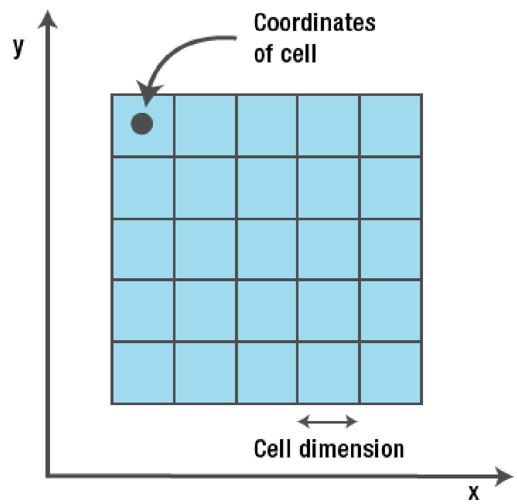


Figure 7 Raster Data Model

Raster Resolution

In GIS, it is important to know the resolution of the raster. The raster resolution is the cell size of each cell of the raster. Unlike how photographers represent resolution, as the number of megapixels their camera uses, in GIS, we are not as concerned about the number of cells, but of how much area on the ground each cell covers. There is a direct trade-off between resolution and file size, the cell coordinate is a center of the point cell, and the coordinate applies to the entire cell area.

It is important to reiterate that each raster cell represents a given area and the value assigned applies to the entire cell area. If there is more than one value, they can fall inside the raster cell, and then the raster cell may contain the average, central, most common, or only value covered by the cell.

Storage Formats

There are three types of storage formats that enable you to store data in various formats: a shapefile, a raster, and a geodatabase. Review each format to learn about how different types of data can be stored.

Shapefiles

A shapefile is a common data format for storing vector data that spatially describes geometries such as points, lines, and polygons. In a shapefile, each object, or record, must also have attributes that describe them. It is important to note that a shapefile can only hold one geometric type. That means, that a shapefile can hold only points, or only lines, or only polygons.

Rasters

Rasters are stored in many different file formats. Common file formats are JPEG, TIF, bitmap, SID, and so on. Each format will have different advantages and disadvantages, therefore you should read the documentation of each one of these file formats when determining which one to store your raster file in.

Geodatabase

A geodatabase is a storage container for GIS data. Databases have many advantages over shapefiles such as speed, ability to hold vector and raster data, ability to restrict attributes to certain formats or range of values, ability to enforce rules on input, and many more.

The standard geodatabase contains two major structures that you should be aware of a feature dataset and a feature class.

- A feature data set is simply a collection of feature classes. You can think of a feature data set as a folder inside of the database. A feature data set will contain one or more feature classes and all those feature classes must have the same coordinate system. In other words, a feature data set is a logical grouping of feature classes inside of a geodatabase.

- A feature class is a collection of common features and is analogous to a single shapefile. Each feature class must have the same geometry, just like shapefile.

Example: National Hydro Data Set

Take a look at the National Hydro Data Set (Figure 8) which is stored in a geodatabase to see how the feature data set and feature classes work together. The geodatabase is the parent level container for all of the information; the feature data set logically grouped together feature classes. Remember that all feature classes inside of the feature data set must have the same coordinate system.

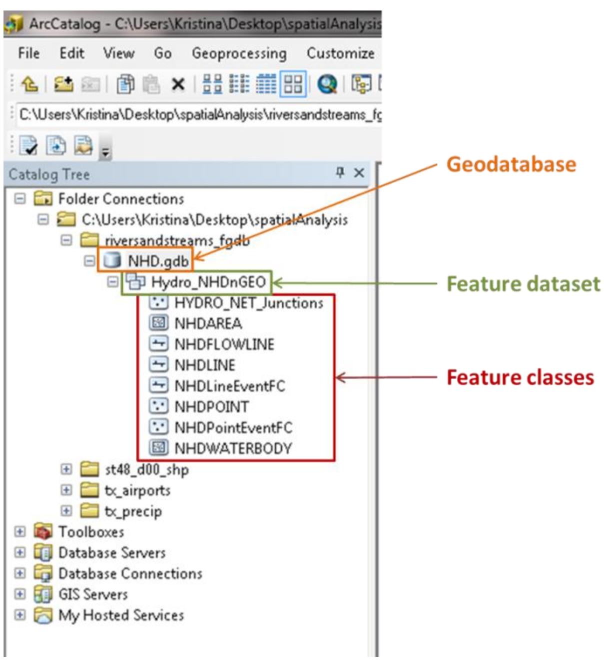


Figure 8 Screenshot of the National Hydro Data Set stored in a geodatabase displaying the feature dataset and classes.

Coordinate Systems and Map Projections

It is extremely important that you gain an understanding of coordinate systems and map projections. GIS works with spatial data and with using data that needs to be placed at locations. You will need to understand the building blocks of how to define and visualize a location. It will be important to focus on map projections because although globes are great for visualization purposes, they are not practical for many uses. One reason globes are not practical is that they are not very portable. Additionally, a round Earth cannot fit without distortion onto a flat piece of paper, so you will need to understand how the earth is distorted in order to flatten it for the creation of a map.

Map Projections

A map projection is defined as a systematic rendering of locations from the earth's curved surface onto a flat map surface. This allows us to flatten the curved surface onto a flat surface such as a piece of paper, or computer screen. The reason why map projections are employed is because globes are not very portable, or practical to use in some cases, therefore, map projections are used to flatten the earth into a map.

Figure 9 is a basic illustration displays the concept of a map projection. The map projection is broadly composed of three parts, the ellipsoid, which models the shape of the earth, a light source which is used to project features on the earth surface, and a developable surface, commonly a flat piece of paper, onto which the Earth's features are projected, and flattened, to be used as a map.

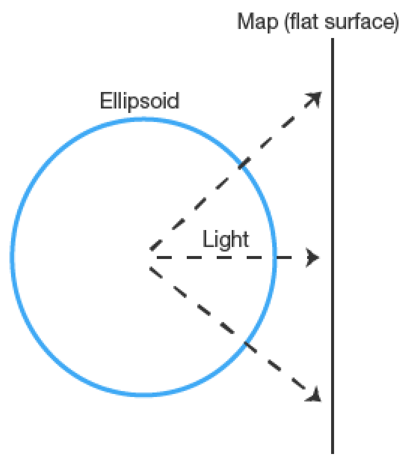


Figure 9 Map Projection

Projection Parameters

There are five map projection parameters:

- Standard Points and Lines
- Projection Aspect
- Central Meridian
- Latitude of Origin
- Light Source Location

A map projection property is defined as an alteration of area, shape, distance, and direction on a map projection. Map projection properties exist based on the conversion from a three dimensional object, for example the Earth, to a two-dimensional representation, such as a flat paper map. This conversion requires the deformation of the three-dimensional object to fit onto a flat map. The three-dimensional spherical surface is torn, sheared, or compressed in order to level it into a flat developable surface.

Map Distortion

In order to create a flat map of a three-dimensional object, the three-dimensional object must be distorted.

- Distortions are unavoidable when making flat maps of the earth.
- Distortion is not constant across the map, as distortion may take different forms in different parts of the map.
- There are few points where distortions are going to be zero, however, distortion is usually less near the points or lines of intersection where the developable surface intersects the globe.
- By determining where the standard points and lines are placed will directly affect where the map will have the least and most amount of distortion.

Map Design and Map Elements

A map is a way of representing our world-based on the knowledge and culture of the mapping society at a particular time in history. The purpose of a map is to transmit knowledge visually. Maps are often considered to be one of the three major modes of communication.

A cartographer is someone who designs and prepares a map for distribution. Cartographers are trained professionals in the field of cartography which is the art and science of making maps. Cartographers also study the philosophical and theoretical bases of the rules for making maps. Cartography is a professional field that has existed for hundreds of years and it takes cartographers many years of apprenticeship to become skilled at their craft. However, with today's computer technology allowing us to quickly manipulate, analyze, and visualize spatial information, mapmaking is now being shared by professions outside of cartography. Therefore, it is important that any user of a GIS be trained in cartography so that their maps will be effective tools for communication.

Map Title

The title of a map should be dominant in size and is typically the largest text on a map. A good map title should focus the user's attention on the purpose of the map. A good map title should also be brief but descriptive. Typically, a map title will include information such as where the map is focused, what information is being focused upon, and the timeframe for which the map is applicable. Map titles are typically placed at the top center of the production medium, however a map title can truly be placed anywhere so long as it is easily found by the map user. If the map is a figure in a larger document, you should not place the title on the map, but instead, place the title of the map in the caption.

Map Body

The map body is the main focus of the map and contains the geographic features that are important to the message of the map (Figure 10). The map body is typically the largest map element on the map, and should dominate the user's attention.

Typically, when designing a layout, the map body is placed first, and other elements are then placed around. However, do not be afraid to move or resize the map body to better accommodate other elements. Being that the map body is the element we want the user to focus on, it should be easy to find, dominant and of adequate size to effectively show the geospatial data.

Figure 10 Example of Map Body

Inset Maps

Inset maps are small ancillary maps that have a larger scale than the main map body. The role of an inset map is to show more detail in a map body of a smaller geographic area. In the example provided (Figure 11), the United States of America is the main map body and the smaller map of Arkansas in the lower right-hand corner, is the inset map which is showing a smaller area in more detail. To make it obvious to the map reader where the inset map is referring to on the main map, you should show an outline of the extents of the inset map of the main map body, or provide leader lines from the main map body to the inset map.

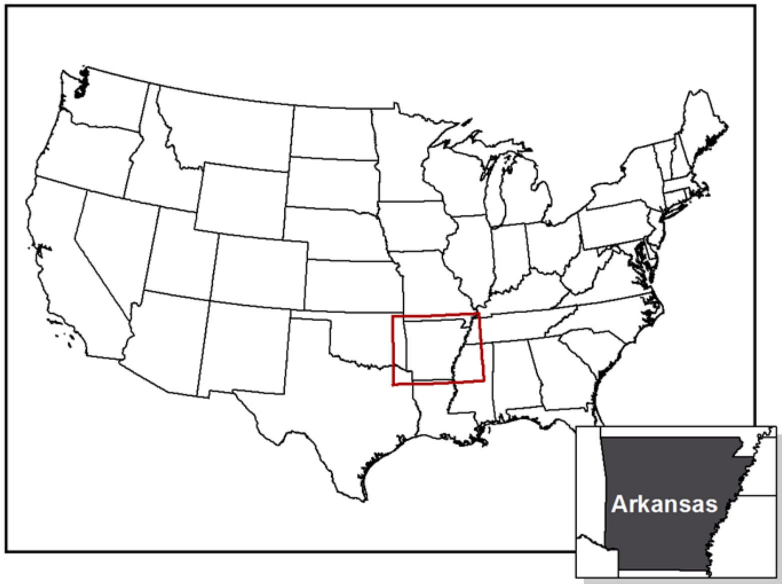


Figure 11 Inset Map

Location Map

A location map is a small ancillary map that is at a smaller scale than the map body. The location map identifies a location of where the main map body is in a larger geographic context. A location map is to be used when the location of an area on the main map body is unfamiliar or not intuitive to map reader. In this illustration (Figure 12), the map of Arkansas is the main map body, and the smaller map of the lower right-hand corner is the location map. Similar to the inset map, there is a visual marker on the inset map that shows the map reader where the main map body is located.

Figure 12 Map of Arkansas

Map Scale

The map scale is used to measure linear relationships on the map. A map scale is typically included on a reference map, but is not required to be included on thematic maps. A map scale should only be included on a map when you want the user to measure the distance on the map, or the scale of the map is not intuitive to the map reader. There are three types of scales that we can place on a map: a graphic scale, a verbal scale, and a representative fraction scale.

Graphic Scale

The graphic scale is perhaps the most common type of scale placed on maps. The graphic scale is a visual representation of the ratio at which the earth has been reduced. The graphic scale typically starts at zero, and measures out to a meaningful, typically round number (Figure 13). One major advantage of a graphic scale is that if the map is enlarged or reduced, say using a photocopier, then the graphic scale will scale with the enlargement or reduction, and will always be correct.

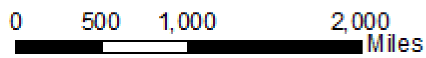


Figure 13 Graphic Scale illustrating zero to 2,000 miles

Verbal Scale

The second type of scale is the verbal scale. The verbal scale is a statement that describes how a distance measured on the map relates to a distance measured on the ground. Again, it is important to use meaningful, typically round measurement units to make it easier for the map user to measure distances.

Example: Verbal Scale

One inch on the map
equals twenty feet on the ground

Representative Fraction Scale

The third type of scale is the representative fraction scale sometimes referred to as the unit of scale. The representative fraction scale is a map scale that is used to represent units in centimeters, inches or feet in the form of a fraction or a ratio. This fraction or ratio, 1: x is used to indicate one unit on the map. The number to the left of the colon indicates that one unit on the map represents x units on the earth's surface indicated by the number to the right of the colon.

Example: Representative Fraction Scale

1:20

Legend Map

The legend map element identifies unknown or unique map features succinctly. A legend may optionally have a title, or contain the title of the entire map. The legend needs to have representative symbols that are found on the map followed by a description of what each symbol represents. The symbols on a legend should be the exact same color, shape, and size of the symbols shown the map. If the symbol on the map varies in size, the symbol in the legend should be the size of an average sized symbol on the map. Let’s look at two examples of legends.

General Reference Map

For a general reference map legend, it would display all symbols found on the map. The representative symbol should be to the left of the short description, and the legend can be organized vertically in one or more columns.

Figure 14 General Reference Map Legend

Thematic Map

For a thematic map and in this case, a graduated symbol map, the graduated circle legend is to show how the size of the symbol changes with the value of the attribute that it is representing. Graduated symbol legends can be placed in a vertical layout, horizontal layout, or a nested layout.

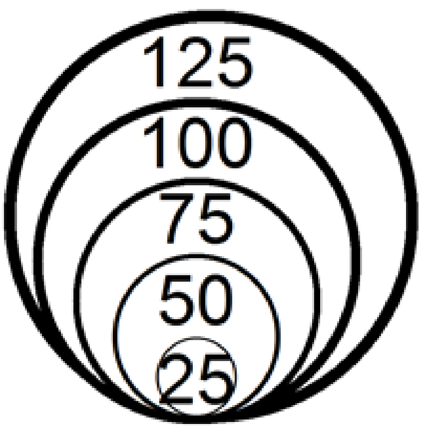


Figure 15 Graduated Symbol Nested Layout

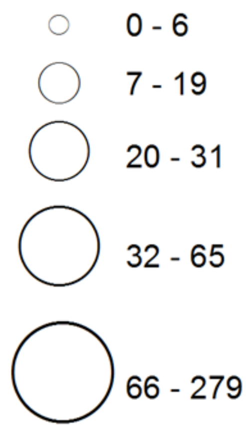


Figure 16 Graduated Symbol Vertical Layout

Directional Indicator

The directional indicator is often considered part of a legend, and may be placed inside the neat line around the legend, near the legend, or elsewhere on a map. The directional indicator, commonly a north arrow, is necessary when north is not at the top of the map, or the map readers are unfamiliar with the area being displayed on the map.

Since map readers are typically familiar with orientation of large landmasses, or their own country or state, north arrows are often not necessary on small scale maps. The directional indicator should be reasonable in size, not dominant at all, but should still be easily found by the map reader.

Labels

Labels communicate attribute or ancillary information directly on the map body, and related to map features on the map body. The purpose of label is to identify features on the map, and help users to orient themselves to the information being displayed on the map. Labels should be placed at locations that allow the map reader to easily associate each label with the feature it is labeling, and should be reasonable in size. For example, see Figure 17.

Figure 17 Labels on a Map

Metadata

Metadata or credits are used to cite the sources of data sets used to create a map. It is also used to provide the map author's information, the date the map was created, and other explanatory information about the creation of the map. Since you want the user's focus to be on other aspects of the map, metadata should not be visually dominating. Metadata is generally placed along a bottom edge of the map and deemphasized. If the map reader wants to read the metadata, they will typically spend a little time searching for it.

Graticule Map

The graticule map element visually represents a coordinate system or location scheme. You should include the graticule on a map if the map reader will be referencing coordinate locations throughout the map. You should use meaningful divisions on the graticule so that it is easy for the map reader to use it. Typically, graticules are omitted from thematic maps as the purpose of a thematic map is not to measure, but to look at spatial distributions and patterns of data.

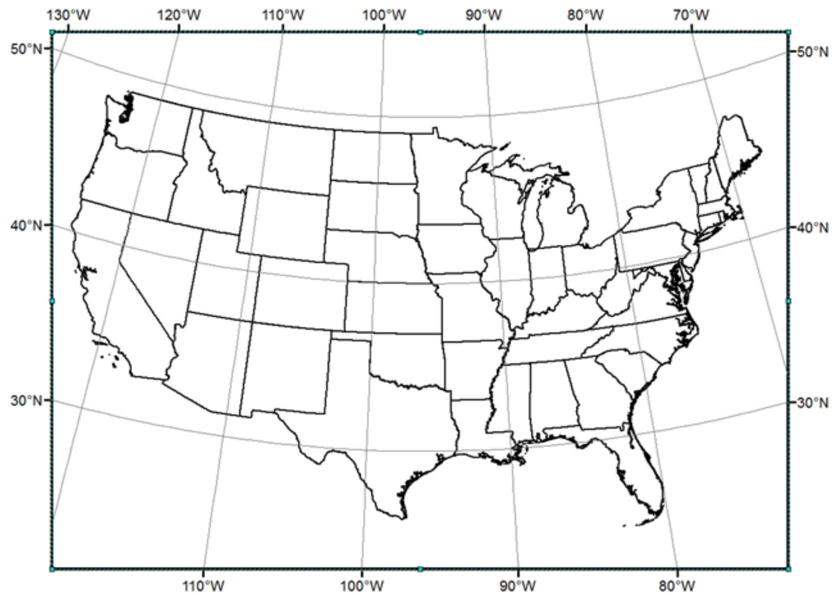


Figure 18 Graticule Map

Neatline

The neatline is considered to be the frame of the map and should encapsulate the map and map elements if needed. The goal of the neatline is to provide a nice, clean frame for the map to live within, and to separate the map from surrounding items on the medium. The neatline is used to direct user's eyes to the center of the map. Generally neatlines should not be visually dominant, but large enough so that the eye can use it as a frame. Review Figure 19.

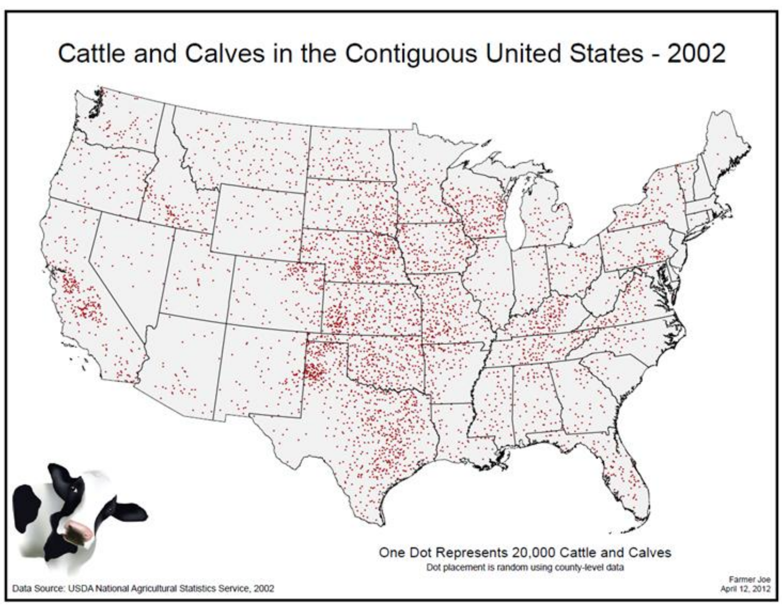


Figure 19 Example of Neatline

Ancillary Text and Objects

Ancillary text and/or objects are additional supporting information which provides a greater understanding of the topic of the map. A few examples of ancillary text or objects are text, pictures, sounds, movies, and graphs. A couple of common reasons to include ancillary text on a map are to indicate data manipulation pertinent to the interpretation of the map and to indicate special cases or missing data.

Map Elements: Best Practices

Best practices dictate that:

- Map elements should be positioned and sized in accordance with their importance. The most significant items should be roughly at the center of the map or placed at the top of the page. This is typically why the title is at the top of the page, and the map body is in the center of the page.

- Map elements should use as much space as possible within the neatline so that white space is reduced.

- Map elements should be placed around the map so that the map has a visual balance that is pleasing to the eye.

SUMMARY

In this lesson you learned about discrete and continuous spatial phenomena. You reviewed attribute types, and attribute data which are typically arranged in tables where a row is equivalent to one entity and a column is equivalent to one attribute or descriptor of that entity. You also learned about vector data and raster data models and how to differentiate between these formats. You gained knowledge about coordinate systems and map projections which is important for understanding how to define and visualize purposes. Lastly, you learned about the importance of map design and map elements. The best practices for map elements provided within the lesson outlines essential information to help you with map design.

ASSIGNMENTS

1. Lab: Reviewing the Basics of Geospatial Data

2. Quiz: Reviewing the Basics of Geospatial Data

Authoring Organization: Del Mar College
Written by: Richard Smith

Development was funded by the Department of Labor (DOL) Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant No. TC-22525-11-60-A-48; The National Information Security, Geospatial Technologies Consortium (NISGTC) is an entity of Collin College of Texas, Bellevue College of Washington, Bunker Hill Community College of Massachusetts, Del Mar College of Texas, Moraine Valley Community College of Illinois, Rio Salado College of Arizona, and Salt Lake Community College of Utah.

Lesson 2: Introduction to Geospatial Analysis

INTRODUCTION

This lesson is an introduction to geospatial analysis. You will learn about descriptive statistics including central tendency of the data, distribution of the data, and the shape of the data. Classification methods are explained and demonstrated with images and content. The lesson provides examples and explanations of table operations. You will also learn about creating graphic statistical models and the elements associated with them.

LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Explore data relationships using geospatial data.
2. Create simple data sets using a table operation method.
3. Classify quantitative data using a variety of statistical methods.
4. Create a scatter plot of data.
5. Analyze scatter plot data to produce a presentation of results.

LEARNING SEQUENCE

Required Reading	Read the following: • Online Lesson Material
Resources	View the following: • None
Assignments	Complete the following: • Lab: Introduction to Geospatial Analysis • Quiz: Introduction to Geospatial Analysis

INSTRUCTION

Descriptive Statistics

Descriptive statistics are quantitative descriptions of data that provide some basic descriptions of the data set. Cartographers use descriptive statistics to explore the character of the data. Descriptive statistics can describe things such as the central tendency of the data, the dispersion of the data, and the shape of the data.

Authoring Organization: Del Mar College
Written by: Richard Smith

Central Tendency

Central tendency of data describes the distribution of the data. There are six measures that we commonly used to describe the central tendency of a data set: maximum, minimum, range, mean, median, and mode.

- The **maximum** measure of central tendency simply reports the maximum value in a data set.

- The **minimum** measure of central tendency returns the minimum value in a data set.

- The **range** measures how high and low a series of numbers span. In order to determine the range you would identify the highest and lowest numbers in the series and then subtract the lowest number in from the highest. The range measure returns the maximum value minus the minimum value of a data set.

- The **mean** measure of central tendency is the average value of the data set. The average value is defined by summing all values in the data set and dividing the sum by the number of values in the data set.

$$\bar{X} = \frac{\Sigma x}{N}$$

- The **median** measure of central tendency reports the middle point of the data. For instance, if we had three observations in our data set, we would sort the observations in ascending order and then look at the value at the midway point which would be the second observation of this case. If there is an even number of observations in a data set, then the median will be the average of the two most central observations.

- The **mode** measure of central tendency reports the most common value found in the data set. If multiple values are tied as the most common value, the data set has multiple mode values.

Dispersion

Dispersion measures the variability of the data set. There are two measures of dispersion: variance, and standard deviation.

Variance

The variance measure of dispersion takes the sum of the squares of the deviations divided by the number of observations.

$$\sigma^2 = \frac{\Sigma(x - \bar{X})^2}{N}$$

The units of variance are identical to the original units of measure. So for instance, if the original units of the observations were in feet, then they variance will report how much the observations vary on average in feet.

Standard Deviation

The standard deviation is similar to the variance except that it takes the square root of the sum of the squares of the variance divided by the number of observations.

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{X})^2}{N}}$$

The standard deviation measures dispersion in a standard way so that two different data sets can be compared. Standard deviation does not report the dispersion in the original units of the observations.

Review the normal distribution of data provided which illustrates that most of the examples of the data set are close to the average value while relatively few examples tend to be positioned at one extreme or the other. The x-axis is the value in question and the y-axis is the number of observations for each value on the x-axis.

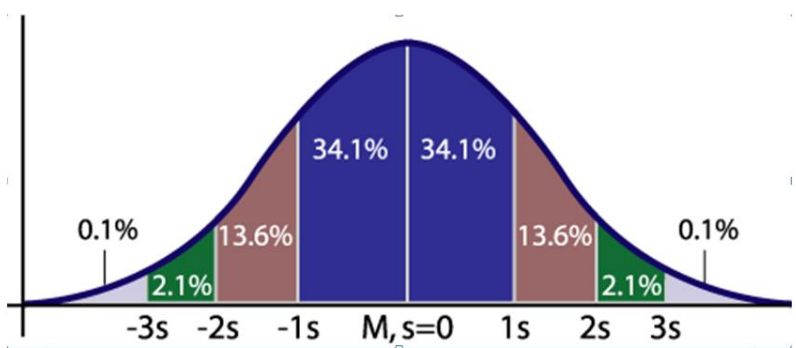


Figure 1 Normal Distribution of Data

The standard deviation tells us how tightly all the various observations are clustered around the mean of the data set. When the examples are pretty tightly bunched together, the bell-shaped curve is steep, and the standard deviation is small. When the examples are spread apart, the bell curve is relatively flat, which tells you that you have a reasonably large standard deviation.
One standard deviation away from the mean in either direction on the horizontal axis accounts for about 68% of the observations in the data set. Two standard

deviations away from the mean of the four areas closest to the center accounts for about 95% of the observations of the data set. Three standard deviations account for about 99% of all the observations of the curve.

Skewness

The skewness measure tells us whether the peak of a distribution is to one side of the mean or the other. If a data set has a negative skew, then the peak of the distribution is above the average, if a data set has a positive skew, the peak of the data set is below the average.

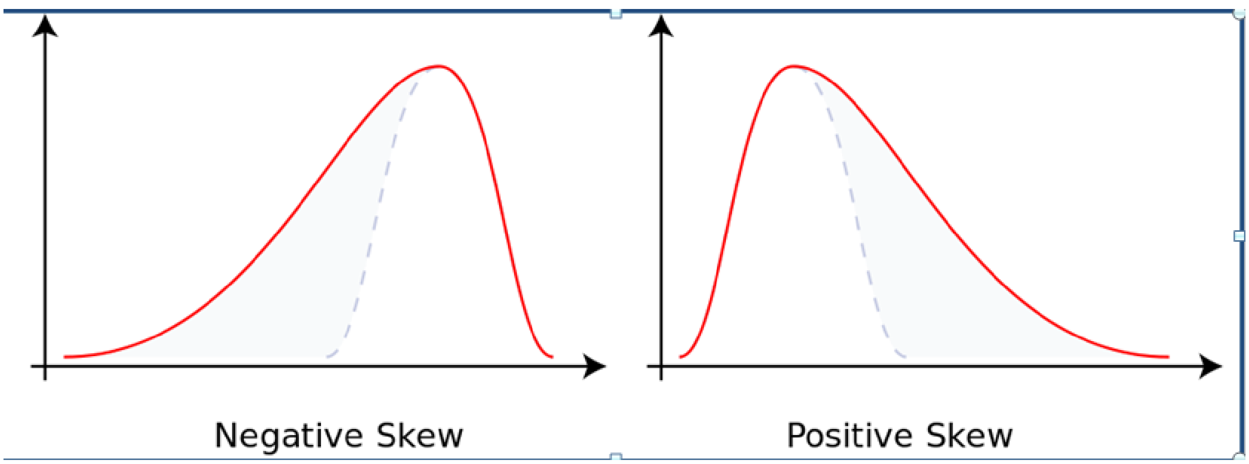


Figure 2 Negative and Positive Skew

Kurtosis

The final measure of central tendency is kurtosis. Kurtosis describes the flatness or peakedness of a distribution. For normal distribution, the kurtosis is equal to the value of 3.0. A normally distributed data set having the kurtosis value of 3.0 is a mesokurtic distribution. A value above 3.0 is a leptokurtic distribution. A value less than 3.0 is a platykurtic distribution.

Below is an illustration of the three types of kurtosis.

- The green line is the normal distribution with a value of 3.0 and is mesokurtic.
- The red line with a strong peak is the leptokurtic distribution and has a value greater than 3.0.
- The blue line which has a flat top is the platykurtic distribution which has a value less than 3.0.

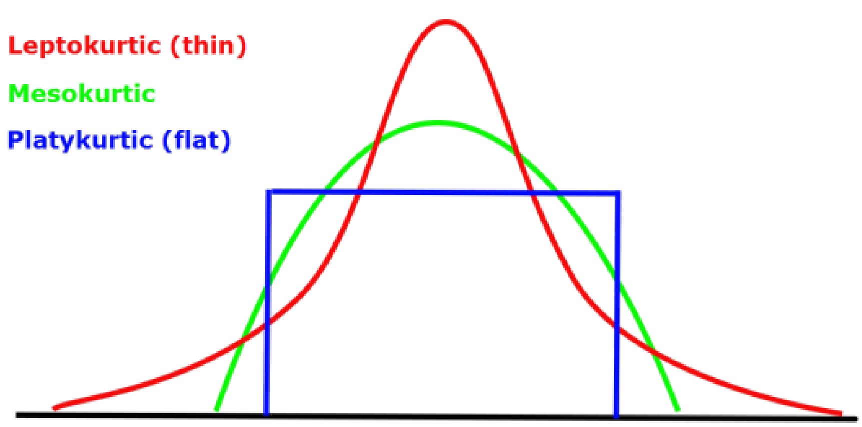


Figure 3 Three Types of Kurtosis

Data Classification

Data classification categorizes objects based on a set of conditions into separate bins or classes. Classification may add to or modify attribute data for each geographic object. For example, a classification could add a new code such as large or small.

Classification could also recode an attribute, such as changing urban to dense. One attribute can yield many different maps depending on which classification method is chosen. Different classification methods will have a direct effect on how the map and data are perceived by the map user, therefore much care must be taken when choosing a data classification method.

Classifying Data

The primary reason to classify data is to simplify the data for visual display. There are three goals for data classification with regards to cartography.

Goal 1: Simplify the visualization so that spatial patterns of distribution can become more easily viewable by the map reader.

Goal 2: Group similar observations together.

Goal 3: Show the difference between the groups.

Review the following examples (Figures 4-6) which illustrate how data can be classified.

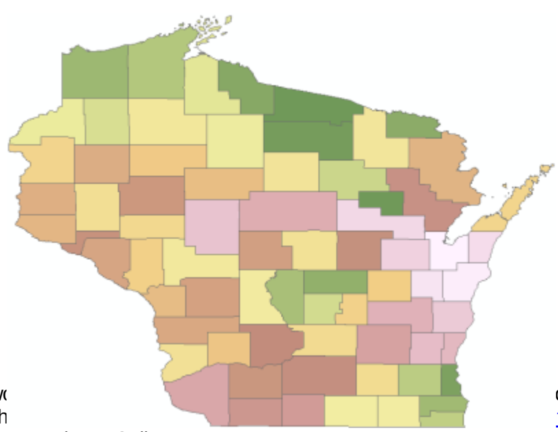


Figure 4 This map assigns each value a unique color, which means, that each county has its own unique color. It is very difficult to look for patterns when each county has its own unique color.

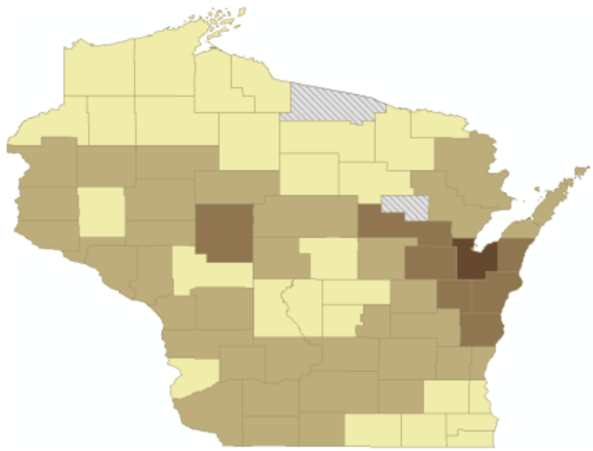


Figure 5 This map uses the same data, but uses a data classification method to categorize the counties to make it easier to visualize the data. The lighter values that represent less of an item and the darker values represent more of an item; spatial patterns of distribution begin to emerge.

Figure 6 This is an example of a classification showing countries with high, medium, and low values of population.

Classification Methods

It is important to be familiar with the following types of classification methods. Review each classification method outlined below.

Binary Classification Method

Binary classification places objects into two classes. The two classes can be the value of 0 and 1, true and false, or any other dichotomy that you can think of. Typically, a binary classification is used to store the results of complex operations when the operation returns either a yes or no answer.

In this example, the states were classified into two binary classes: one class for the northern states and one class for the southern states.

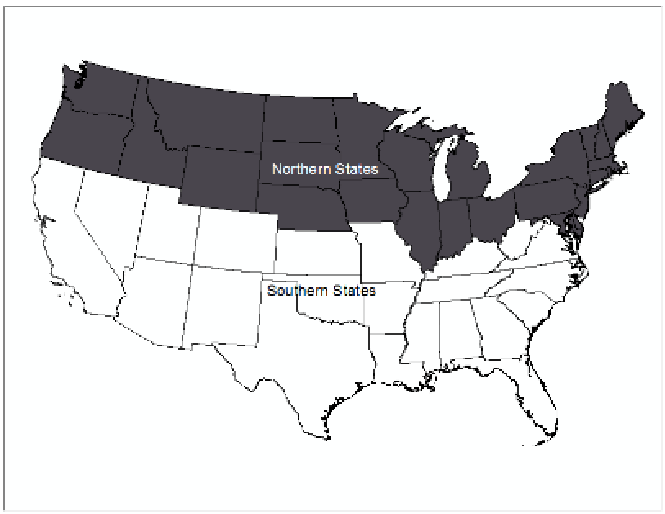


Figure 7 Binary Classification Method

Dissolve Classification Method

The dissolve method combines similar features within a data layer based on a shared attribute. For example, continuing the northern and southern states example, we can dissolve all the states based on whether they are a northern state or a southern state. The dissolve operation creates new geometry and in this case, two polygons. In regards to this pair of new polygons, one represents the combined extent of all states classified as northern, while the other represents the combined extent of all states classified as southern. Take note that the new geometry does not transfer any of the attributes of the dissolved.

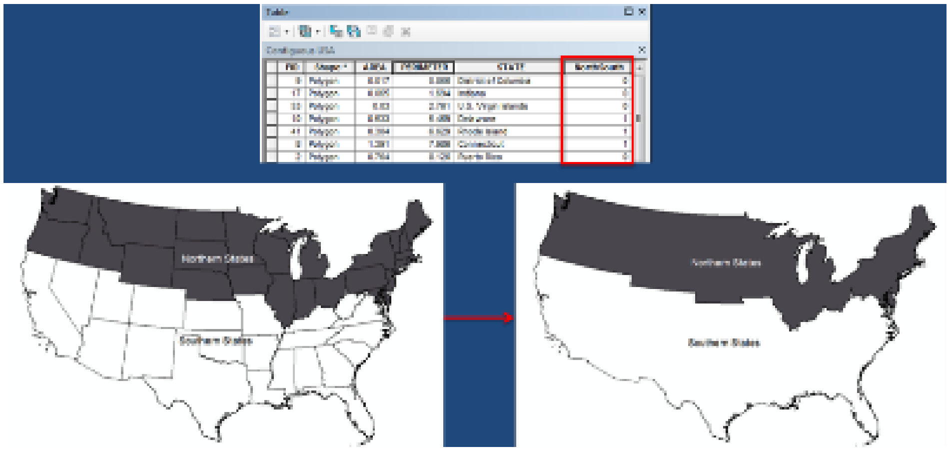


Figure 8 Dissolve Classification

Automatic Classification Method

An automatic classification method is where you set up rules for the computer to follow and then the computer executes the data classification. For example, let's pretend we are building a vending machine. In order for our machine to determine whether the user has inserted enough money, we need the machine to determine which type of coin was inserted. Therefore, we set up rules for the vending machine to sort the coins based on their diameter and width. With those rules in place, the vending machine will be able to sort the coins into different bins, thereby determining the amount of money the user inserted. This is an example of automatic classification. In a GIS, it is a similar operation. We tell the computer that we want data within a certain range of values to go into different bins, thereby sorting them and preparing them for visualization on a map.

Jenks Natural Breaks Classification Method

The Jenks Natural Breaks classification method aims to maximize homogeneity in classes. It uses breaks in a histogram as class breaks and assumes that group data are alike. Review Figure 6, which includes an example of the Wisconsin map, classified using the natural breaks method. The map displays the result of the classification visually and shows the five classes varying from a light brown to a dark brown in a sequential color. This map is interesting to look at, and displays interesting patterns in the data. The histogram illustrates the values where the x-axis is the observation value, and the y-axis is the number of observations for each value. The vertical blue lines show the extent of each class and are considered to be the class breaks. Looking at the histogram and the blue lines, you can see that natural breaks tend to place the class breaks where there are natural valleys or vacancies in the data set.

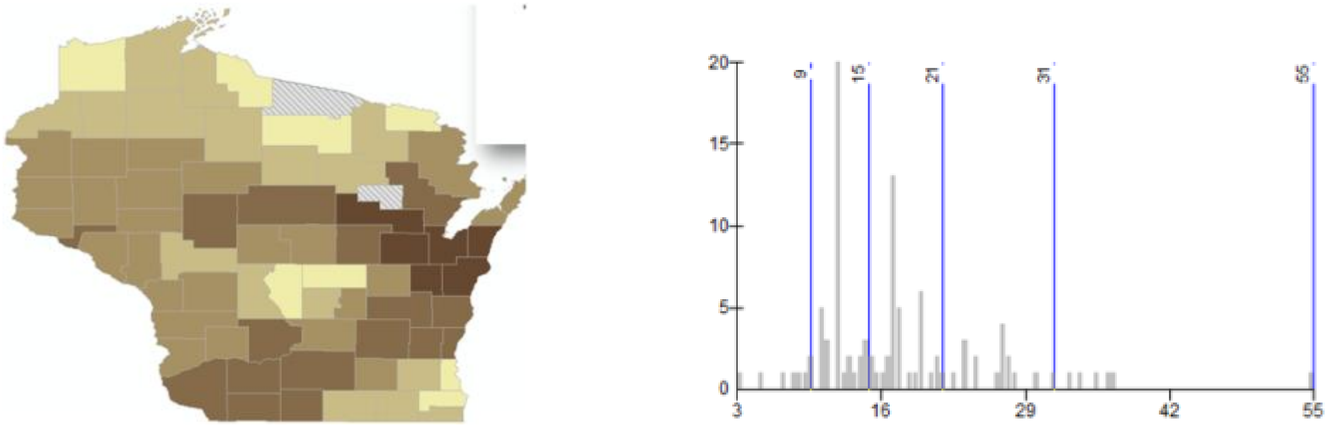


Figure 9 (left) Jenks Natural Breaks, (right) Data Set

Consider the following advantages and disadvantages of the Jenks Natural Breaks classification method.

Advantages

- Considers distribution of data
- Minimizes in-class variance
- Maximizes between class variance
- Produces classification with high accuracy

Disadvantages

- Complicated
- Difficult to understand procedure for grouping

Nested Means Automatic Classification Method

The nested means automatic classification method creates classes about the arithmetic means of the data set. Additional means can be calculated about the first mean to create additional classes. The way the nested means works is at the first two classes are created 1 above the mean and 1 below the mean. The third and fourth classes are created above and below the means from the first two classes. In this case, it creates a reasonably interesting map, for each class has about the same number of observations.

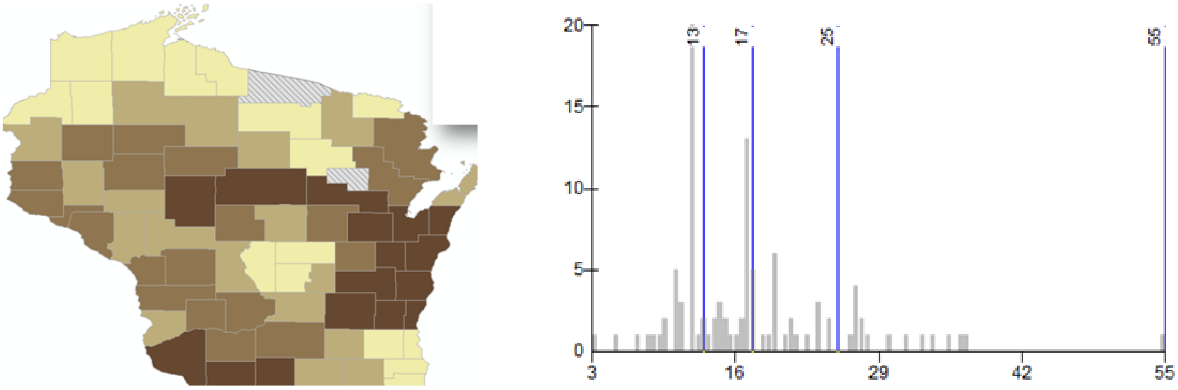


Figure 10 (left) Nested Means, (right) Data Set

Consider the following advantages and disadvantages of using the Nested Means classification method.

Advantages

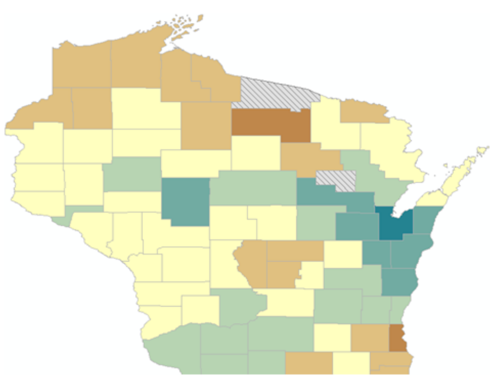
- Easily Computed
- Mathematically intuitive

Disadvantages

- Limited to classes: 2, 4, 8 ...
- Does not consider distribution of data

Mean and Standard Deviation Automatic Classification Method

The mean and standard deviation automatic classification method creates classes about the arithmetic mean and standard deviations above and below the mean. For this method, since we are looking at divergent behavior, you should use a diverging color. Figure 11 is an example of a diverging color map created using colors tending toward brown as being identified as observations below the mean, and observations towards blue being above the mean. The data for the state has a positive skew, it is not normally distributed; therefore, this method is not ideal for this data set, but does still tend to display interesting geographic patterns.



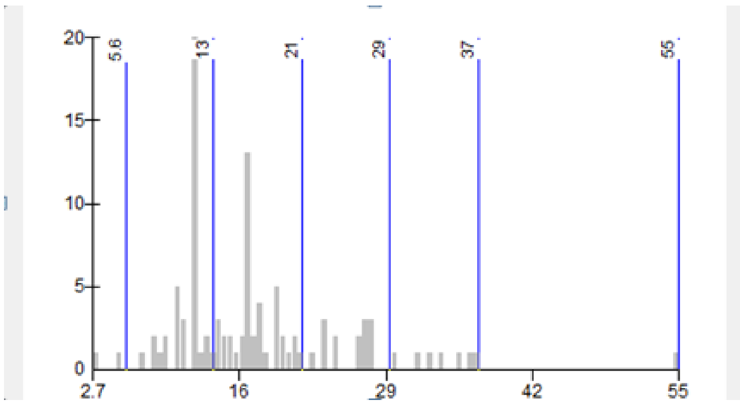


Figure 11 (left) Diverging Color Map, (right) Data Set

Consider the following advantages and disadvantages of using the Mean and Standard Deviation classification method.

Advantages

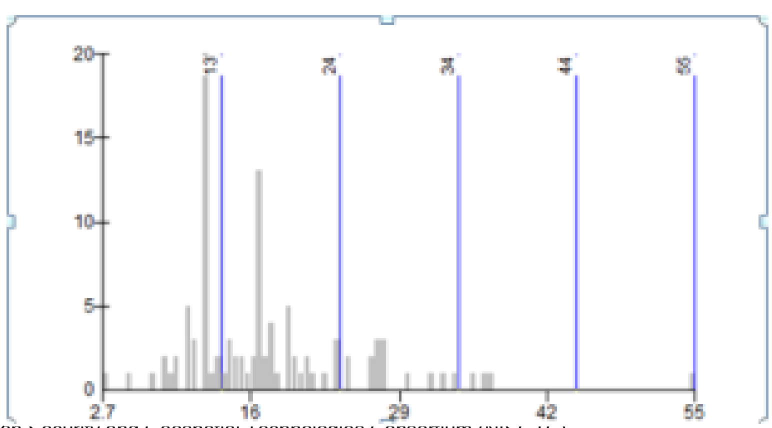
- Good for data with normal distribution
- Considers distribution of data
- Produces constant class intervals

Disadvantages

- Most data are not normally distributed
- Requires understanding of statistics
- Not easily understood by map reader

Equal Interval Automatic Classification Method

The equal interval automatic classification method creates classes with equal ranges. The class range is calculated by taking the maximum value of the data set, subtracting the minimum value from it, and then dividing that by the number of observations in the data set. interesting patterns are now hidden since the class ranges are including data that is different from each other. Review Figure 12. Looking at the map you can see interesting patterns are now hidden since the class ranges are including data that is different from each other. In the histogram of the data, the class breaks are evenly distributed which is the goal of the equal interval classification method.



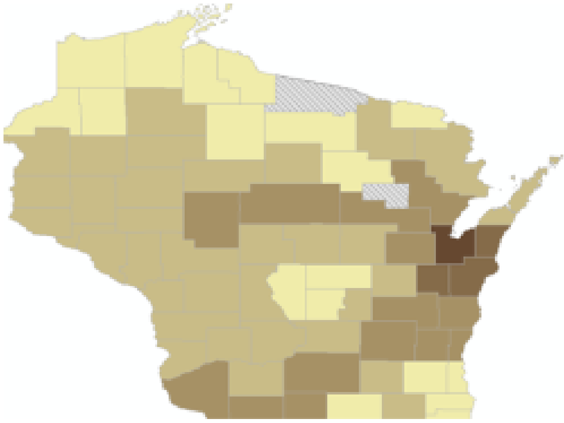


Figure 12 (left) Equal Interval, (right) Data Set

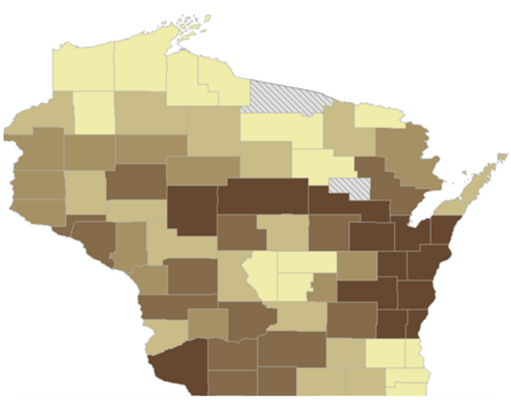
Consider the following advantages and disadvantages of using the Equal Interval classification method.

Advantages
- Easily understood
- Simple to compute
- No gaps in legend

Disadvantages
- Does not consider distribution of data
- May produce class with zero observations

Equal Frequency Automatic Classification Method
The equal frequency automatic classification method is also known as the quantile classification method. The goal of the equal frequency method is to distribute observations equally among classes, that is, each class will have the same number of observations. If the number of observations does not divide equally into the number of classes, then you should place the extra observations across the lower classes, thereby overloading the lower classes.

As the equal frequency automatic classification method aims to have the same number of observations in each class, there should be about the same amount of each color on the map, which permits a balanced looking map. Again, a major negative, is that a class may contain data values that are dissimilar, which goes against the purpose of classifying data.



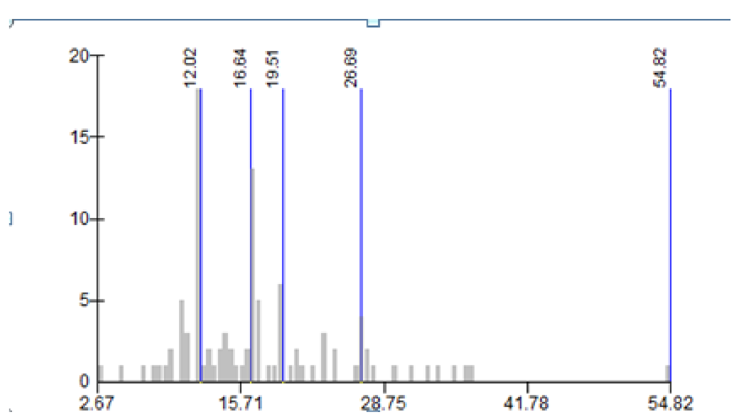


Figure 13 (left) Equal Frequency Map, (right) Data Set

Consider the following advantages and disadvantages of using the Equal Frequency classification method.

Advantages

- Easily calculated
- Applicable to ordinal data
- No classes with zero observations

Disadvantages
- Does not consider distribution of data
- Gaps in legend
- Observation value may be closer to value in a different class

Arithmetic and Geometric Intervals Automatic Classification Method
The arithmetic and geometric intervals automatic classification methods create class boundaries that change systematically with a mathematical progression. This classification method is useful when a range of observations are significant and the observations follow some sort of mathematical progression that can be followed with the classes. Take a look at Figure 14. On the map, this classification method creates a reasonably interesting map. Additionally, it does not over fit the data but in this case does seem to over fit the data creating patterns where they do not exist. If we look at the histogram, you can see that the class ranges slowly increase along the x-axis as this classification method is fitting the data in increasingly larger classes.

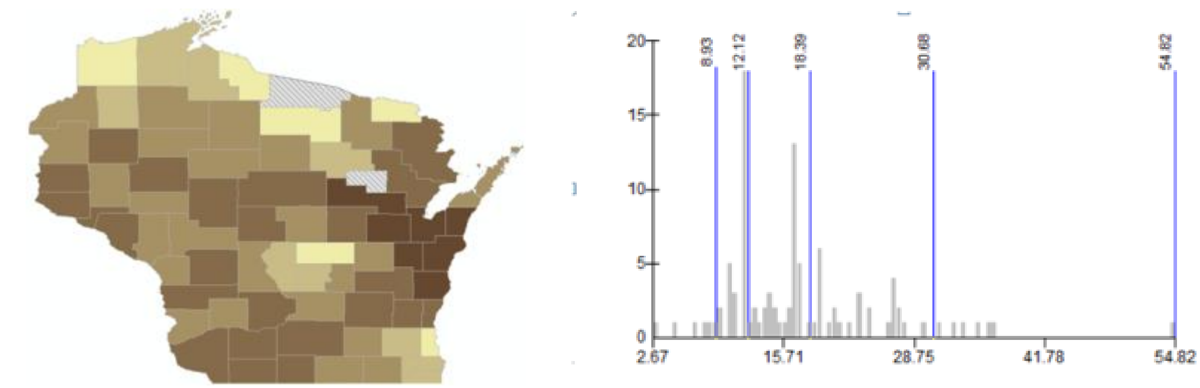


Figure 14 (left) Arithmetic and Geometric Intervals map, (right) Data Set

Consider the following advantages and disadvantages of using the Arithmetic and Geometric Intervals classification method.

Advantages

- Good for data with large ranges
- Break points determined by rate of change in the data set

Disadvantages
- Not appropriate for data with small ranges or linear trends

Common Table Operations

Combining tables of attributes is a powerful analysis technique. There are three common types of table operations: intersect, union, and join. Each one of these operations combines tables in different ways, and for different purposes.

It is important to note that in order for multiple tables to be inputs for intersection or union operations, the tables must have the same fields and field types, which are more commonly referred to as the schema. Let's take a look at each one of these three table operations in more detail.

Intersect

The intersect table operation produces an output table that only contains records that all of the input tables had in common. That means that every attribute for the entire row was identical. As an example (Figure 15), we have two input tables that have the same schema, as all input tables must have the same schema for the intersect operation. The output table only includes the two records that were found in both input tables. In this case, the row where the ID is 1, the region is A, and the salesman is Will, and in the second row the ID is 5, the region is W, and the salesman is Claire.

Input Table 1

ID	Region	Salesman
1	A	Will
2	D	Adam
3	A	Billy
4	M	Alice
5	W	Claire

Input Table 2

ID	Region	Salesman
1	A	Will
5	W	Claire
11	M	Jean

Intersect

Output Table

ID	Region	Salesman
1	A	Will
5	W	Claire

Figure 15 Output table only contains records that all input tables have in common.

Union

The union table operation produces an output table that contains all records that existed in all input tables. For example, we have the same two tables as before, and notice at the output table combines all the rows from both input tables. Again, note that both input tables have the same schema, as this is required for union.

Input Table 1

ID	Region	Salesman
1	A	Will
2	D	Adam
3	A	Billy
4	M	Alice
5	W	Claire

Input Table 2

ID	Region	Salesman
1	A	Will
5	W	Claire
11	M	Jean

Union

Output Table

ID	Region	Salesman
1	A	Will
2	D	Adam
3	A	Billy
4	M	Alice
5	W	Claire
11	M	Jean

Figure 16 Output table only contains all records that exist in all input tables.

Join

The join operation joins together two tables based on a shared attribute. Based on the shared attribute, attributes of one input table are appended to the attributes of the other input table. For example in Figure 17, there are two input tables (Input Table 1 and Input Table 2) that share the common attribute of region. Based on the region attribute, the manager and area attributes of Input Table 2 are appended to the rows of Input Table 1 in order to create the Output Table.

The region attribute, matches its value between the two tables, which determines which attributes of which row from Input Table 2is copied over. For example, if we look at the Output Table, the row with ID 1 had a region A, which means that the manager is John, and the area is 20 sq. miles. In row where the ID was 2, since the region was the value of D, then the row with the same region of value D in the second input table is copied to the output table.

Also note, that for the record where the ID was 3, the region is A again, so once again, the manager John, and the area 20 are copied to the output table for that row.

Input Table 1

ID	Region	Salesman
1	A	Will
2	D	Adam
3	A	Billy
4	M	Alice
5	W	Claire

Input Table 2

Region	Manager	Area (sq. mi.)
A	John	20
D	Felicia	25
W	Alan	16
M	Seneca	19

Join

Output Table

ID	Region	Salesman	Area (sq. mi.)
1	A	Will	John
2	D	Adam	Felicia
3	A	Billy	John
4	M	Alice	Seneca
5	W	Claire	Alan

Figure 17 Based on a shared attribute, attributes of one input table are appended to attributes of the other input table.

Graphic Statistical Models

A graphic statistical model is another form of cartography that provides visualization for analyzing data in a form other than a map. Graphic statistical models are excellent at allowing the user to identify trends in the data.

Additionally, graphic statistical models provide an insight into the distribution of the data over the entire data set. There are many different graphic statistical models available for use, but here we will only cover the four most common, which are scatterplots, bar graphs, vertical line graphs, and histograms.

Scatter Plots

A scatter plot is a good method of viewing relationships between variables. A scatter plot allows the reader to identify relationships between two variables where each variable is on its own axis. With both variables on separate axis, relationships are visible by a pattern in the plot which is commonly known as a trend and is traced using trendlines.

If there is a trend inside the data set, its scatter plot will visually describe it for you. It may also describe no relationship as the data seems randomly scattered. Figure 13 is an example of a scatter plot showing the GDP per capita of the United States in the year 2009 and the economic complexity index of the United States in year 2008. As we can see in the scatter plot, as the GDP increases, so does the economic complexity index. This is a good example of positive correlation between the two variables. Also, since there seems to be a trend in this data set, you could draw a trend line starting at the bottom left through the center of the data points to represent the general trend of the data.

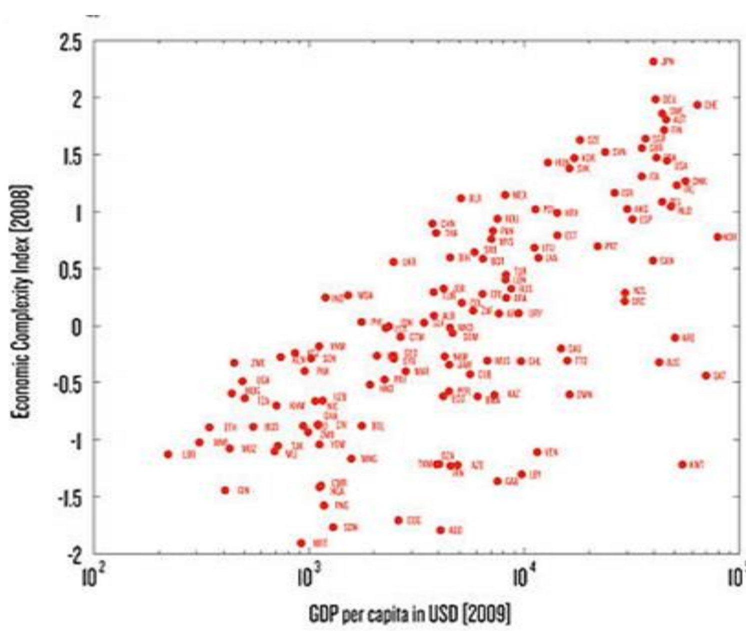


Figure 18 Example of a Scatter Plot

Bar Graph

A bar graph shows rectangular bars that are in proportion to the values they represent. This allows for comparison between categories of items in a nice visual form. On a bar graph, one axis shows the categories of the items while the other shows a discrete value for each one of those items.
Figure 19 is an example of a bar graph showing the wind generated electricity as a percentage of a state's energy generation for the year 2011. On the vertical axis we see the name of each item, in this case state names. On the horizontal axis, we see numbers that represent percentages. For each state, a bar is drawn that is proportional in size to the percentage value of a state's wind generated. With all of the states' bars drawn, the bar graph presents the user with a nice visual medium for comparing the percentages between states. It should be noted that while this bar graph is ordered in descending order from top to bottom, this is not required.

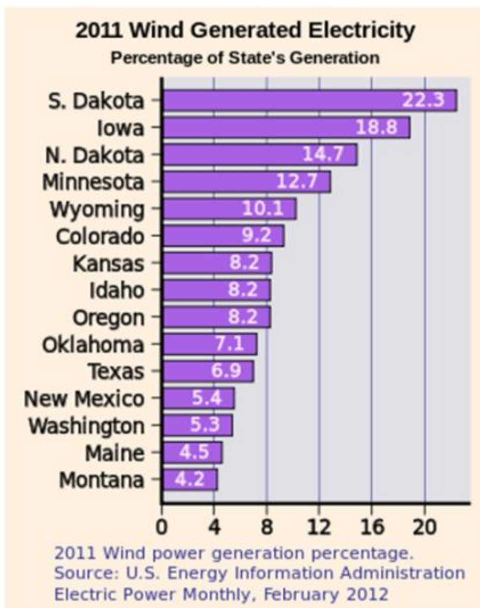


Figure 19 Bar Graph

Vertical Line Graph

A vertical line graph displays information through plotting a series of data points that are then connected with lines. Vertical line graphs are excellent for visualizing trends within a data set and are very simple to create and interpret. In this vertical line graph (Figure 20), on the vertical axis are temperature values between 80 and 115. On the horizontal axis, the 12 months of the year are represented. For each month, a dot is placed above that month at the location where the temperature was recorded. This is done for all 12 months. Next, a line is drawn between all 12 dots to show continuity.

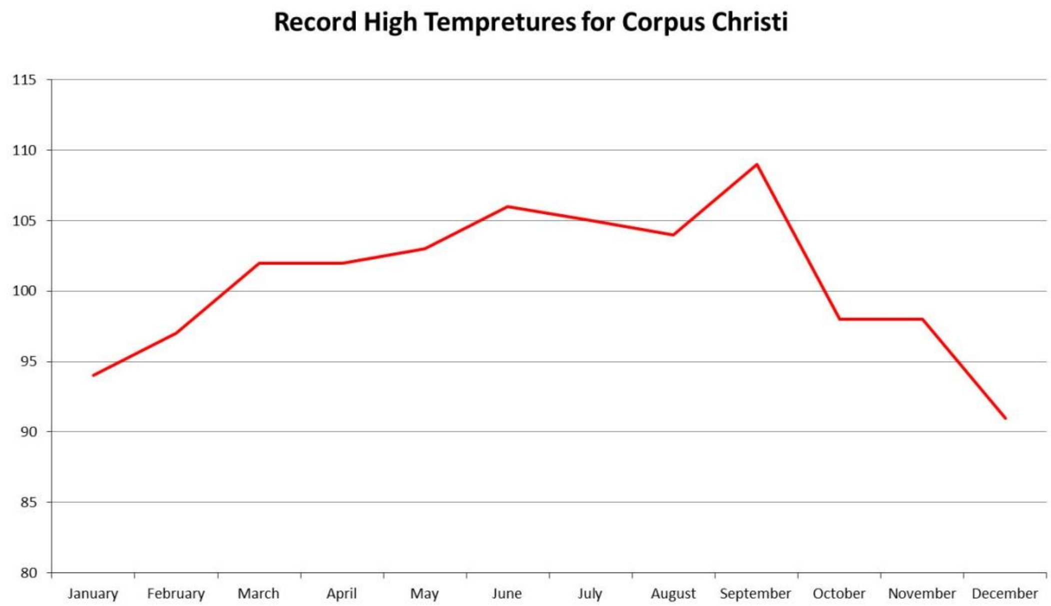


Figure 20 A vertical line graph of record high temperatures in Corpus Christi, Texas.

Histogram

A histogram is used to plot the density of data. On a histogram, for each range, or value of ranges, a bar is extended for each observation that falls on the value or range of values. A histogram allows the user to see the frequency of observations in relation to specified ranges within a dataset.

Figure 21 is an example of a histogram showing the percentages of 2000 Models of Automobiles at Age 5-to-6 Years with the Consumer Reports overall reliability rate. On the vertical axis, we see the percentage value. On the horizontal axis, we see five qualitative measures of reliability. The height of the bars varies depending on how many values fall within each measure of reliability. As can be seen here, the majority of values fell within the average reliability rating.

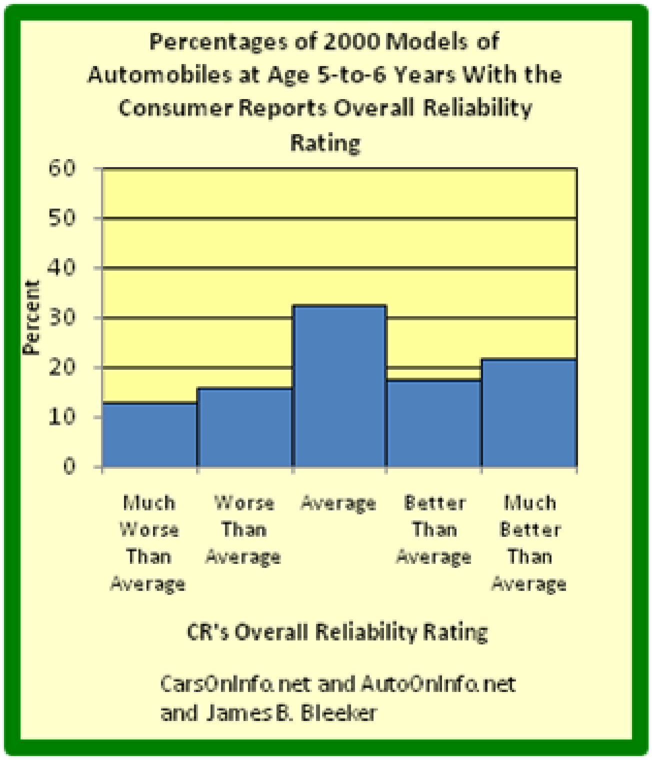


Figure 21 Histogram

SUMMARY

This lesson introduced you to geospatial analysis. You learned about descriptive statistics including central tendency examples. Standard deviation was explained as were the three types of Kurtosis. The lesson explained how to classify data and the methods used in this process. There were many examples of equal intervals, equal frequency, and arithmetic and geometric intervals. Table operations were covered in this lesson as was creating graphic statistical models and the various types.

ASSIGNMENTS

1. Lab: Introduction to Geospatial Analysis
2. Quiz: Introduction to Geospatial Analysis

Authoring Organization: Del Mar College
Written by: Richard Smith

Development was funded by the Department of Labor (DOL) Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant No. TC-22525-11-60-A-48; The National Information Security, Geospatial Technologies Consortium (NISGTC) is an entity of Collin College of Texas, Bellevue College of Washington, Bunker Hill Community College of Massachusetts, Del Mar College of Texas, Moraine Valley Community College of Illinois, Rio Salado College of Arizona, and Salt Lake Community College of Utah.

This workforce solution was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration. The solution was created by the grantee and does not necessarily reflect the official position of the U.S. Department of Labor. The Department of Labor makes no guarantees, warranties or assurances of any kind, express or implied, with respect to such information, including any information on linked sites, and including, but not limited to accuracy of the information or its completeness, timeliness, usefulness, adequacy, continued availability or ownership.

Lesson 3: Using Attribute and Spatial Queries for Data Exploration

INTRODUCTION

In this lesson you will learn about using attribute and spatial queries for data exploration. Information presented includes set algebra and Boolean algebra. Data dictionaries and deciphering coded values are explained and exhibited in the lesson. Buffering and dissolving are discussed as well as the importance of each as it relates to spatial queries and data exploration.

LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Perform advanced query to prepare data for use in analysis.

2. Use a data dictionary to decipher coded data.

3. Determine how to use queries to address a question.

LEARNING SEQUENCE

Required Reading	Read the following: • Online Lesson Material
Resources	View the following: • None
Assignments	Complete the following: • Lab: Using Attribute and Spatial Queries for Data Exploration • Quiz: Using Advanced Attribute and Spatial Queries for Data Exploration

INSTRUCTION

Using Attribute and Spatial Queries for Data Exploration

Attribute queries are an extremely common GIS aspatial operation. Attribute queries select a subset of records based on values of specific attributes. Each attribute query must specify three things: an attribute field, a set algebra operator, and an attribute value. An example of an attribute query would be if we had a data set of land parcels for sale.

If we were interested in selecting parcels that are at least six hundred acres in size, our attribute query would be: Acres greater than six hundred where 'Acres' is the attribute field, 'greater than' is the set algebra operator, and 'six hundred' is the attribute value we wish to evaluate.

Attribute queries can also select records based on multiple attributes combined together using Boolean operators, such as 'and', 'or', and 'not'.

Set Algebra Operations

Set algebra uses operations to determine whether two values are equivalent or not. The four basic set algebra operations are less than, greater than, equal to, and not equal to.

Set Algebra Operations	
<	Less than
>	Greater than
=	Equal to
< >	Not equal to
> and < may not be applied to nominal attributes	

< Less than: The less than operation checks to see whether the value on the left is less than the value on the right. The less than operation is represented by the left angle bracket symbol.

> Greater than: The greater than operation checks to see where the value on the left is greater than the value on the right. The greater than operation is represented by the right angle bracket symbol.

= Equal to: The equal to operation checks to see whether the values on both sides are equal to each other. The equal to operation uses the =.

< > Not equal to: The not equal to operation checks to see if the values on both sides are different from each other, and is equivalent to the combination of less than and greater than. The not equal to operation is represented by both the left angle bracket and right angle bracket used together.

Applied Alone or in Combination

The three operations: greater than, less than, and equal to can be applied alone, or in combination. So for instance, you can perform the test of whether the value in the left side is less than or equal to the value of the right side. The symbol that would represent this operation, would be both the left angle bracket followed by the =. The results of all of these operations are either the value of "true", or "false".

As a quick challenge question, I am making the claim that greater than and less than may not be applied to nominal attributes. Why do you think this is the case? The answer is because nominal attributes are only descriptors, and it is illogical to compare them with respect to magnitude or rank. It is, however, logical to compare them using the equal to, or not equal to operators. Let's look at some examples of table queries using simple set algebra operations.

If we consider this attribute table representing different buildings, their square footage, number of floors, use, and zone, we can perform some simple attribute queries to select a subset of these buildings. For example, if we use the attribute query floors greater than 1, buildings A, B, D, and E would be selected as their floors attribute are all greater than the number 1.

Attribute Table: Floors > 1

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

If we change our attribute query to now read floors less than or equal to 2, then buildings A, B, C, E, and F are all selected, leaving building D as the only building not selected. This is because all the selected buildings have an attribute value of 2 or less for the floors attribute.

Attribute Table: Floors <= 2

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

As a final example, if our attribute query reads floors not equal to 2, then only buildings C, D, and F are selected, as those buildings have a floor attribute that is not the number 2.

Attribute Table: Floors <> 2

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

Boolean Algebra

Boolean algebra has multiple conditional operators. The three most common Boolean operators are:"OR", "AND", and"NOT". The Boolean operators evaluate values on the left and right side of the operator, and then assign an outcome.

The outcome is a Boolean, or binary result such as true/false, 0/1, on/off, or any other dichotomy. In Boolean algebra, order of operations do matter, therefore it is not uncommon to use many sets of parentheses to force a particular order of operations. It is also important to note that Boolean operators are not distributable inside of parentheses.

AND Boolean Operator

For the "AND" Boolean operator, the queries to the left and to the right of the "AND" Boolean operator must both evaluate to true for the entire statement to be considered true. If one of the two, or both criteria, does not evaluate to true, the entire state will return false, and therefore, the record will not be selected. The only way the record will be selected using the "AND" Boolean operator is if both queries evaluate to true.

Consider an example using the buildings attribute table provided. Note that our statement has two queries combined with an "AND" Boolean operator. The left query selects all buildings where the zone attribute is equal to commercial. The right query selects all buildings where the square feet attribute is greater than 40,000. The "AND" Boolean operator will only select a record if a building is owned to commercial and has a square footage greater than 40,000. If either of these two queries is not met, then the record is not selected. Therefore, based on this query, building C is the only building that is both zoned commercial and has a square footage greater than 40,000 feet.

Attribute Table: Zone = Commercial AND Sq. Ft > 40,000

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

OR Boolean Operator

The "OR" Boolean operator again considers the results of two queries, one on each side. In order for the entire statement to be considered true, at least one of the attribute queries must return true. If only one of the attribute queries returns true, the record will still be selected as the entire statement is considered true. If all of the attribute queries return true, then the record will still be selected as the entire statement is considered true. However, if neither of the attribute queries return true, in other words they both return false, then the record will not be selected. Therefore, the only way a statement is considered false using two attribute queries and it or Boolean operator, is if both attribute queries return false.

Consider the statement that reads floors > 1 OR use = department store. This statement will select all records where the building has more than one floor, or the building is used as a department store. Therefore, building F is the only record not selected because it has one floor and is used as a grocery store and does not meet either of the two queries on each side of the "OR" Boolean operator.

Attribute Table: Floors > 1 OR Use = Department Store

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

NOT Boolean Operator

The "NOT" Boolean operator selects attributes that do not meet the attribute query following the "NOT" Boolean operator. In other words, the attribute query following the "NOT" Boolean operator will have its evaluated value switched. This means that attribute queries that evaluate to true, will be switched to false and vice versa.

Consider the query "NOT" zone equals residential. Parentheses placed around zone equals residential help clarify the Boolean operator and the attribute query. For the selection, each record is evaluated on the zone attribute to see if it is equal to residential. If the zone equals residential, it will return a value of true. When we apply the "NOT" Boolean operator to this query, we would select all records not returned for the initial query or those records that had values not equal to residential Therefore, you could read the statement as select all buildings where zone is not equal to residential. Based on this query, buildings A, C, D, and F are selected as none of them are zoned residential.

Attribute Table: Zone NOT Residential

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

Set and Boolean Algebra Practice

Now it is time to practice. Complete Situations 1 and 2.

Situation 1: Take a few moments to determine what an appropriate attribute query would be to look for a home with more than 2,000 square feet using the Attribute Table: Buildings A – F.

Attribute Table: Buildings A – F

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

Situation 1 Answer

Situation 1: Looking for a home with more than 2,000 square feet.

Based on the situation where we want to find a home with more than 2000 square feet and appropriate attribute query we query on two different fields, use, and square feet. As we want both attributes to be true in order to select the record, we use the "AND" Boolean operator. Therefore, our attribute query reads use equals home "AND" square feet are greater than 2000. This selects building B as it is a home and has more than 2000 feet. It does not select record E, because even though the use is home, the square footage is exactly 2000 and is not greater than the requested 2000 ft.^{2}

Use = Home AND Sq. Ft > 2,000

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

Situation 2: Take a few moments to determine what an appropriate attribute query would be to look for a commercially zoned building that has more than 10,000 square feet and more than one floor. Use the provided Attribute Table: Buildings A – F.

Attribute Table: Buildings A – F

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

To see how well you did on Situation 2 scenario go here.

Situation 2: Answer

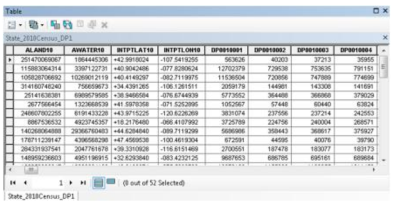
For this attribute query, we are using 3 attribute queries on 3 different attribute fields, and two Boolean operators. The query used here to satisfy the situation is Zone = Commercial AND Square Footage > 10,000 OR Floors > 1. To evaluate this attribute query we must keep in mind the order of operations for Boolean operators. "NOT" is evaluated before "AND" is evaluated before "OR". Once true or false values have been evaluated for these queries, these two values should be evaluated according to the rules associated with the "AND" Boolean operator. Once this true or false value has been determined, it should be placed to the left of "OR" floors greater than 1. Buildings C and F are the only two building selected as a both meet the requirements of the zone equaling commercial and square footage greater than 10,000. Note that building C and F only have one floor, but since the attribute queries surrounding the "AND" Boolean operator evaluated to true, even if floors evaluates to false, the record is still selected. In other words, no building meets every single attribute query in this statement. The C and F buildings still meet the criteria set forth in the statement.

Zone = Commercial AND Square Footage > 10,000 OR Floors >

Building	Sq. Ft	Floors	Use	Zone
A	75,000	2	Medical	Hospital
B	4,000	2	Home	Residential
C	50,000	1	Department Store	Commercial
D	100,000	3	Medical	Hospital
E	2,000	2	Home	Residential
F	20,000	1	Grocery Store	Commercial

Data Dictionaries

Often times an attribute table's field names are coded, having a system set up for the title of each attribute. For example, in a census data set, the field name DP0010001 represents total population. Naturally, DP0010001 is not a self-evident field name; therefore, we needed dictionary to decipher its meaning. The purpose of the data dictionary is to provide the descriptive field names for attribute tables. The data dictionaries are often provided in a text file format (.txt) or Microsoft Excel format (.xls).



As an example, here is the state census data for the 2010 United States Census. Looking at the attributes field names, they are all coded and do not provide easily identifiable information.

So the natural question is why do they do this?

The reasons why they use these sometimes hard to decipher field names are:

- Sometimes titles of fields can be too long to fit into the allotted space for that particular data. Additionally, a proper explanation can be made in the space allotted.
- It is sometimes better to create a symbolized system which puts the coded value for the field name, and be cross-referenced to a data dictionary which will explain the field in more detail. This also makes it easier for querying. For example, if you are searching for the total population of females, it is easier to specify DP0020003 in your query.

Deciphering Coded Values

To decipher the coded values, you should cross-reference the attribute table's field name with the data dictionary to find its definition. For example, in this image, the field name DP0010021 represents males under the age of five as found in Excel file serving as the data dictionary.

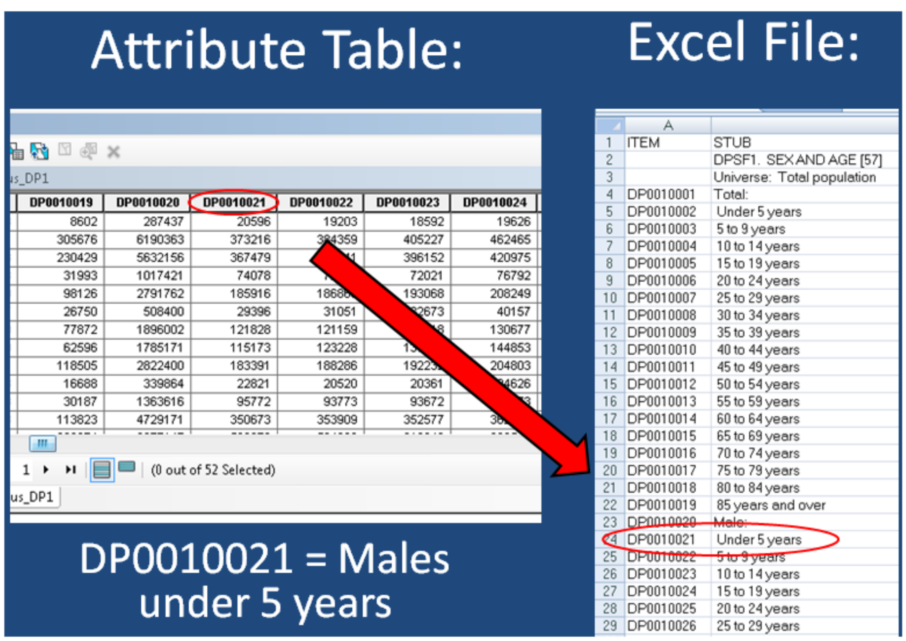


Figure 1 Deciphering Coded Values

Spatial Queries

The spatial query is where features are selected based on their location relative to other features. For example, let's say there was a power outage and you digitized the area with no power. You could use that digitized polygon to select all the facilities within this digitized area to perhaps dispatch utility crews. Much like attribute queries, you can combine multiple spatial queries to construct more complex queries. For the remainder of this section, specific types of spatial queries will be discussed.

"intersect"

The "intersect" spatial query selects any features that geometrically share a common part with the source feature. For the examples, the light red squares serves as the source, selecting feature, and the points, white lines, and white fill polygon are the target features available for selection.

In the image, Examples of Intersections, all three white lines, both points, and the white polygon are selected. All six of these features touch in some way within the light red polygon. In the image, Examples of Non-Intersection, the point and lines to the polygon do not touch the source light red polygon; none of them are selected as no intersections exist.

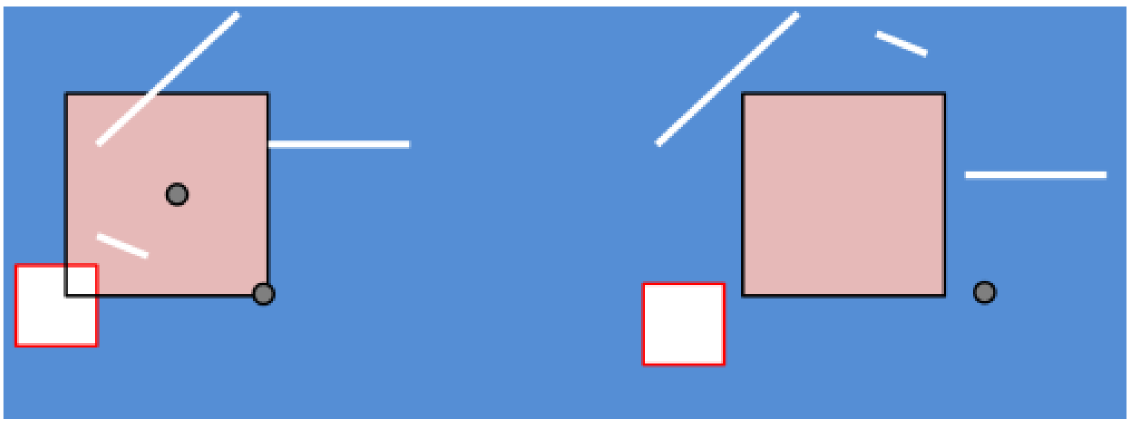


Figure 2 (left) Examples of intersections, (right) Examples of non-intersection

"are within a distance"

The next spatial query is "are within a distance". For this particular spatial query, it is necessary to specify a distance from the source feature in order to create a buffer. The query then selects all features intersecting the buffer.

For example in the image, Examples of Selection within a Distance, the dashed line represents the buffer at a set distance from the light red source polygon. The two lines, point, and white fill polygon are all selected as they all touch the buffer in some way. In the image, Examples of Selection not within a Distance, we see examples where the point, line and polygon are not selected as they do not intersect with the buffer.

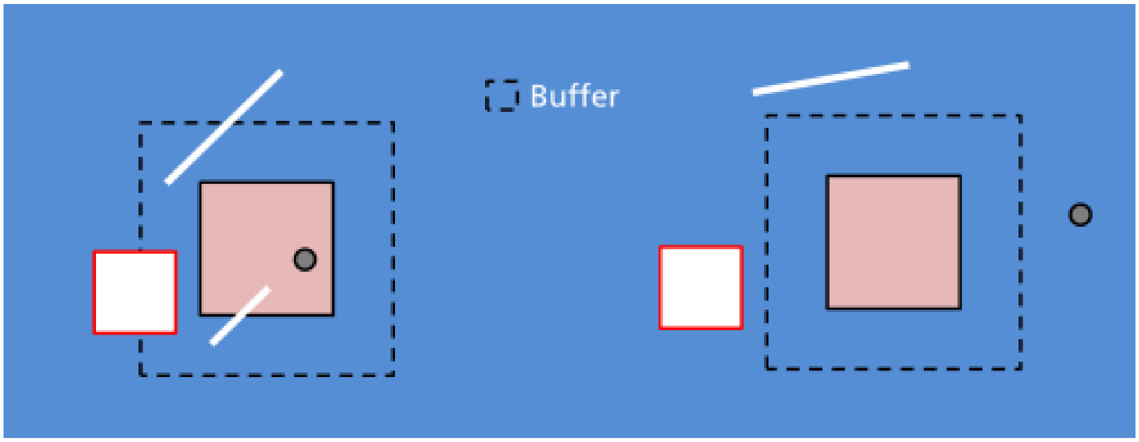


Figure 3 (left) Examples of selection within a distance, (right) Examples of selection not within a distance

"completely contain"

For the "completely contain" spatial selection, each point in the geometry of the target features must fall inside the source feature, including its boundaries. This is also considered to be when the source or selecting feature completely contains another feature.

The image, Example of Completely Contained Section, illustrates how the light red source feature completely contains the point, white line, and white fill box. In the image, Example of Not Completely Contained Section, none of the points, lines, or polygons are completely contained within the source polygon, therefore none of them are selected.

It is important to note the white line running across the top of the border of the source polygon, and the point on the bottom left edge of the source polygon. As both of these features exist on the boundary of the source polygon, they are not considered to be contained within, and are not selected.

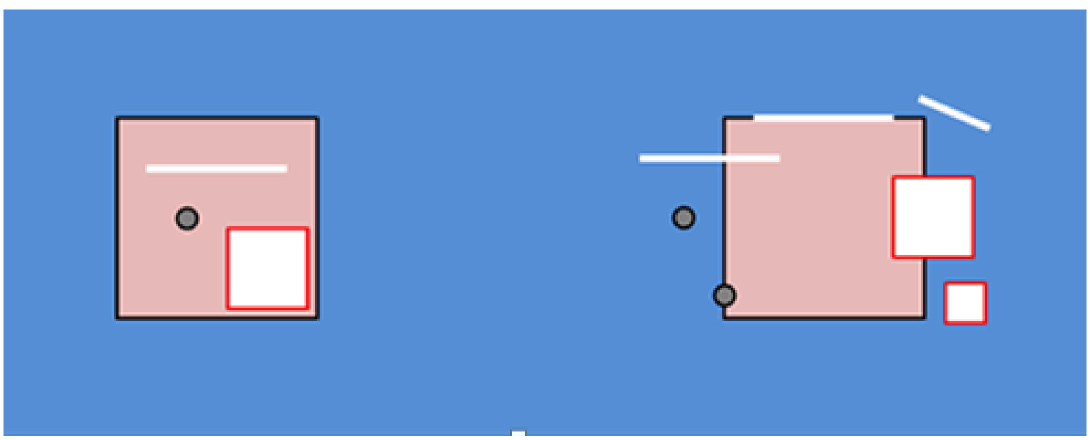


Figure 4 (left) Example of completely contained section, (right) Example of not completely contained section

“are completely within”

The next spatial selection is “are completely within”. For this spatial selection, target features must fall within the geometry of the source feature. This is the reverse operator of completely contain. For example the image, Example of Completely within Selection, the point, line, and white fill polygon all are completely contained within the light red source polygon, and therefore they are selected. In the image, Example of a Not Completely within Selection, none of the target features are completely within the light red source polygon, therefore none of them are selected.

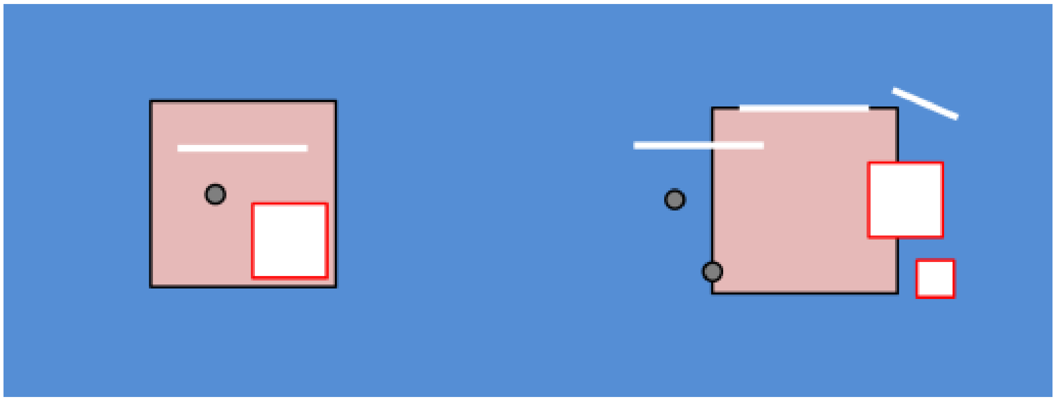


Figure 5 (left) Example of completely within selection, (right) Example of a not completely within selection

“have their center in”

“Have their center in” selects all features whether their centers fall inside the geometry of the source feature. In our examples, the center of each feature is denoted by a red X. In the image, Example of have Center in Selection, it illustrates how all three lines, point, and white fill polygon, have their centers falling inside of the light red source polygon, therefore they were all selected. In the image, Example of not having center in the selection, none of the target features have their center inside the source polygon; therefore none of them are selected.

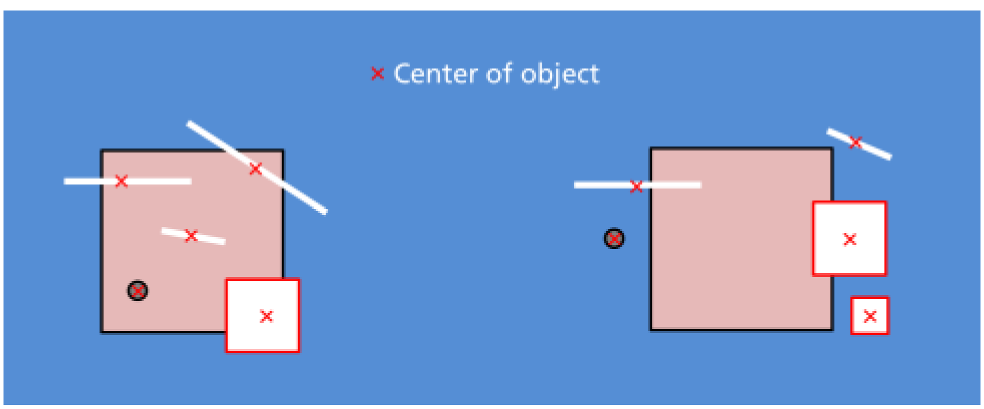


Figure 6 (left) Example of have center in selection, (right) Example of not having center in selection

"share a line segment with"

The next spatial selection operator is "share a line segment with". For this spatial selection operator, the geometries of the source and target features must have at least two contiguous vertices in common in order for a selection to take place.

The image, Example of Share a Line Segment With, depicts a scenario where all features would be selected because the two lines and the white fill box share line segments with the light red source polygon. In the image, Example of Do Not Share Line Segment With, none of the features are selected, even though the features do have one vertex falling on the boundary of the source polygon, remember that they must share least two contiguous vertices, and simply touching at one point does not cause a selection.

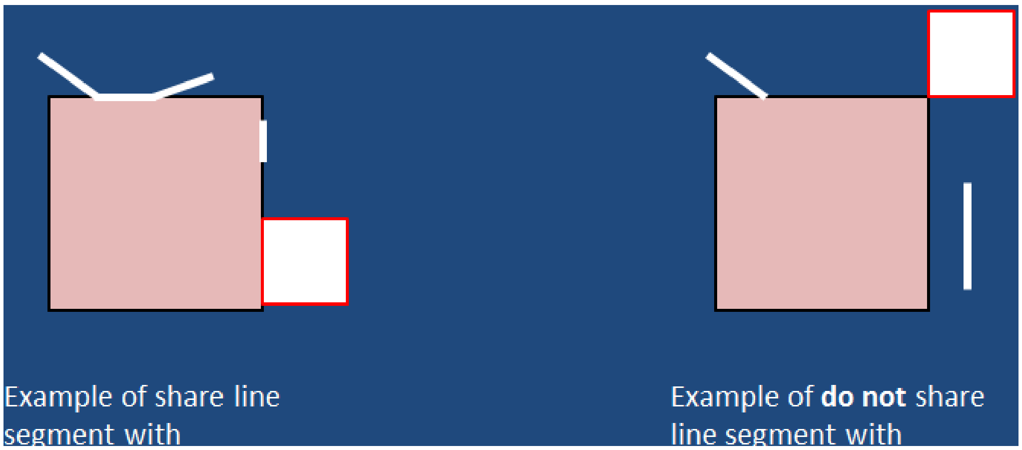


Figure 7 (left) Example of share line segment with, (right) Example of do not share line segment with

"touch the boundary of"

The "touch the boundary of" spatial selection operator selects polygon or line features where the intersection of the target feature with the geometry of the source feature is not empty.

The intersections of their interiors must be empty; otherwise they will not be selected. For example, the image titled Example of Tough Boundary of shows how the two lines, and white fill box touch the boundary of the source polygon, but neither of the lines or the box extend inside of the source polygon.

Therefore, both lines and the box are selected. For example in the image, Example of Do Not Touch Boundary Of, since the two lines and white box extruded into the geometry of the source polygon, they are not selected even though they touch the boundary of the source feature.

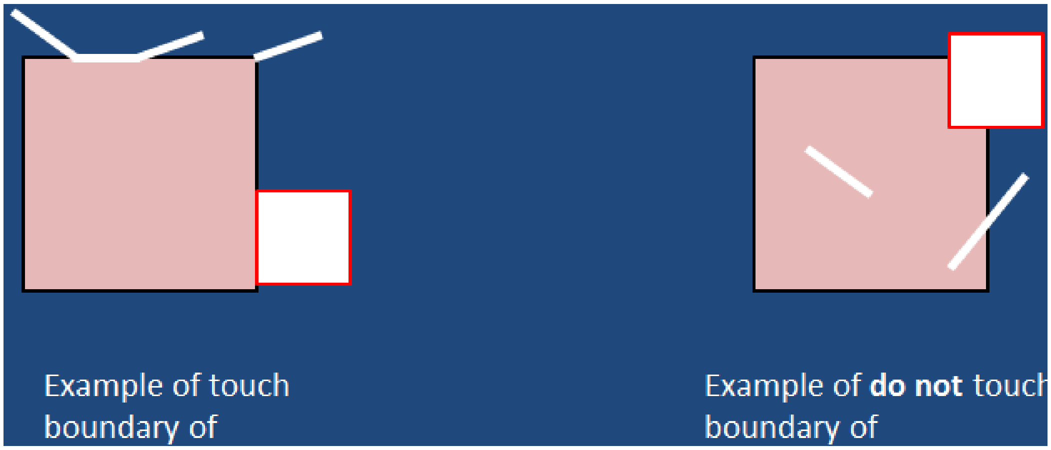


Figure 8 (left) Example of touch boundary of, (right) Example of do not touch boundary of

"are identical to"

The "are identical to" spatial selection operator selects only geometries of features that are strictly equal. For example, the image titled Example of Identical to, shows a thin black line within the figure of the white line both have exactly the same vertices, and are considered equal.

The fact that they are symbolized differently is not relevant for the selection, as it is only the location of the nodes and vertices that determine whether the features are identical. This applies to the two points; the light red circle and the gray circle have the same center point, and are considered to be identical.

Lastly, the light red box with the black outline shares the exact same vertices as the white fill the box with red outline. The image, Example of Not Identical to, provides an illustration of where the geometry between these points, lines, and polygons are not identical, therefore no selections occur.

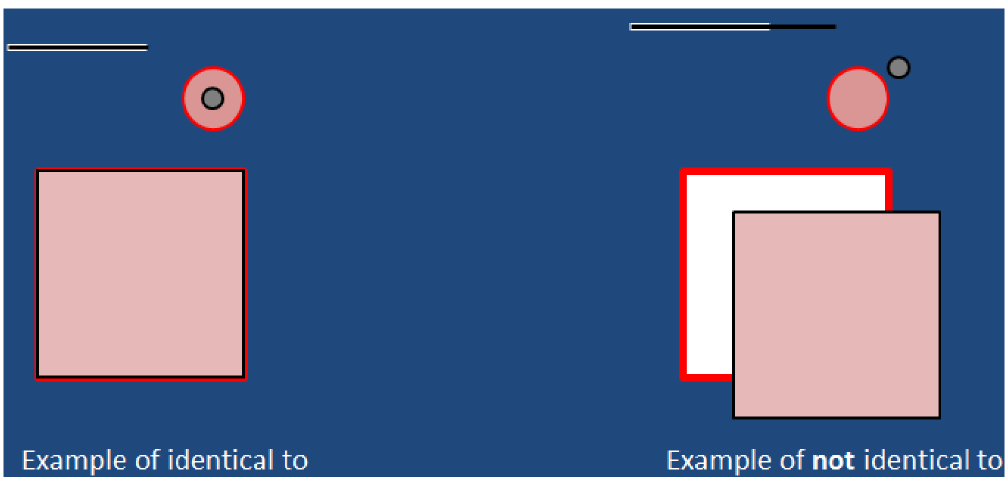


Figure 9 (left) Example of identical to, (right) Example of not identical to

“are crossed by the outline of”

For the “are crossed by the outline of” spatial selection operator, the boundaries of both features must have at least one vertex, and point, or edging common for a selection to take place. It is important to note that the target and source features cannot share a line segment, and must be either lines or polygons.

In the image, Example of Crossed by the Outline of, the white line and white fill box both cross the outline of the light red source polygon and do not share any line segments. In the image, Example of Not Crossed by the Outline of, the line on the top of the source polygon sharing a line segment is not selected, the line in the of top right is not selected because it does not cross the boundary of the source polygon, and the white fill box is not selected because it is only sharing a line segment along the boundary.

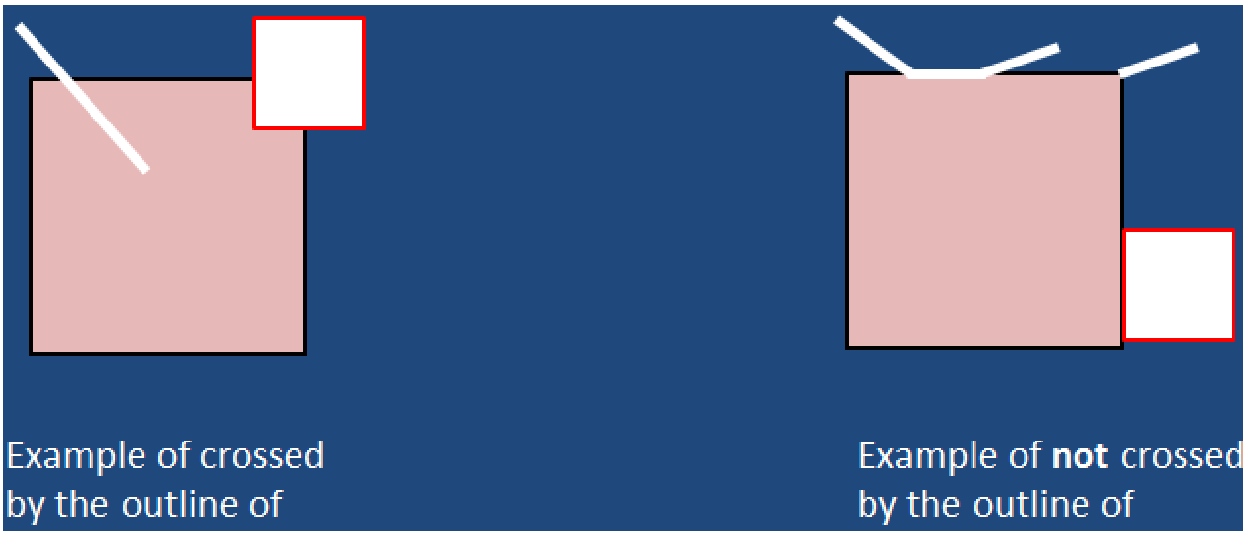


Figure 10 (left) Example of crossed by the outline of, (right) Example of not crossed by the outline of

“is contained by”

“Is contained by” selects target features where the geometry of the source feature falls inside the geometry of the target feature including its boundaries.

Figure 11 provides an example. On the left, the example of is contained by, the thin black line is considered the source feature and is completely contained within the longer white line; therefore the white line is selected. The light red source box is completely contained within the white fill box; therefore the white fill box is selected. On the right, the example of is contained by, illustrates how the black source lines extend past the extents of the white lines, which means no selections occur. Similarly, as the light red source polygon extends across the boundaries of the target white fill box feature, no selection occurs there either.

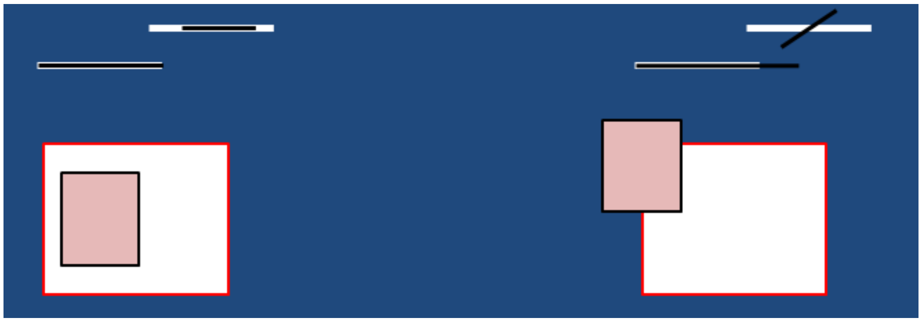


Figure 11 Examples of (left) is contained by; (right) is not contained by

“contains”

The “contains” spatial selection operator selects geometry of the target feature that falls within the geometry of the source feature, including its boundaries. Refer to Figure 12. The image on the left, is an example of contains. In this example, all target features are selected because the source features (represented by the black lines and the light red polygon) encapsulate the target features (represented by the white lines and the white fill box). On the right, the example of not contains, you will find that the target features are not contained, and therefore, are not selected.

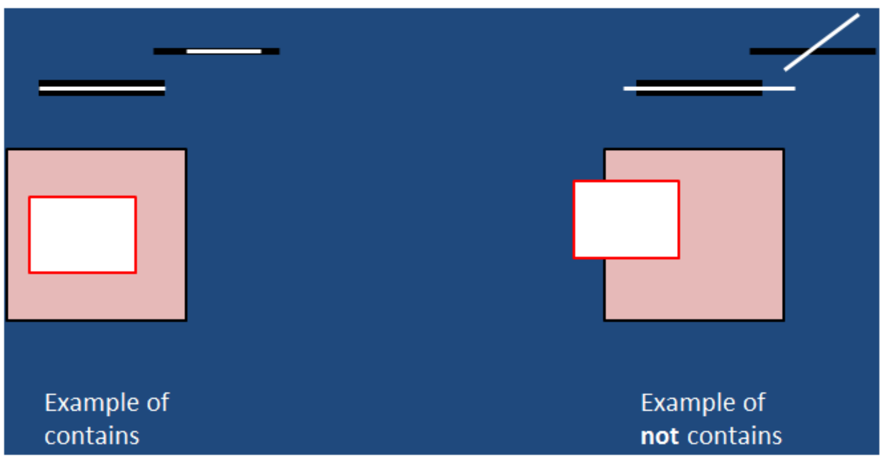


Figure 12 Examples of (left) contains, (right) not contains

Buffering

A buffer is a region that is less than or equal to a distance from one or more input features. Buffers can be created from points, lines, polygons, and raster geospatial data sets. Buffers are typically used to identify areas or objects “inside”, or “outside” the threshold distance. Can you think of any examples where you might use a buffer? One example could be to determine how many homes are within one mile of the coastline. In this case, the coastline is the input and one mile is the threshold distance, which results in a polygon that extends one mile out from the coastline. We can then utilize spatial selection to select all houses that are inside the buffer.

To illustrate a buffer, the polygon to the left is the original input feature. After specifying a buffer distance and running the buffer tool, a new polygon is created at the set buffer distance which surrounds and follows the outline of the original input feature.

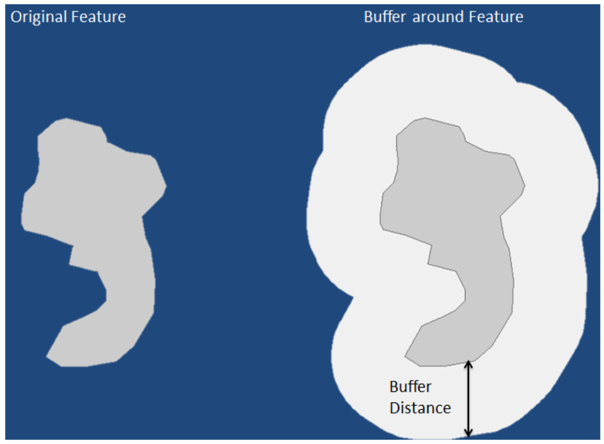
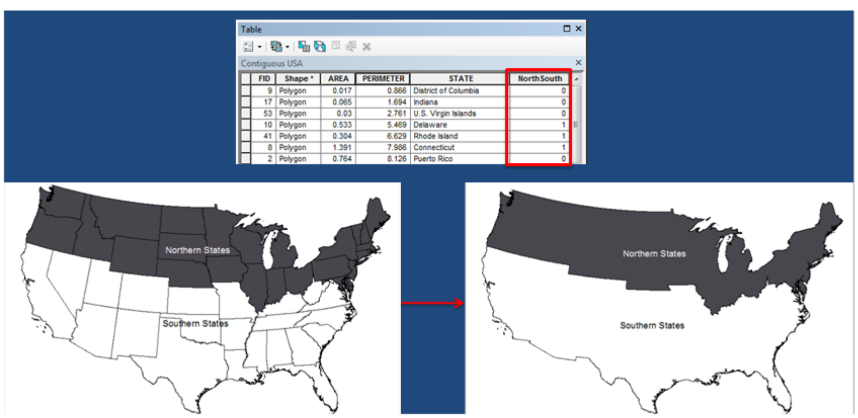


Figure 13 Examples of (left) original feature, (right) buffer around feature

Dissolving

The dissolve operation combines similar features within a data layer based on a shared attribute. For example, the data set used to represent the map on the left has the attribute of north-south. If the state is a northern state it has the value of one, and if the state is a southern state it has a value of zero.

If we use this data set, and dissolve on the north-south attribute field, the result would be demonstrated by the map on the right. All states that have the value of one, are dissolved into a single polygon. All states that have the value of zero for the north-south attribute are dissolved into a second single polygon. The dissolve field can have as many different values as you choose; however, only records with identical values for that attribute will be combined into a single feature.



FID	Shape *	AREA	PERIMETER	STATE	NorthSouth
9	Polygon	0.017	0.866	District of Columbia	0
17	Polygon	0.065	1.694	Indiana	0
53	Polygon	0.03	2.761	U.S. Virgin Islands	0
10	Polygon	0.533	5.469	Delaware	1
41	Polygon	0.304	6.629	Rhode Island	1
8	Polygon	1.391	7.986	Connecticut	1
2	Polygon	0.764	8.126	Puerto Rico	0

Figure 14 Example of Dissolve

SUMMARY

This lesson focused on using attributes and spatial queries for data exploration. You learned about attribute queries and how to create them as well as the role algebra has in this process. The lesson also provided information on data dictionaries, deciphering coded values, and examples of spatial queries to understand how to apply this knowledge.

ASSIGNMENTS

1. Lab: Using Attribute and Spatial Queries for Data Exploration

2. Quiz: Using Advanced Attribute and Spatial Queries for Data Exploration

Authoring Organization: Del Mar College
Written by: Richard Smith

Development was funded by the Department of Labor (DOL) Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant No. TC-22525-11-60-A-48; The National Information Security, Geospatial Technologies Consortium (NISGTC) is an entity of Collin College of Texas, Bellevue College of Washington, Bunker Hill Community College of Massachusetts, Del Mar College of Texas, Moraine Valley Community College of Illinois, Rio Salado College of Arizona, and Salt Lake Community College of Utah.

This workforce solution was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration. The solution was created by the grantee and does not necessarily reflect the official position of the U.S. Department of Labor. The Department of Labor makes no guarantees, warranties or assurances of any kind, express or implied, with respect to such information, including any information on linked sites, and including, but not limited to accuracy of the information or its completeness, timeliness, usefulness, adequacy, continued availability or ownership.

Lesson 4: Vector Data Analysis – Overlay Techniques

INTRODUCTION

In this lesson you will learn about vector data analysis – overlay techniques. The content includes information on environmental settings. Coverage data format and its elements are explained in the lesson. Vector overlay information is explained and examples are presented to further your understanding on this subject.

LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Identify vector data analysis overlay techniques.

2. Convert coverage data format to a modern GIS data format.

3. Explain how environmental settings are used to enhance data organization.

LEARNING SEQUENCE

Required Reading	Read the following: • Online Lesson Material
Resources	View the following: • None
Assignments	Complete the following: • Lab: Vector Data Analysis - Overlay Techniques • Quiz: Vector Data Analysis – Overlay Techniques

INSTRUCTION

What are Environment Settings?

Environment settings are a set of parameters that a user will set prior to executing a tool or operation. Environment settings allow the user to:

- Determine settings for the input, processing, and output of the tool or operation.
- Set defaults for parameters such as output location, analysis extents, cell sizes, and so on.
- Remain organized as well as make tasks quicker and easier as environment settings can be shared between tools and operations.

Location

In ArcMap, and ArcCatalog, you can access the environment settings dialog box by clicking Geoprocessing from the menu bar followed by environments.

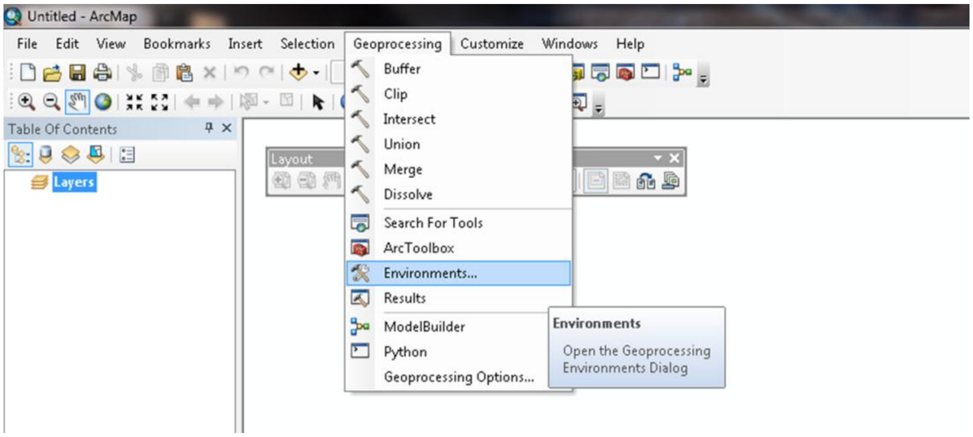


Figure 1 Screenshot of Geoprocessing tab within ArcMap: Accessing Environment Settings

Parameters

This is the environment settings dialog box as opened in either ArcMap or ArcCatalog. You can click the double down chevrons to expand each environment setting and change its values.

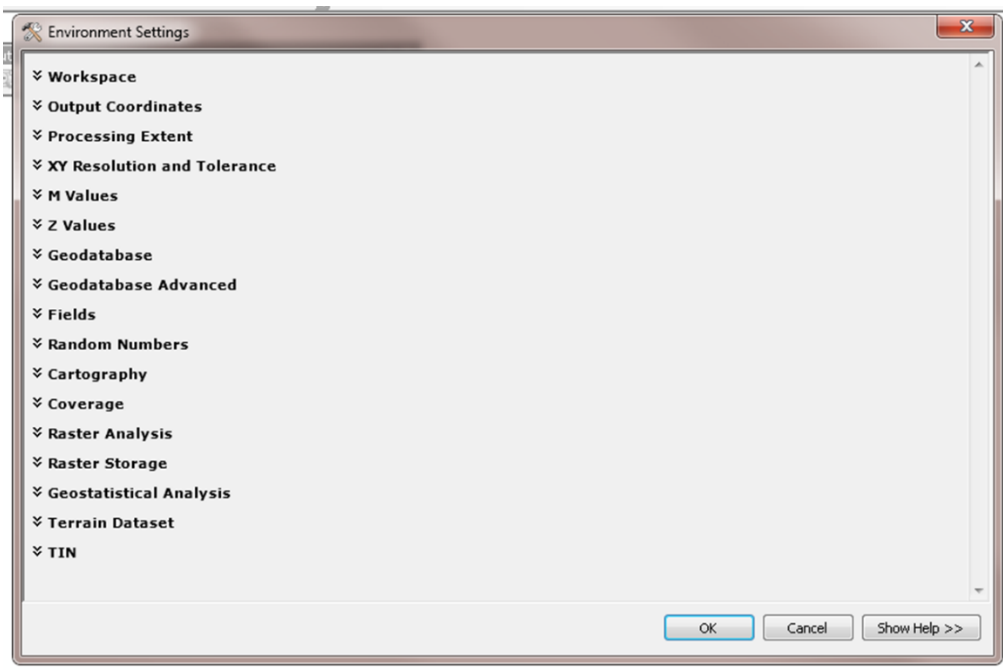


Figure 2 Screenshot of Environment Settings dialog box: List of Parameters

Review a few environment settings.

- The workspace environment setting sets the current and scratch workspace. The workspace is where the tool will look for input values by default, and where the tool will place intermediate and output files by default.

- The output coordinates environment setting sets a geographic transformation for all output coordinates when the tool or operation is ran. You can find more information about other environment settings in ArcGIS desktop help.

Coverage Data Format

The coverage data format stores vector data such as points, lines, and polygons. Coverage files contain the spatial and attribute data for the geographic features that it represents. Coverage is similar to a shapefile, however, the file format is quite different, and the data is stored in separate feature classes.

Storing Coverages

To store a coverage, there are a few restrictions to keep in mind. A coverage is stored as a directory containing multiple files. The name of a coverage cannot be longer than 13 characters, it cannot contain spaces, it cannot start with a number, it does not have a filename extension, and it must be named using all lowercase letters.

The reason for these restrictions is that a coverage is considered a legacy format, and was created during a time when computers and operating systems did not support longer and more complex filenames.

Coverage Composition

A single coverage is composed of multiple files, similar in the way that a shapefile is composed of multiple files. Each file contains information concerning one part of the entire feature class. The contents of a coverage will appear different in Windows Explorer versus ArcCatalog, and it is not recommended that you ever manually move or change any of the files inside of the coverage folder. It is recommended that you manage the coverages using ArcCatalog, as it simplifies the coverage files into a single entry in the ArcCatalog file viewer.

Converting Coverages

The reason coverages still need to be learned is that quite a bit of data is still provided in this legacy format. Therefore, while it is not recommended that you save newly created data into the coverage format, you must know how to convert the coverages into a more contemporary GIS data format.

To make things slightly more complicated, coverage files are often delivered in an interchange format, known as the E00 format. In ArcGIS, a script tool is provided to convert coverages in the interchange format to a feature data set. This tool is located Arctoolbox under the conversion tools toolbox, to coverage toolset, and is named import from E00.

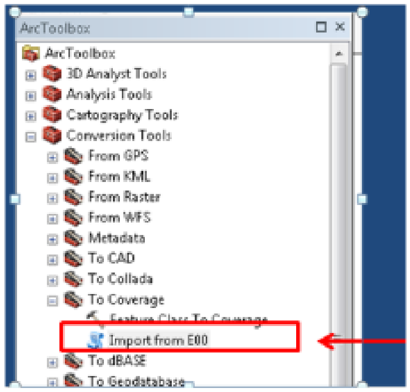


Figure 3 ArcToolbox: Import from E00

Overlay Operations

Overlay operations involve combining spatial and attribute data from two or more spatial data layers. In other words, we are stacking the data, and looking for where the layers overlap. While the overlay operations may seem simple when used alone, they can actually be quite powerful operations when combined in series.

One thing to note about overlay operations is that overlays require that the data be in the same coordinate system. This is so that the software can perform the highest accuracy of overlay. Some software packages automatically re-project on-the-fly if the data is not in the same coordinate system, but it is preferable if you convert all data sets to the same coordinate system before performing the overlay operation.

An example of an overlay operation would be the clip operation. For example, say I have a data set that contains all the streets in the United States of America. However, I only want to see the streets in Arkansas. I can use the clip operation to cut out all the streets that are only in Arkansas by providing the state outline of Arkansas as the clipping feature.

Vector Overlay

A vector overlay involves combining point, line, or polygon geometry and their associated attributes. All overlay operations create new geometry and a new output geospatial data set. You should be cautioned that with certain overlay operations, very large attribute tables may result if the overlay operations combine many layers that each possess.

Additionally, it might be possible that the combined attribute tables would cause duplicate attribute fields to exist. In these cases, you should consider reducing the number of transferred attributes to the minimum required, and renaming duplicate fields so that there is no ambiguity.

Basic Cases of Overlay - Clip

The first basic case of overlay is the clip function. The clip function defines the area for which features will be output based on a "clipping" polygon. I like to think of it in terms of cookie dough and a cookie-cutter. The cookie dough is the data layer that we want to reduce in size based on our cookie-cutter.

The clipping layer is the cookie-cutter that will determine how we cut out the part of the cookie dough we are interested in. It is important to note that in the clip operation only the geometry and attributes of the data layer are transferred to the result layer.

Neither the geometry nor attribute of the clipping layer are included in the output, just as you would not make your cookies in the oven with the cookie-cutter still surrounding the cookie dough.

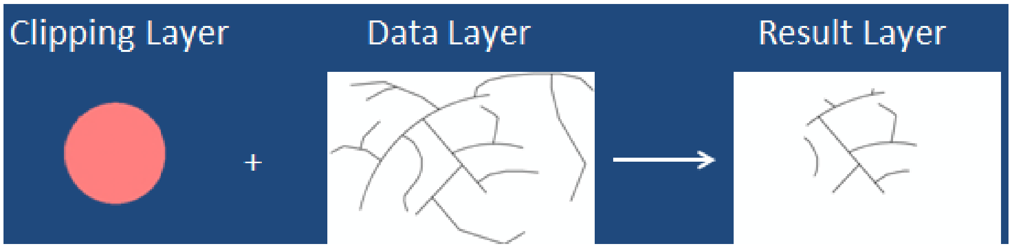


Figure 4 Clip Function

Basic Cases of Overlay - Intersection

The next basic case of overlay is the intersection operation. The intersection operation combines data from two or more input layers, but the combination only exists for the regions where all input layers contain data. This is similar to the clip function; however the attributes and geometry of all input data layers are transferred to the result layer.

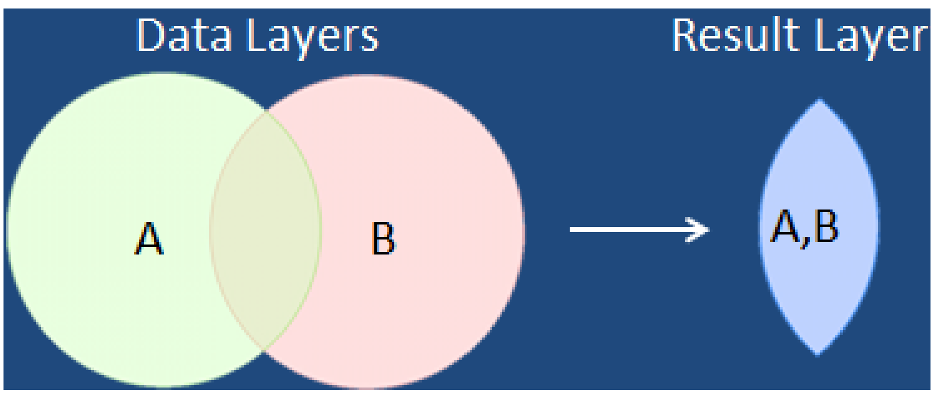


Figure 5 Intersection Operation

Basic Cases of Overlay - Union

The third basic case of overlay is the union operation. The union operation is an overlay that includes all data from all of the input data layers. Since the union operation combines all attributes, and all geometry, no geographic data are discarded in the union operation, and all corresponding attribute data are saved for all regions. Figure 6 is an illustration of a union overlay where we initially had two separate data sets, A and B that, when union is applied, result in a single combined data set.

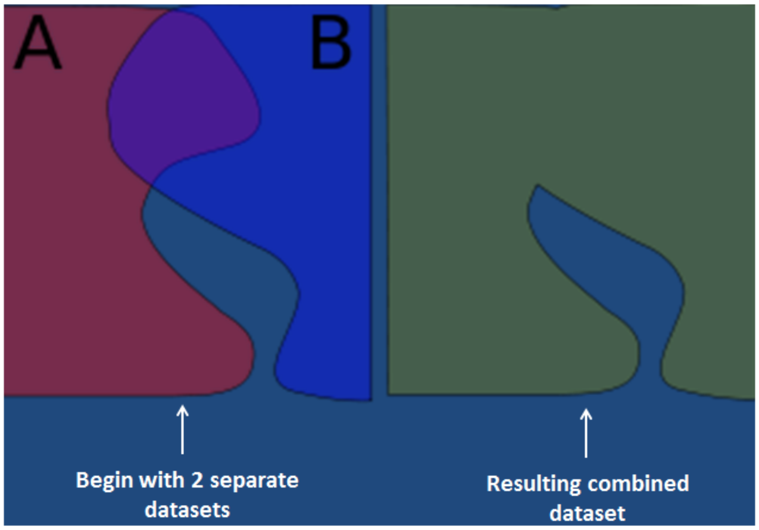


Figure 6 Union Operation

Vector Overlay Problems

When performing vector overlays, there are a few problems that you should be aware of. In the case where the input features represent common features that are represented in both layers, they may have slightly different geometry. This creates sliver polygons when the overlay operations are performed.

Example 1

Imagine we have one input data set that represents the national boundary of Canada created at a 1 to 5,000,000 scale, and that we have a second input data set of the national boundary of the United States of America created at a 12 to 1,000,000 scale. If we stack these two layers on top of each other, it is probable that the boundary lines will not overlay exactly, as they were created at different scales, and therefore the Canadian boundary line has probably been simplified more than the United States boundary line. The areas where the two input data sets do not match will result in very small and typically insignificant sliver polygons that need to be dealt with.

There are several methods that exist to reduce the occurrence of sliver polygons, such as manually snapping the boundaries together, or introducing fuzziness to the edges of the overlay operation, or simply manually deleting the sliver polygons.

Example 2

To illustrate the idea of a sliver polygon, Figure 7 includes two different data sets, one outlined in a bold black line (Data Set 1), and the other outlined in a thinner gray line (Data Set 2). Where the two data sets overlay, they do not match exactly for whatever reason. If we perform an intersect, clip, or union operation on these two input data sets, we are going to create three sliver polygons.

The sliver polygons created between these two input data sets are highlighted in red. The sliver polygons are probably insignificant, and are the results of data sets being produced by different agencies, at different times, and probably at different scales. It is up to you to determine how you would like to solve for the sliver polygons.

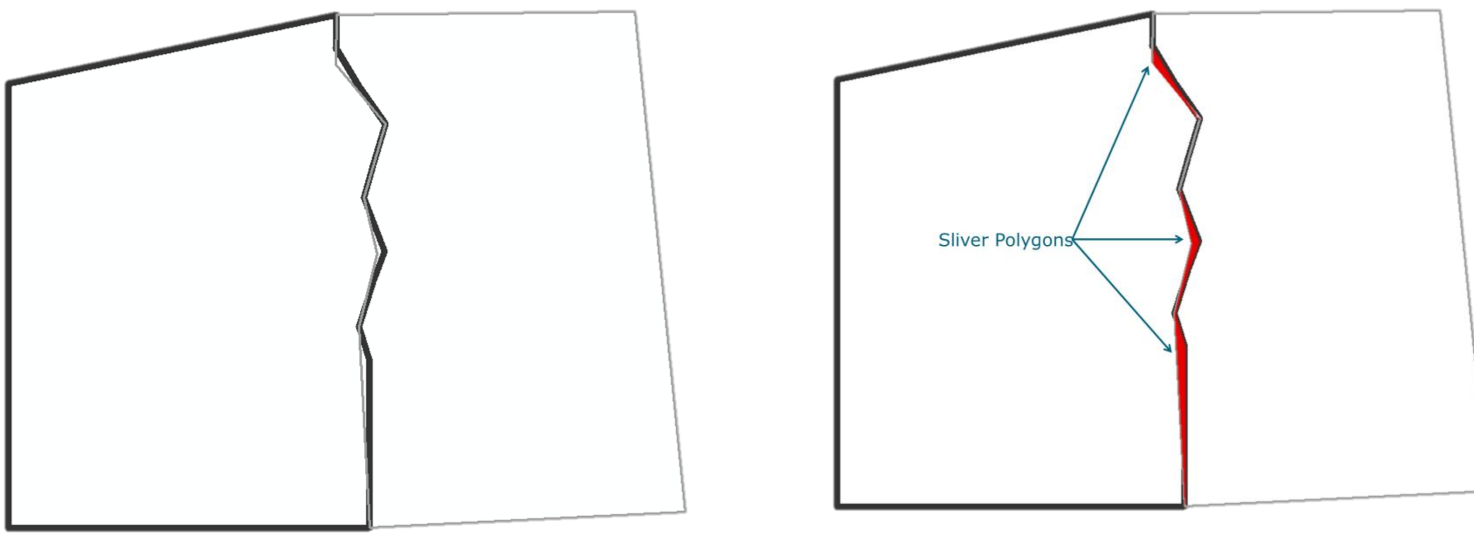


Figure 7 (left) Data Set 1, (right) Data Set 2

SUMMARY

In this lesson you learned about vector data analysis – overlay techniques. You learned about environmental settings and how to apply them in the ArcMap and ArcCatalog. Coverage data format was explained to help you understand storing coverages, coverage composition, and converting coverages. Vector overlays were explained with examples to demonstrate the steps used in this process. The lesson also provided examples of overlay problems to show you what issues can arise when performing vector overlays.

ASSIGNMENTS

1. Lab: Vector Data Analysis - Overlay Techniques
2. Quiz: Vector Data Analysis – Overlay Technique

Authoring Organization: Del Mar College
Written by: Richard Smith

Development was funded by the Department of Labor (DOL) Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant No. TC-22525-11-60-A-48; The National Information Security, Geospatial Technologies Consortium (NISGTC) is an entity of Collin College of Texas, Bellevue College of Washington, Bunker Hill Community College of Massachusetts, Del Mar College of Texas, Moraine Valley Community College of Illinois, Rio Salado College of Arizona, and Salt Lake Community College of Utah.

This workforce solution was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration. The solution was created by the grantee and does not necessarily reflect the official position of the U.S. Department of Labor. The Department of Labor makes no guarantees, warranties or assurances of any kind, express or implied, with respect to such information, including any information on linked sites, and including, but not limited to accuracy of the information or its completeness, timeliness, usefulness, adequacy, continued availability or ownership.

Lesson 5: Vector Data Analysis - Creating a Site Selection Model

INTRODUCTION

In this lesson you will learn about vector data analysis – creating a site selection model. The information will explain what proximity analysis is and tools that use this information. You will learn about ModelBuilder which is a tool in the ArcGIS desktop suite and how to use it. Examples of site selection are include with explanations to help you gain a better understanding of site selection and the role it has in vector data analysis.

LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Identify elements of vector data analysis used for creating a site selection model.

2. Apply the method of proximity analysis for buffering elements.

3. Develop a model that satisfies multiple location criteria.

LEARNING SEQUENCE

Required Reading	Read the following: • Online Lesson Material
Resources	View the following: • None
Assignments	Complete the following: • Lab: Vector Data Analysis – Creating a Site Selection Model • Quiz: Vector Data Analysis – Creating a Site Selection Model

INSTRUCTION

Proximity Analysis

Proximity analysis is a set of tools used to analyze the relationship between a selected feature and its neighbors.

Consider the following questions:

- How close is the nearest gas station?
- What is the distance between your house and the candy store?
- What is the shortest route to get to Starbucks?

- Is there a mechanic a mile away?

Questions like these can be answered with a GIS by using proximity analysis tools. Typically, there are different proximity analysis tools for vector and raster data sets, as the algorithms may be quite different even though the desired result is the same. Proximity analysis tools output tables and or features and can take one or more inputs.

Proximity Toolset

In an ArcGIS desktop, you will find the "Proximity" toolset within the "Analysis Tools" inside of "ArcToolbox". The "Proximity" toolset contains several tools.

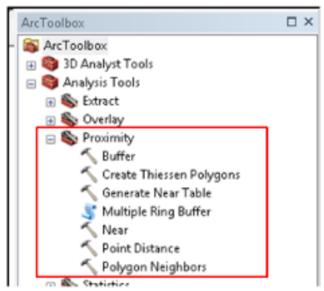


Figure 1 Arc Toolbox: Proximity

Buffers

The "Proximity" toolset includes the "Buffer" tool, and the "Multiple Ring Buffer" tool script. There are two types of buffers available: single buffers and multiple-ring buffers.

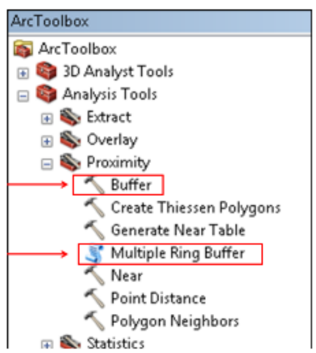


Figure 2 Arc Toolbox: Buffer and Multiple Ring Buffer

Single Buffers: Area Feature

A single buffer creates an area feature at a single specified distance from the input feature. An example of a use for a buffer is if you wanted to specify a distance around a rocket test site to identify areas where sound levels will be disturbing to the residents.

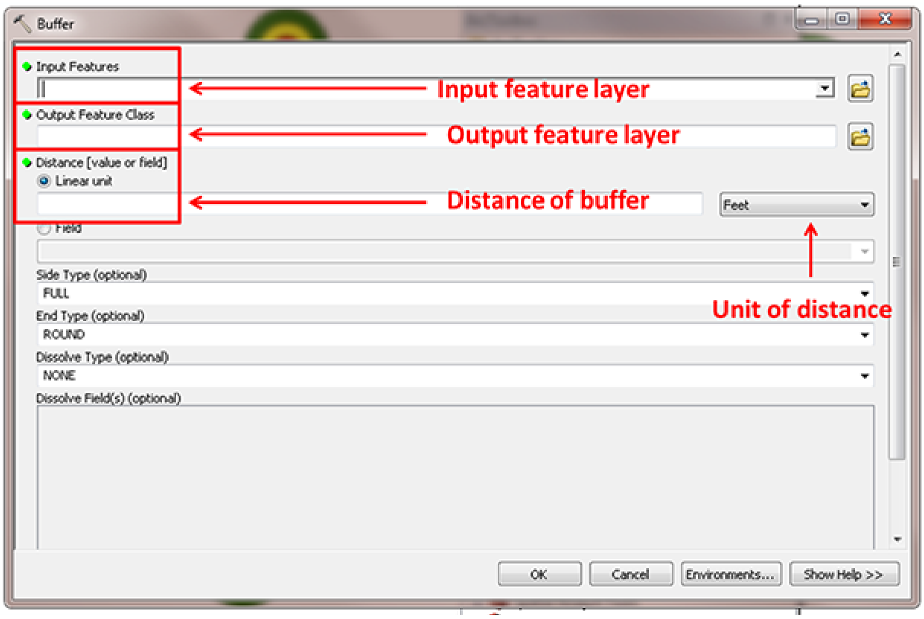
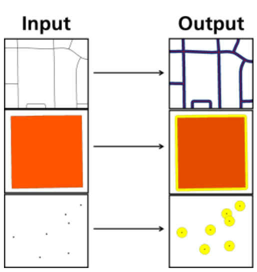


Figure 3 Screenshot of how to create an area feature.

Single Buffer Tool: Three Parameters



The single "Buffer" tool requires three parameters to be filled out before it can run. The first parameter is the input feature layer, which will be the center point of each buffer. The second parameter is the output feature layer which will contain all of the buffer areas. The third parameter is the linear distance in a distance unit, such as feet, miles, or kilometers, that determines the radius of the buffer around the input features. Another option for the third parameter is to select a field which contains numeric values that represent the radius of the buffer you wish to have around that individual feature.

Multiple-Ring Buffers

A multiple-ring buffer creates area features at multiple specified distances from the input feature. For example, we could use multiple-ring buffers to buffer multiple distances from a business to see how many customers live in each buffer range.

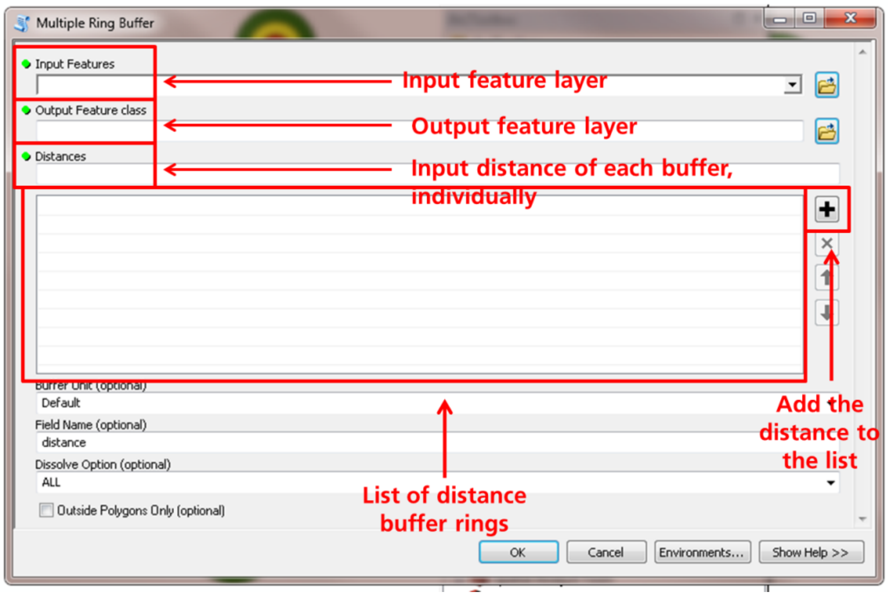
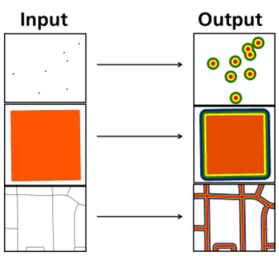


Figure 4 Screenshot of Multiple Ring Buffer window where it is possible to create area features at multiple specified distances from the input feature based on selections.

Multiple-Ring Buffer Tool: Three Parameters



The "*Multiple-Ring Buffer*" tool requires three parameters to be filled out before it can run. The first parameter is the input feature layer, which will be the center point of each buffer. The second parameter is the output feature layer which will contain all of the buffer areas. The third parameter is the linear distance in a distance unit, such as feet, miles, or kilometers, that determines the radius of the buffer around the input features. You can add multiple distances by pressing the plus button in the tool to create a new entry in the list of distance buffer rings.

ModelBuilder

Modelbuilder is a tool in the ArcGIS desktop suite that allows you to create, edit, and manage Geoprocessing models. Modelbuilder preserves a workflow that you can execute multiple times. A model chains together multiple tools, allowing for the output of one tool to be the input of a subsequent tool. Building models provides the benefit of automating the process of running multiple tools in series, and easily saving, sharing, and rerunning the model.

Custom Toolbox

In order to create a model, the model must first reside inside of a new, custom toolbox. To create a custom toolbox in "*ArcToolbox*", open "*ArcToolbox*", then right-click on "*ArcToolbox*", and then click "*Add Toolbox*".

Figure 5 Screenshot of ArcToobox Tab: Creating a Custom Toolbox

Inputting Tools

When you create a new model in your new toolbox, you can "*Add Data*" or "Tool" to Modelbuilder by simply dragging and dropping tools from "*ArcToolbox*". Another way to add tools or data is to click "*Insert*" from" the menu bar followed by "*Add Data*" or "*Tool*".

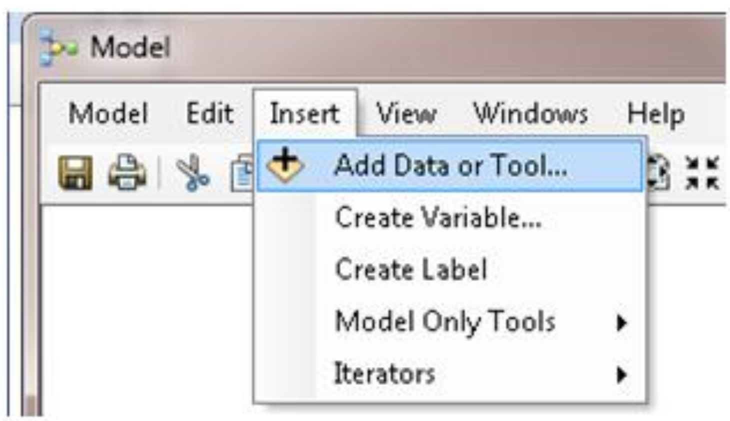


Figure 6 Screenshot of Insert Tab: Adding Tools or Data

ModelBuilder provides an easy to use and understand graphical user interface. Tools and data are represented by rectangles and ovals, respectively. After you initially drag a tool into ModelBuilder, the shapes will look hollow. They are hollow because the proper parameters have yet to be filled out. Once you fill in the proper parameters, the shapes will be colored and the data names will be displayed.

Flow Charts

ModelBuilder represents the algorithm that you are designing as a flowchart. The boxes represent operations. Ovals represent the data that is inputted and outputted. Boxes and ovals are connected by arrows to set the order of execution and data flow. If you follow the arrows from the starting point or points of the flowchart, it will reveal the sequential order in which the operations must be executed.

Running the Model

After you have designed your model, and all of the ovals and boxes are colored, you can run the model by choosing "*Model*" from the menu bar followed by "*Run Entire Model*". The model run and the tools and output data will display a drop shadow showing that they have been run. Once all of the data and tools have a drop shadow, you can view the results. If there are any errors in execution, ModelBuilder will prompt you for more information.

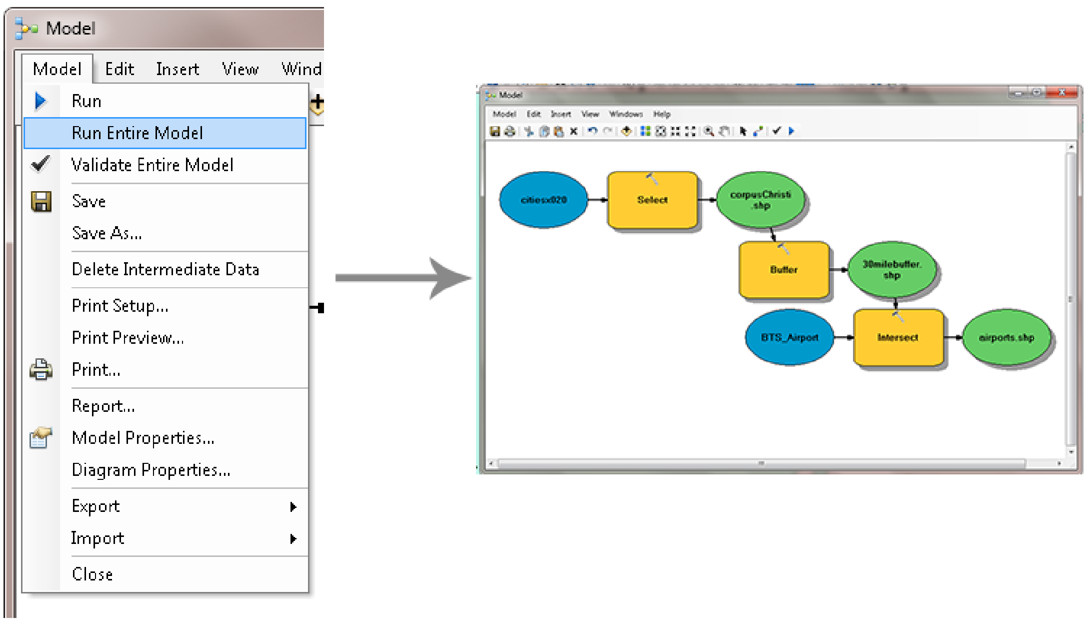


Figure 7 Screenshot of Model Tab: Run Entire Model

Saving and Accessing the Tool

Once you verify that your tool runs correctly, you can save the model by clicking the "Save" button. The model is saved inside of the custom toolbox that you created. If you wish to share your model with a colleague, simply provide your colleague with a copy of the toolbox file, and tell them to add it to their "*ArcToolbox*". As a reminder, to add a toolbox to "*ArcToolbox*", right-click "*ArcToolbox*", then select "*Add Toolbox*", and navigate to the toolbox and "select it". The toolbox will then display in "*ArcToolbox*" and your model will be listed inside of it.

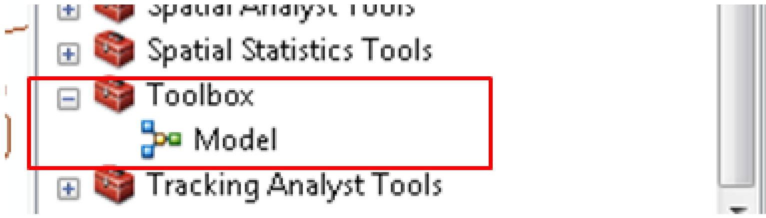


Figure 8 Screenshot: Saving and Accessing the Tool

Site Selection

Site selection is where you measure the needs of a project against the virtues of potential locations. There are many uses for site selection, such as finding an ideal location for a new facility, finding an existing facility for use, or identifying homes for sale within a certain neighborhood at a certain price range.

Site Selection Example

I want to fly to Corpus Christi, Texas and want to find an airport that is within 30 miles of Corpus Christi, so I do not have to drive too far once I arrive. My goal is to identify all the airports within 30 miles of Corpus Christi, Texas.

Data

The first thing I must determine is what data I require. For me to solve this problem, I need a data set containing cities in Texas, and the data set containing airports in Texas.

Tools

The second thing to determine is which tools I should use to identify the airports within 30 miles of Corpus Christi. I have identified three tools that I will need to use: "*Select*", "*Buffer*", and "*Intersect*". The "*Select*" tool will allow me to select Corpus Christi, Texas from all of the cities in Texas. The "*Buffer*" tool will create a 30 mile radius buffer around the selected city of Corpus Christi, Texas. Finally, the "*Intersect*" tool will collect the airports from the airport layer that intersects with the 30 mile radius buffer.

Solving the Site Selection Example: Setting Parameters

Review the five steps outlined below which were used to set the parameters for the Site Selection Example.

Step 1: Creating a Model

Now that I have all of my tools identified, I drag them one by one from "*ArcToolbox*" into an empty model. I will then review how to set the parameters for each of the three tools: "*Input*", "*Buffer*", and "*Intersect*". Remember, all the tools and data outputs are not filled in because the tools do not have all the required parameters to run yet.

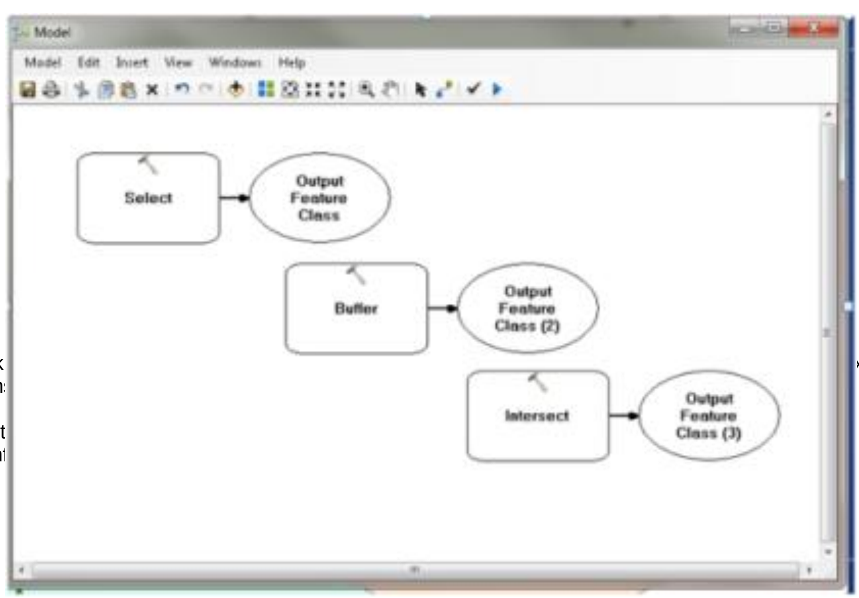


Figure 9 Tools Added to ModelBuilder

Step 2: Select Input Parameters

For the "*Select*" tool, I set the "*Input Features*" as my city's feature class. I set my "*Output Feature Class*" to the location I want to save it at with the feature class name I wish to use. Last, I use "*Query Builder"* to form the input features.

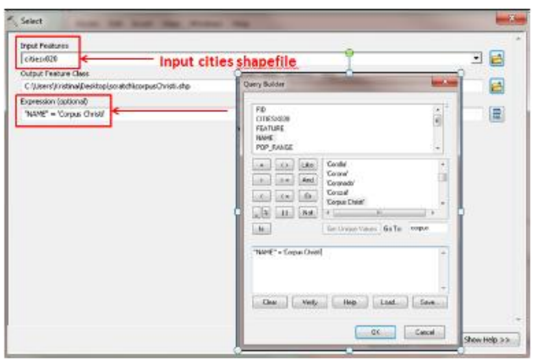


Figure 10 Select Input Parameters

Step 3: Buffer Input Parameters

For the "*Buffer*" tool, the "*Input Features*" are the output file from the "*Select*" tool that I just set the parameters for. I set the "*Output Feature Class"* name, and also set the "*Linear unit*" for the "*Buffer*", which is 30 miles.

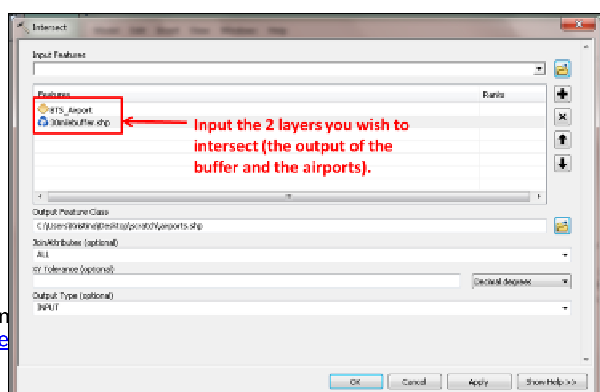


Figure 11 Buffer Input Parameters

Step 4: Intersect Parameters

Lastly, for the input tool, I set the "*Input Features*" to the output of the "*Buffer*" tool that I just set, and the airports feature class. I then set the "*Output Feature Class*" name which will represent all the airports within 30 miles of Corpus Christi.

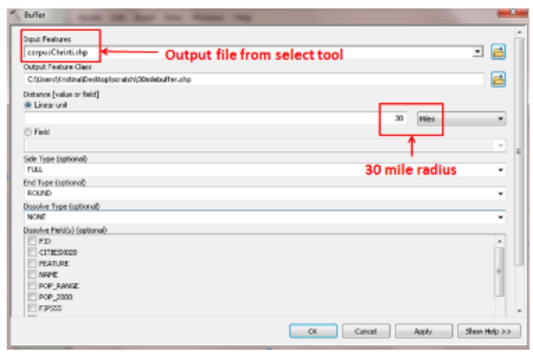


Figure 12 Insect Parameters

Step 5: Ready to Run

This is what the model looks like with all the parameters set. The blue ovals represent data that is only used as input. The green ovals represent data that is the output of tools that may or may not be used as input to other tools. This model is now ready to run.

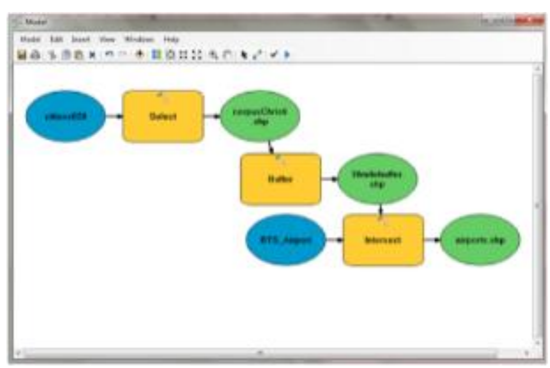


Figure 13 Model: All Parameters Set

Before I run the model, I will "Save" the model inside of my custom toolbox. Now it is time to run the model by choosing "Model" from the menu bar followed by "Run Entire Model".

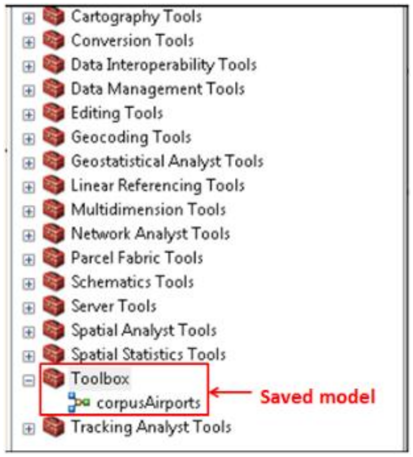


Figure 14 Saving the Model

Figure 15 illustrates what a model looks like after it has completely executed. All the tools and outputs have drop shadows indicating that they have successfully executed. Now that the model has completed, we can look at the results in ArcMap.

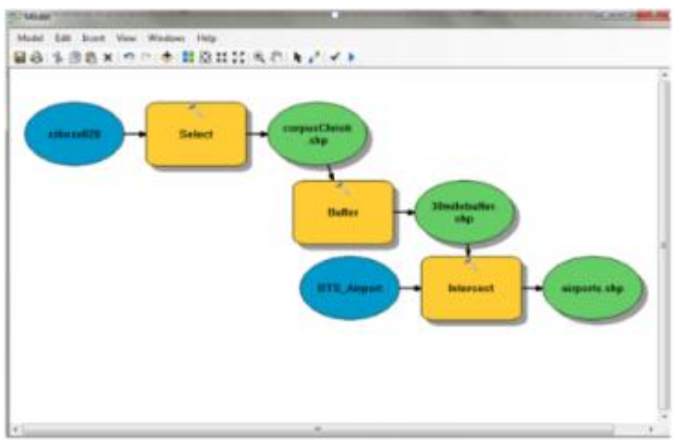


Figure 15 Model Run Complete

The results of the model can be viewed in Figure 16. The green dots represent all the airports within 30 miles of Corpus Christi, Texas. The tan oval represents the 30 mile radius around Corpus Christi, Texas. The green background represents the coastal bend of Texas.

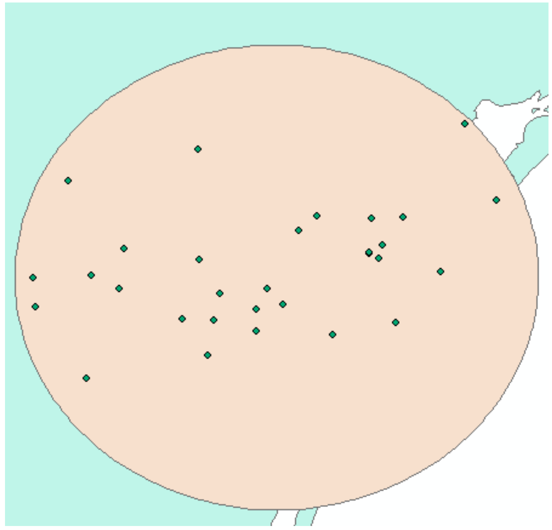


Figure 16 Airport Locations

SUMMARY

This lesson explained vector data analysis – creating a site selection model. You learned what proximity analysis tools are such as the proximity toolkit and buffers in the ArcGIS desktop. The content includes examples and explanations of different buffer types used in vector data analysis. ModelBuilder tools and steps on how to use ModelBuilder were illustrated in the lesson to aid your understanding of this tool of the ArcGIS desktop. Site selection was part of the lesson and you also learned about setting parameters before running the system.

ASSIGNMENTS

1. Lab: Vector Data Analysis – Creating a Site Selection Model

2. Quiz: Vector Data Analysis – Creating a Site Selection Model

Authoring Organization: Del Mar College
Written by: Richard Smith

Development was funded by the Department of Labor (DOL) Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant No. TC-22525-11-60-A-48; The National Information Security, Geospatial Technologies Consortium (NISGTC) is an entity of Collin College of Texas, Bellevue College of Washington, Bunker Hill Community College of Massachusetts, Del Mar College of Texas, Moraine Valley Community College of Illinois, Rio Salado College of Arizona, and Salt Lake Community College of Utah.

This workforce solution was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration. The solution was created by the grantee and does not necessarily reflect the official position of the U.S. Department of Labor. The Department of Labor makes no guarantees, warranties or assurances of any kind, express or implied, with respect to such information, including any information on linked sites, and including, but not limited to accuracy of the information or its completeness, timeliness, usefulness, adequacy, continued availability or ownership.

Lesson 6: Vector Data Analysis - Network Analysis

INTRODUCTION

In this lesson you will learn about vector data analysis – network analysis. Applications for network analysis and benefits are examined in the content. You will learn about network analysis tools and how to use them. Topology examples are included as are the advantages and disadvantages to help you understand topology as it relates to network analysis.

LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Prepare vector data sets for use in network routing.
2. Apply network techniques to create efficient routes including impedances.
3. Generate service areas based on network analysis.

LEARNING SEQUENCE

Required Reading	Read the following: • Online Lesson Material
Resources	View the following: • None
Assignments	Complete the following: • Lab: Vector Data Analysis – Network Analysis • Quiz: Vector Data Analysis – Network Analysis

INSTRUCTION

Network Analysis

Network analysis allows us to solve everyday questions.

- What is the quickest route to a restaurant?
- Which homes are 15 minutes from an elementary school?
- Which patrolman is closest to an incoming accident?
- Where is the best place for a business to open a new branch?

You may have noticed that all of these questions have some sort of distance measurement, route selection, or service area component. Network analysis tools are used to address questions relating to these concepts.

Benefits of Network Analysis

Network analysis can provide information needed to make strategic decisions. For instance, a business that transports goods based off of transportation networks would benefit by using optimal route selections that could reduce transportation costs. With response emergency management, network analysis can analyze current traffic patterns and the street network to determine the quickest path for an emergency vehicle. Additionally, if there are multiple destinations, and multiple travelers, network analysis can work to route each of the travelers in the most optimal way to each of the destinations without each traveler having to cover too much ground and staying within a reasonable service area. Lastly, network analysis allows the calculation of more accurate results by using network distances such as street networks, or navigable rivers, rather than simply using straight-line distances.

What is Network Analysis?

Network analysis is a system of interconnected points and lines that represent possible routes from one location to another. A network is composed of locations connected by links. For example, in a neighborhood, locations would be houses, and the links would be the roads throughout the neighborhood connecting the houses. These network data sets are provided in vector format only, as network links are best represented by line features. Examples of network data sets are telephone cables, roads, water distribution pipes, and power lines.

Impedances

All networks include impedances, which are obstacles that slow travel time. Examples of impedances are a traffic accident on a road, a poor connection in a computer network, and power loss over long stretches of power lines. Impedances may also be expressed as the cost to travel some distance along the network, such as the amount of gasoline to travel between two cities, the cost of maintaining a vehicle to travel cross-country, the salary of a bus driver ferrying passengers across the city, or any other cost related with traversing the network. Typically, impedances are taken into account with traversing a network by determining which path is optimal through minimizing the time spent, money consumed, distance traveled, or some other specified cost.

Network Analysis Tools

The closest facility network analysis tool measures the cost of going between facilities or determines which facility is nearest to an incident. When using the closest facility tool, you can specify impedances along the way. The result of the closest facility tool is that it will display the best routes between facilities, reports costs, and returns directions. A common example of when to use the closest facility tool would be to dispatch a fire truck to a fire from the closest fire house.

Closest Facility Example

In this example, the warehouse manager wants to know from which of the two warehouses should the shipment originate while utilizing the most efficient path for the delivery truck to take. In this case, the warehouse on the left was chosen, as it is closer to the first customer. Since it must already deliver to the first customer, it makes sense for that truck to drive just a little further to deliver to the second customer instead of dispatching a second truck from the other warehouse to make the delivery needed for the second customer.

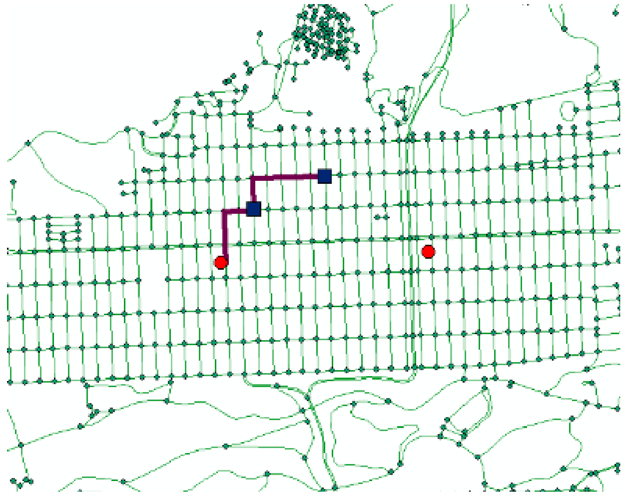


Figure 1 Example of Closest Facility

Route Selection

Another network analysis tool is the route selection tool. The route selection tool identifies the best route to take based on certain criteria. The route selection tool is often used to find the least costly route that visits a number of facilities. In other words, the route selection tool will find a path that uses the least amount of time and gas, assuming gas and time are the two impedances you wish to minimize the cost of. In addition to the graphic output from the route selection tool, a list of directions is provided from the origin to the destination.

Route Selection Example

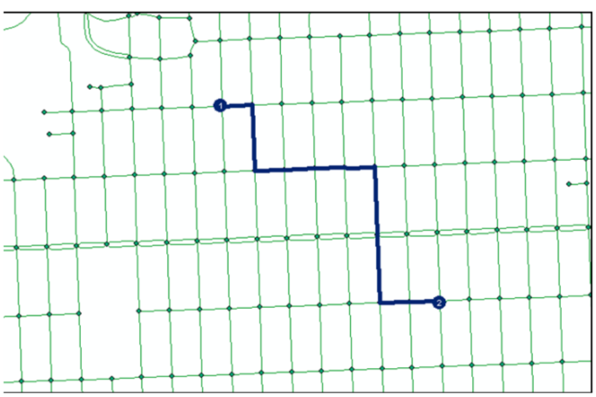


Figure 2 The origin is at the blue circle labeled number one. The path that is outlined that minimizes distance and gas when traveling from the origin to the destination blue circle labeled two.

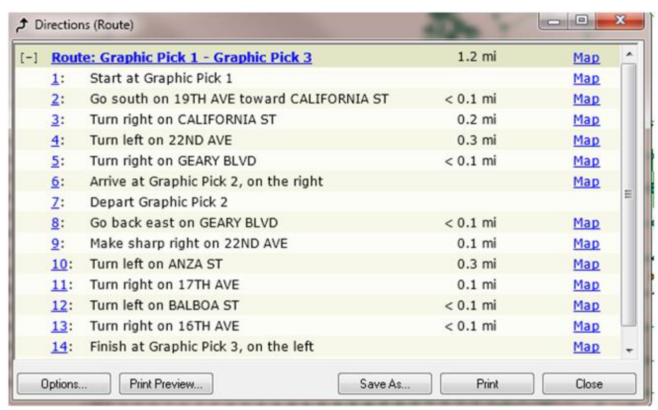


Figure 3 Directions are provided from the origin to the destination.

Service Networks

The service networks tool identifies accessible streets within specified impedances. For instance, if we want to select all homes that are within 10 miles of a school, the service networks tool takes into account travel costs along the network from the origin outwards.

Example: Output of a Service Network Tool

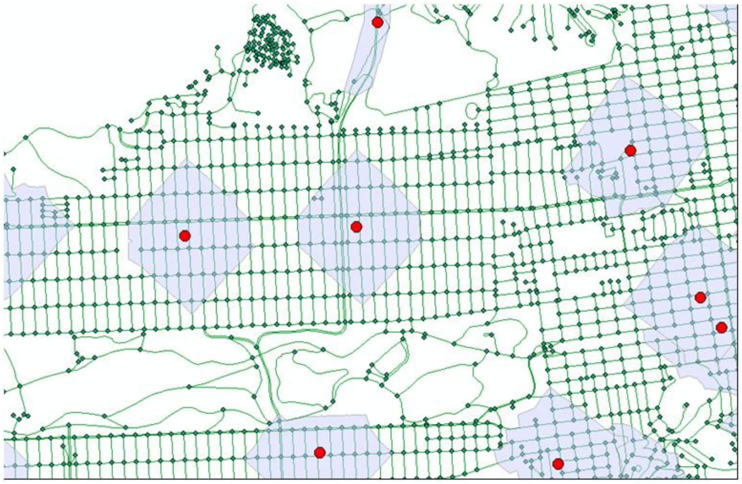


Figure 4 This is an example of the output of a service network tool showing the extent where they can travel 1000 feet from the origin along the network.

Location Allocation

The location allocation tool provides a means to perform site selection analysis. By using appropriate input parameters, the location allocation tool locates the best location for new facilities by taking into account the demand points. You can also take into account locations of existing and competing facilities if it has a bearing on the allocation.

Example: Location Allocation Problem

In order to minimize travel time for customers, where should a fast food restaurant be located?

In our location allocation problem, the customers are the demand points, and the proposed fast food restaurant locations are the candidate facilities. You can also enter the facilities that already exist, whether they are another one of this fast food chain's restaurants, or whether the location is a competitor's fast food restaurant. You may want to include these so that you do not place your new fast food restaurant very close to one of your existing fast food restaurants.

Once all of the inputs have been specified, the location allocation tool will perform an analysis based upon the time traveled attribute in this network, in order to specify which candidate facility would be the optimal site for your new fast food location.

Topology

What is Topology?

Topology is the study of the geometric properties that do not change when features undergo transformations or modifications. The topology represents and enforces the geometric relationships between features. In this brief introduction to topology, it is perhaps best explained visually. Let us continue this brief introduction to the concept of topology by examining its two types: planar topology and non-planar topology.

Planar Topology

In planar topology, the geometric representations are considered to exist on a single two-dimensional surface. No overlaps of polygons are allowed without overlaps creating another polygon. With respect to lines, if lines cross, there must be an intersection at each line crossing. Additionally, if one line in topology moves, then the node representing the intersection also moves thereby maintaining the property of the intersection between the two lines. The same also applies to polygons. As one polygon moves, connected polygons will deform with it if the topology requires that the polygons always be jointed.

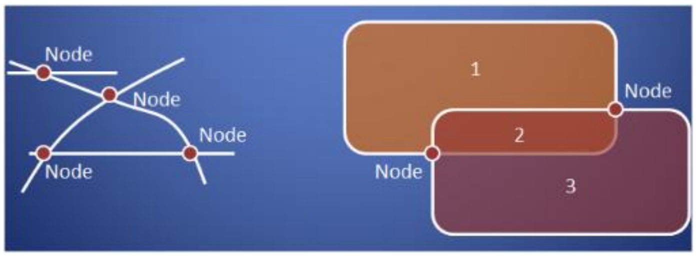


Figure 5 Planar Topology

Non-Planar Topology

In non-planar topology, features may be considered to exist on different planes, overlapping slightly at an edge. Therefore, it is not necessary to have intersections where lines cross if they are on different planes. The same goes for polygons.

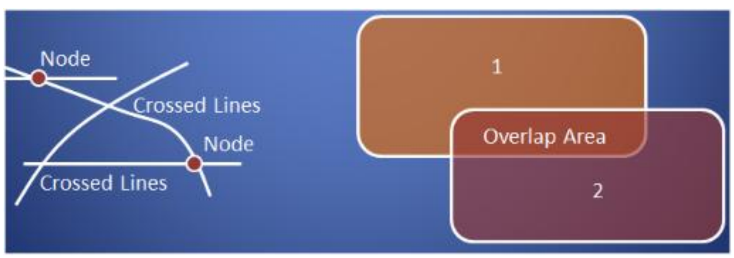


Figure 6 Non-Planar Topology

Advantages of Topology

There are several advantages to enforcing topology. The first advantage is that it ensures data quality. For instance, if you are representing houses built on parcels of land, you can set up topology rules that do not allow houses to be built across parcel boundary lines.

A second advantage of enforcing topology is that it can prevent sliver polygons. For instance, if you are digitizing county outlines, then it is important that adjacent counties share a common boundary. In this case, you can set topology rules that require adjacent polygons to share common vertices, and if one vertex is moved, the boundaries of all counties vertices are moved.

A third advantage of enforcing topology is that it ensures line connectivity. For example, if you are representing a street network, you would want all street intersections to fully connect, without any spacing between the end of one road and the beginning of the intersection. Topology assists in maintaining the integrity of data by enforcing logical rules.

Disadvantages of Topology

Although enforcing topology has its advantages, there are a few disadvantages. First, enforcing topology is a time-intensive process where humans must ensure that all lines connect, all polygons close, and all data begins and ends with node.

Second, humans must set up the topological rules for a computer to follow. The computer must expend additional processing time to build topological tables, and to maintain the connectivity and adjacency information. Additionally, the computer has to assign codes to features so that the individuals enforcing topology can keep up with the bookkeeping required.

SUMMARY

Vector data analysis – network analysis was the topic of this lesson. The lesson explained what network analysis is, how it works, benefits of network analysis, and impedances that affect the process. Network Analysis tools were explained with examples of choosing the closest route and administering the route selection through software. Various types of Topology were explained with examples and the advantages and disadvantages of each type.

ASSIGNMENTS

1. Lab: Vector Data Analysis – Network Analysis

2. Quiz: Vector Data Analysis – Network Analysis

Authoring Organization: Del Mar College
Written by: Richard Smith

Development was funded by the Department of Labor (DOL) Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant No. TC-22525-11-60-A-48; The National Information Security, Geospatial Technologies Consortium (NISGTC) is an entity of Collin College of Texas, Bellevue College of Washington, Bunker Hill Community College of Massachusetts, Del Mar College of Texas, Moraine Valley Community College of Illinois, Rio Salado College of Arizona, and Salt Lake Community College of Utah.

This workforce solution was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration. The solution was created by the grantee and does not necessarily reflect the official position of the U.S. Department of Labor. The Department of Labor makes no guarantees, warranties or assurances of any kind, express or implied, with respect to such information, including any information on linked sites, and including, but not limited to accuracy of the information or its completeness, timeliness, usefulness, adequacy, continued availability or ownership.

Lesson 7: Raster Data Analysis – Working with Topographic Data

INTRODUCTION

This lesson focuses on raster data analysis – working with topographic data. You will learn about triangulated irregular networks, Delaunay triangulation, and terrain analysis techniques. This includes information about aspect analysis, hillshades, and viewshed analysis. Reclassifying rasters is explained with content and examples to help you understand how to change information when necessary.

LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Create slope, aspect, and hillshade surfaces using raw elevation data.

2. Analyze environmental issues using elevation and derived data sets.

3. Reclassify raster data and use in a map algebra-based model.

4. Apply viewshed analysis to enhance site selection.

LEARNING SEQUENCE

Required Reading	Read the following: • Online Lesson Material
Resources	View the following: • None
Assignments	Complete the following: • Lab: Vector Data Analysis – Working with Topographic Data • Quiz: Vector Data Analysis – Working with Topographic Data

Vector Data Analysis – Working with Topographic Data

A triangulated irregular network is a vector-based data structure constructed by triangulating between points. It is an interconnected grid of triangles that represents typically elevation values for topographic data. As lines are drawn between all the input points, this creates a connected network of irregularly- shaped triangles.

Provided below are two examples of triangulated networks. Figure 1 is an example of multiple points connected together by edges to create contiguous triangles. Collectively, this is known as a triangulated irregular network. Figure 2 illustrates that if each point in the triangulated irregular network has a seed value, or what is commonly referred to as an elevation value, then the triangulated irregular network can be rendered in three dimensions.

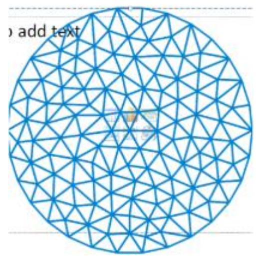


Figure 1 Points Connected into Triangles

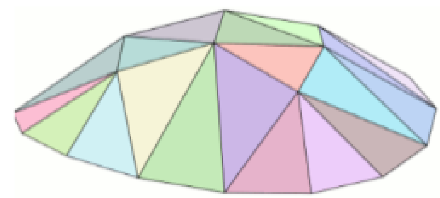


Figure 2 Example of a how a Triangulated Irregular Network can be rendered into three dimensions.

Data Sizing

The major advantage of a triangulated irregular network over a raster is that it can handle non-uniform spacing between the input points, thereby creating non-uniform triangle sizes. Additionally, over an area that is flat and smooth, fewer input points are required to accurately represent that area. The rougher terrains will require more data collection and more data points for an accurate representation. Review the following examples of data sizing and their outcomes.

Examples	**Outcomes**
Elevation of a Surface	This is a collection of input points representing the elevation of a surface. When the state is converted into an irregular network, a continuous surface is created.
Continuous Surface	This is the continuous surface created from the input points and is represented as a triangulated irregular network.
Triangulated Irregular Network	This is a triangulated irregular network of Lake Michigan. Note that in the middle of the lake, there are fewer points which saves space, and as you near the shores, there are more points, and smaller triangles, as the terrain becomes more complex.
Triangulated Regular Network	Triangulated irregular networks can also be used to represent more than elevation, such as a using a triangulated regular network to represent a dolphin. Notice how the nose has a denser amount of points requiring more triangles, but the body is smoother and therefore less sample points can be used thereby creating larger triangles.

Delaunay Triangulation

To create a triangulated irregular network, Delaunay triangulation is performed using all the input points. Delaunay triangulation is the computational geometry for triangulating points. No point can be inside the circumcircle of any triangle. A circumcircle is a circle that passes through all of the vertices of a polygon. Additionally, lines are not allowed to overlap in a Delaunay triangulation. There are many uses for triangulated irregular networks. Common uses are mapping physical land surfaces, mapping the sea bottom, executing slope analysis and aspect analysis, and three-dimensional rendering. Figure 3 is an illustration of the circumcircles in a Delaunay triangulation. If an input point does happen to exist inside of a circumcircle then a new set of triangles would be introduced to include that point.

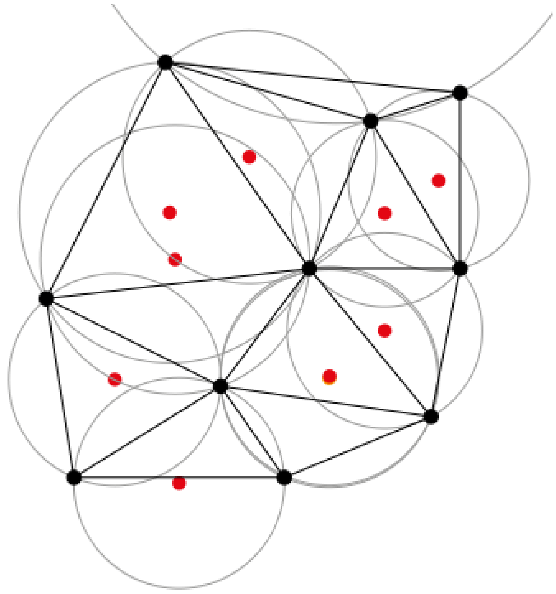


Figure 3 Circumcircles

Terrain Analysis Techniques

The slope analysis method creates a surface that displays the amount of slope of the terrain. The slope surface will display where the terrain is flat, moderately steep, and very steep. Slope is defined as rise over run where rise is the change of elevation and run is the horizontal distance. Therefore, if there is a large change in elevation over a short distance, then the slope is severe. Alternately, if there is a small change in elevation over a short or long distance, then the slope is closer to being flat. The slope analysis tool requires that the input be a raster and it will output a raster in return.

The following examples are provided to help explain when you might use slope analysis.

1. You may want to build a home on a large hill overlooking a lake; however you need to know the flattest portions of the hill to know where a house could be placed.

2. Perhaps you need to place a highway going up the sides of a mountain. In this case, you need to know the paths with the least severe slope so that the cars can climb the mountain. In both cases, the slope analysis tool will output a raster which defines severity of the slope in each cell of the raster.

Options for Customizing Slope Analysis Output

In addition to having the slope outputted in a percentage, you can also choose the slope grade to calculate. Additionally, you can classify each type of slope to simplify the display. For example, we can consider a slope that is greater than 15% to be considered a bad slope that we cannot use, and we could consider all slopes equal to or less than 50% to be good slopes, thereby creating a binary raster.

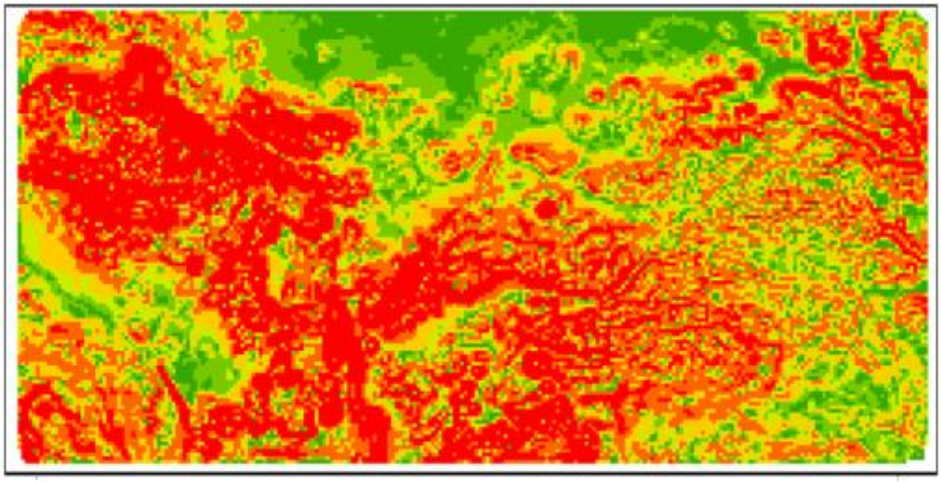


Figure 4 This image is an example of a classified slope analysis. The darker green color indicates areas where the slope is near 0% and considered to be flat. The darker orange color is used to indicate areas that have extremely steep slopes that have been measured near 90°.

What is Aspect Analysis?

Surface aspect displays the orientation of the terrain in each cell. It is also known as the slope direction, which tells us which direction the slope was spacing. When you run the aspect analysis tool, it represents the slope direction as an azimuth calculated from 0°, which represents north. Therefore, east would be 90°, south would be 180°, and west would be 270°.

The aspect analysis tool would provide output needed to answer questions such as:

- Which side of the mountain gets the most sun?
- Will I wake up with the sun in my face if we build on this side of the hill?
- Which side of the mountain will give me a better breeze in the evening?
- Which side of the hill will be better for my crops?

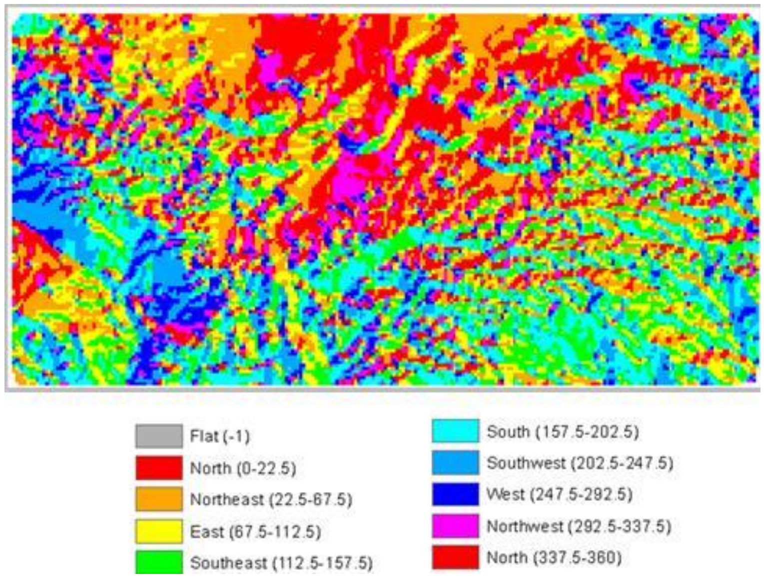


Figure 5 This is an example output from the aspect tool. For instance, green means that the terrain is facing southeast at an azimuth between 112.5 and 157.5.

What are Hillshades?

Hillshades model shadows on terrain given the location of the sun. A hillshade is also known as a shaded relief. One example for utilizing a hillshade would to apply a three-dimensional feel to a two-dimensional map. You could accomplish this by placing the hillshade at the bottom of the stack of layers, and making layers above the hillshade semi-transparent. Another use would be to model shadows in an urban area to determine whether a building would be placed in a shadow at any particular time of day based on its surrounding environment.

Figure 6 is an example of hillshade, the sun is located in the upper left-hand corner which is why the terrain that is facing northwest is lighter than the terrain to the southeast of ridges is darker as it is in the shadows of the ridges.

Figure 6 Hillshade

What is Viewshed Analysis?

A viewshed shows areas that are visible from a specific source point. It uses elevation values to decipher what is visible and what is hidden. It also calculates what features are impeding views and uses those when calculating line of sight to determine what is visible from the observer's position. One use for viewshed analysis could be to determine the view from a house before building begins. A viewshed can also let you know if a certain pasture is visible from a farmhouse, or for determining the view from our proposed highway.

Figure 7 is an example of a binary raster created from a viewshed analysis. Everything in black is considered as not being seen as there is no line of sight which would include any of the terrain from the observer's location. This is represented by the red circle in the upper right hand corner. The green cells in the raster are considered to be visible from the observer's location.

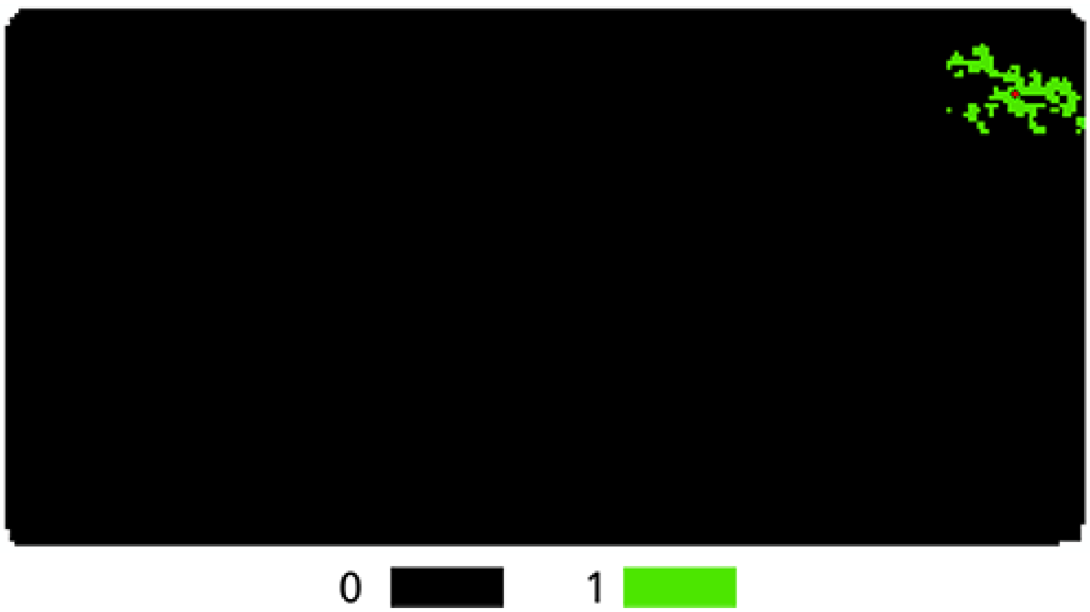


Figure 7 Example of a Binary Raster

Reclassifying Rasters

Raster reclassification is a process to change the values of each cell in a raster to other values. It will assign the new values based on user input parameters. You may wish to perform raster reclassification in situations where you want to change the current values based on new information. You can also use raster reclassification to replace specific values with no data values. Lastly, you may wish to simplify a raster by separating the cell values into classes, thereby reducing the number of unique values in the raster.

Example 1

In this example of a raster reclassification (Figure 8), two inputs are required: the input raster, and the reclassification table. The reclassification table tells the computer how to translate the input values into the output values. The raster recalculation tool uses two inputs to output a brand-new raster with the reclassified values. In this example, we are simply changing one unique value in the input raster to a different unique value in the output raster.

K	L	F
L	O	N
X	F	K

Input	Output
L	2
K	9
F	4
X	7
O	5
N	3

=

9	2	4
2	5	3
7	4	9

Figure 8 Example 1 Raster Input and Reclassification Table

Example 2

In this example, again we have an input raster and a reclassification table as the input. However, in this case, notice that we only have two unique output values to the six unique input values. Therefore, the output raster will represent the data in simpler terms.

K	L	F
L	O	N
X	F	K

Input	Output
L	Yes
K	Yes
F	Yes
X	No
O	No
N	No

=

Yes	Yes	Yes
Yes	No	No
No	Yes	Yes

Figure 9 Example 2 Raster Input and Reclassification Table

Map Algebra

Map algebra is the combination of raster layers in a cell by cell process. Map algebra can be a simple operation such as multiplication, subtraction, and addition, or more complicated operations that may or may not use set algebra or Boolean algebra.

In this example of map algebra (Figure 10), there are two input rasters to the left of the equals sign and an output raster located to the right of the equals sign. The map algebra operation is subtraction; therefore the second input raster subtracts its values from the first input raster where the cells overlap resulting in a new raster. The new raster shows the results of the subtraction that occurred between the two input asters on a cell-by-cell basis.

1	5	2	3
5	2	6	1
2	6	3	7
3	1	7	4

-

4	2	8	6
3	1	3	5
2	0	7	1
1	9	3	2

=

-3	3	-6	-3
2	1	3	-4
0	6	-4	6
2	-8	4	2

Figure 10 Map Algebra Example

Weighting

When analyzing data, some feature values may be more important than others. In order to reflect the importance of some features over others, we can weight them differently according to their level of importance.

Examples of Weighting:

- Slope is often more important when discussing issues of where runoff will flow.
- Aspect can be more important when deciding how much sunlight is necessary to plant certain crops.
- Viewshed can be even more important when deciding where to place an establishment depending on where you will be able to see it from.

Ranking the Values

Different layers have different levels of importance. It is important to rank the criteria for each layer. The ranks can be binary or they can even be incremental. Consider the following example.

Suppose you want to build a house and the slope of the terrain needs to be less than 4°, you want your bedroom window to face west, and you want to be within a distance of 20 miles to a grocery store.

In order to complete the weighting, you must first rank the criteria.

Step 1: Decide which factor is the most important. For example, when you are determining where you can build a house and understand that you cannot build a house on a slope greater than 4°, slope becomes the most important factor. Slope is often more important when discussing issues of where runoff will flow.
Step 2: Determine if you plan on growing your own crops or if you would like to obtain food by driving to a grocery store. A food source becomes the second most important factor, if you do not plan on growing your own crops

Step 3: Determine the importance of the view from your bedroom window. When you are overlaying these rasters for analysis, you can apply a large multiplier to accentuate some layers, and apply a small multiplier, or even a fraction to diminish the importance of other layers. The results will be biased towards the layers that have higher ranks and multipliers.

SUMMARY

In this lesson you learned about raster data analysis – working with topographic data which includes triangulated irregular networks. Examples explained the various types of triangulated irregular networks and data sizing elements including the elevation of a surface and continuous space. The lesson also demonstrated Delaunay triangulations and how they are created. Slope analysis and options for customizing slope analysis output were explained in this lesson. The reclassification of rasters process was explained and examples were presented to further your understanding of this process.

ASSIGNMENTS

1. Lab: Vector Data Analysis – Working with Topographic Data

2. Quiz: Vector Data Analysis – Working with Topographic Data

Authoring Organization: Del Mar College
Written by: Richard Smith

Development was funded by the Department of Labor (DOL) Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant No. TC-22525-11-60-A-48; The National Information Security, Geospatial Technologies Consortium (NISGTC) is an entity of Collin College of Texas, Bellevue College of Washington, Bunker Hill Community College of Massachusetts, Del Mar College of Texas, Moraine Valley Community College of Illinois, Rio Salado College of Arizona, and Salt Lake Community College of Utah.

This workforce solution was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration. The solution was created by the grantee and does not necessarily reflect the official position of the U.S. Department of Labor. The Department of Labor makes no guarantees, warranties or assurances of any kind, express or implied, with respect to such information, including any information on linked sites, and including, but not limited to accuracy of the information or its completeness, timeliness, usefulness, adequacy, continued availability or ownership.

Lesson 8: Raster Data Analysis – Density Surfaces

INTRODUCTION

In this lesson you will learn about raster data analysis – density surfaces. Map density is defined and explained with examples to further your understanding on this topic. You will learn about converting between vector and raster data and view examples of how this process occurs. Polygon to raster conversion and raster to raster to vector conversions are also part of this lesson.

LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Construct data density surfaces from point data using appropriate methods.

2. Convert between vector and raster formats.

3. Develop approach to address questions using density techniques.

LEARNING SEQUENCE

Required Reading	Read the following: • Online Lesson Material
Resources	View the following: • None
Assignments	Complete the following: • Lab: Raster Data Analysis – Density Surfaces • Quiz: Using Advanced Attribute and Spatial Queries for Data Exploration

INSTRUCTION

Density

Mapping density allows the user to see where large numbers of observations or higher values occur. Densities show the spatial relationship amongst the different locations of data and also allow us to create predictions from the data as to where other locations or values will occur.

Here are a few examples of why and when density maps would be created.

- You would create a density map if you wanted to locate the area with the most or least hospitals which may be an indicator of health care coverage.

- The survey they ask you to do at the store asks your ZIP code, phone number, address, or other things which allow the store to map the density of where their customers are mostly coming from which may allow them to do targeted marketing.

- Mapping population densities can be used to make determinations of where the population lives, where public services may be needed, or where workers are traveling to and from.

Density Analysis: Point Density

Point density calculates the density of point features across a specific area. It accomplishes this by adding the point to the specified study area and dividing the total by the area of the study area. For example, take a look at Figure 1. If we consider this data set showing the locations of cities in Texas, we can visually see that there are dense groupings of dots; however, it may be difficult to visually discern which of the three major areas are denser.

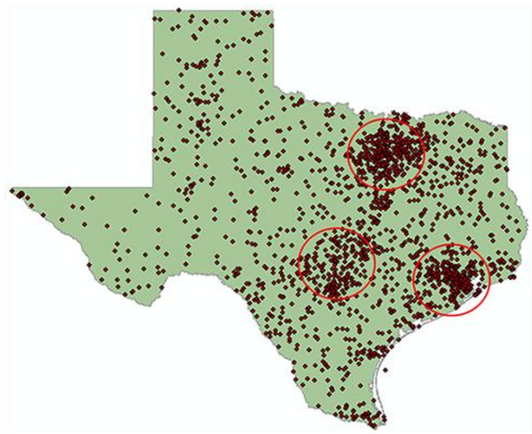


Figure 1 Data set showing locations of cities in the state of Texas.

f we run a density analysis, the results assist us in visually determining where the densest areas are located. This density map (Figure 2) is derived from the point locations of all the Texas cities, and clearly show that the Houston area as the most concentrated density of cities with the Dallas-Fort Worth area coming second and the Austin/San Antonio area coming third.

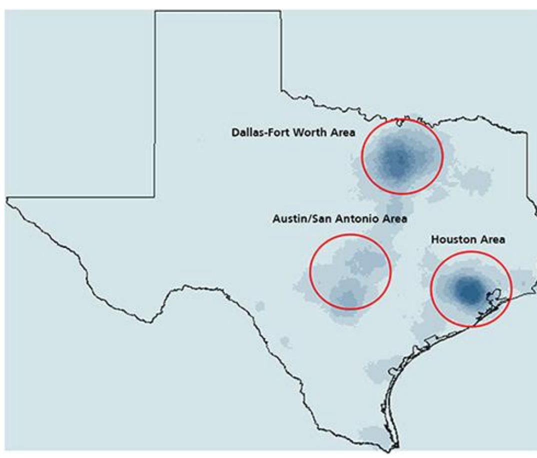


Figure 2 Density Map

Density Analysis: Line Density

Line density is similar to point density as it calculates the more dense areas versus less dense areas containing line features. In Figure 3, the data set shows all the locations of streets in Nueces County, Texas. Circled in red are what appear to be three most densely paved portions of the county. If we run a line density analysis on the streets, we can get a simpler representation of where the roads are most dense in this county.

Figure 3 Data Set: Streets in Nueces County, Texas

In Figure 4 the line density raster confirms that the three circles encompass the most densely paved portions of the county. What is interesting to note on this density map, is that it generally shows the connectivity between the different cities in the county and how strong those connections are with regards to roads.

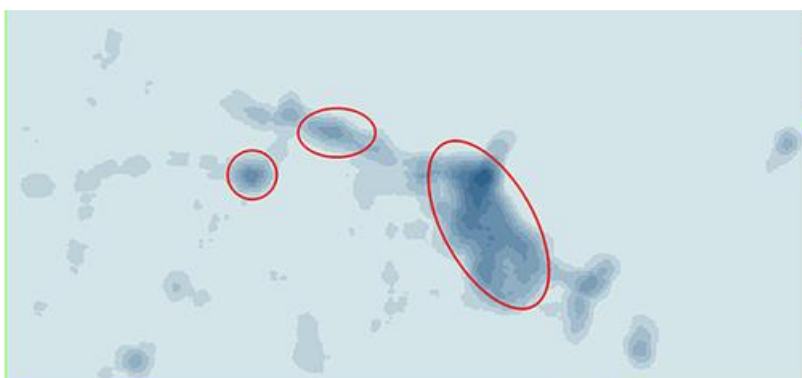


Figure 4 Line Density Raster

Density Analysis: Kernel Density

The kernel density function can calculate density of either line or point features using a kernel function. The kernel density function allows some features to hold more weight than other features based on specific values. For example, when searching for the density of fire hydrants in an area, some hydrants provide more water pressure, thus covering a wider area than others; these fire hydrants will hold more weight than those that have a lower water pressure, when the kernel density function is executed.

Differences in Density Functions

There are a few important differences between kernel density and point and line density functions.

- Kernel density tends to create smoother density formations.
- Kernel density can also omit features with no data values, essentially giving them a weight of zero.
- The kernel density function assigns the highest values to the center of the point or line in question and tapers the values down to zero as you move out to the edge of the search radius

To illustrate kernel density and how it produces different results than point density, take a look at example maps provided in Figure 5. Note that the kernel density function creates a smoother density map as the densities are not as dispersed as seen within the point density map.

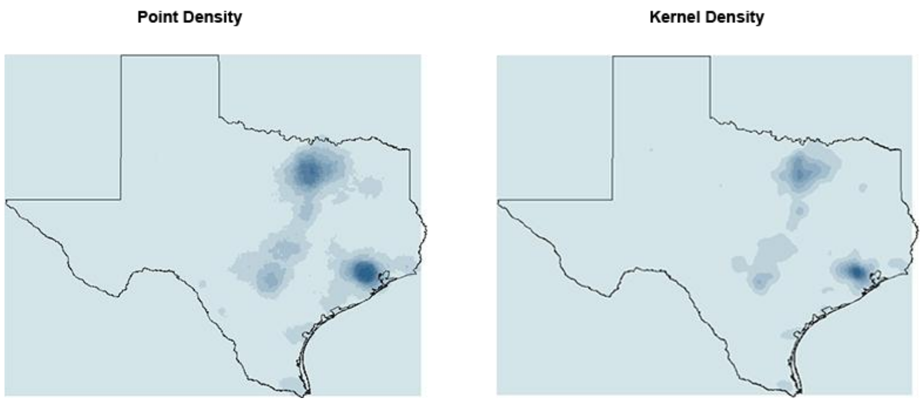


Figure 5 (left) Point Density, (right) Kernal Density

Converting between Vector and Raster Data

You can convert points, lines, and polygons to rasters. You may want to do this if you wish to allow the user to get a continuous view of the previously discrete data. One major issue to keep in mind in regards to converting vector data to raster data, is that the overall accuracy of the existing data may be degraded due to the logistics of expanding a small point or line to an entire cell or cells. This could create misconceptions regarding where certain areas truly exist.

Point to Raster Conversion

When converting a point to a raster, if there is more than one point in a cell when creating the raster, the point with the most common value will be selected. If there are multiple points with the same amount of values, the points with the lowest "Feature ID" (FID) will win out. Otherwise, the value of the cell can be determined in an alternate manner of your choosing. Priorities of which point will be set as the raster's cell value can be set. Points with the highest priority will be selected above all others.

The example provided illustrates converting points to a raster (Figure 6). Note that when there are multiple points in a cell, the cell takes on the value of the points with the most common value. If there is a tie, then the cell's value will be equal to the point with the lowest FID value.

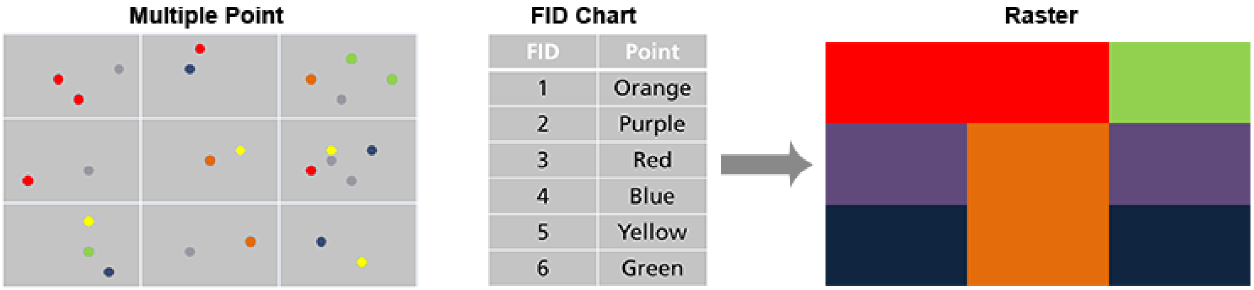


Figure 6 Converting a Point to a Raster

Line to Raster Conversion

When converting lines to a raster if there are multiple lines in a cell, the lowest FID will again be selected. If there are priorities, again, the highest priority is selected. A common method of prioritization is to select lines based on the longest length as being the highest priority. The example provided illustrates converting lines to a raster.

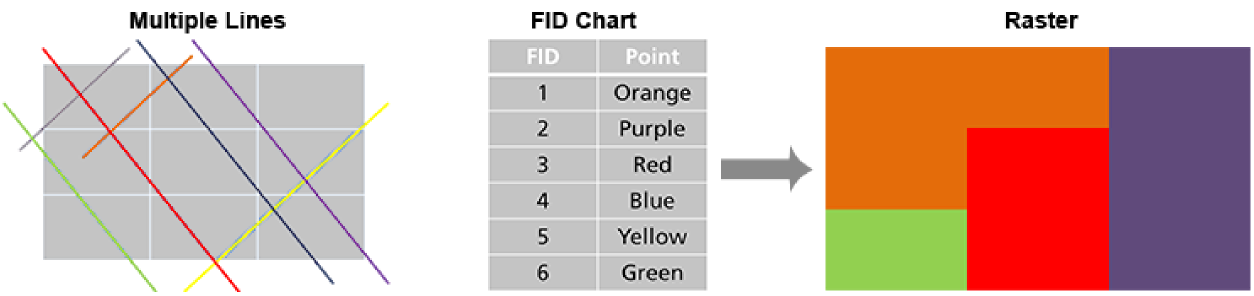


Figure 7 Converting Lines to a Raster

Polygon to Raster Conversion

You can convert polygons to a raster, however the determination of which polygon's value to set to the output rasters cell value is a little more complicated. There are three predominant methods of making this determination. The three methods are maximum area, maximum combined area, and cell center.

Maximum Area

The maximum area method chooses the polygon with the highest total area in a cell, provided it covers over 50%"of the total cell area and has no other selection priorities set. If there are multiple polygons with the same area covered, then the polygon with the lowest FID is selected. There are other priorities that you can set, and those priorities always take precedence over area covered.

Maximum Combined Area

The maximum combined area method considers how much coverage a feature layer has within a cell. This means that, if one feature layer has two polygons inside the cell that cover more than 50%, but neither feature covers more than 50% alone, then neither of the features are selected.

If more than one feature has the same size the lowest FID wins. Highest priorities are always selected first, and if the priorities of the same as well in the area with the lowest FID is once again selected.

Cell Center Method

In this method, the polygon that covers the center of the cell is selected as long as no user-set priorities are set to justify. If multiple polygons fall within the center of the cell, the polygon with the lowest FID is selected.

Consider the Input Polygons chart, and its' associated Entry Table, as well as the methods of Maximum Area and Cell Center images provided. Note that even though the input is the same, choosing "the maximum area method" versus "the cell center method" creates significantly different output rasters.

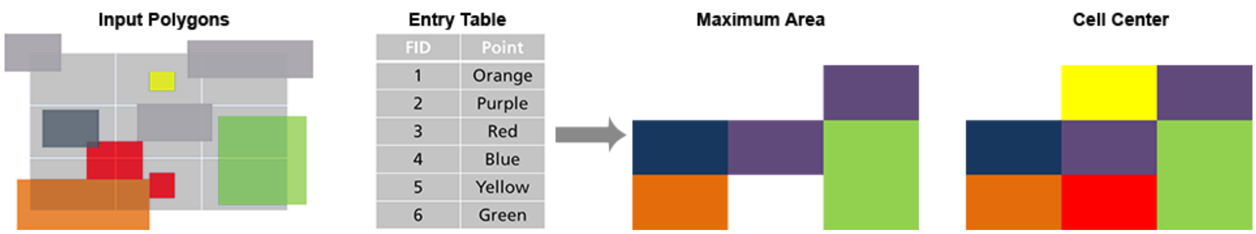


Figure 8 Maximum Area Method versus Cell Center Method

Raster to Vector Conversion

It is possible to convert a raster to a point, line, or polygon. This allows the user to complete different analysis by having any vector format.

Raster to Point

When converting a raster to points, each individual cell in the raster is used to create a corresponding point. The point will be created at the center of the cell. Any cell size will work for this conversion. Outputted points will contain the value of the cell. Figure 9 is an example of a raster converted to a point. Note that each point that has been created maintains the same value of the cell it was derived from, and if a raster cell had no data, there was no point created.

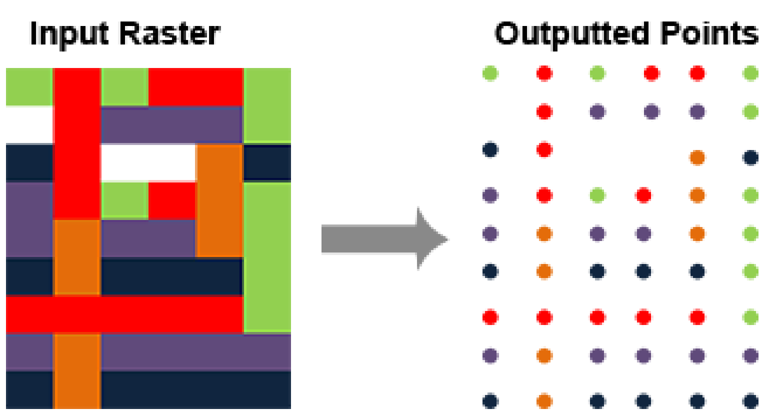


Figure 9 Raster Converted to a Point

Raster to Line

When converting a raster to a line, any raster cell that has a value greater than zero will be successfully converted. The outputted lines will pass through the center of the cells. If there are no data values they do not become features. Figure 10 is an example of a raster that is the converted to a line. Note that the lines do not exist were cells had a no data value.

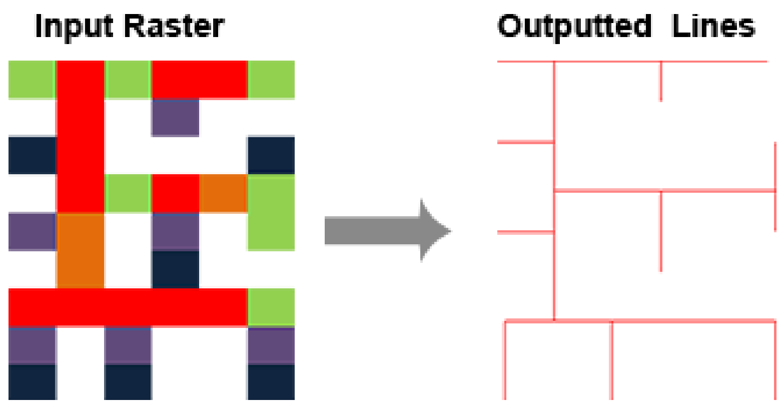


Figure 10 Raster Converted to a Line

Raster to Polygon

When converting a raster to a polygon, you can choose which attribute from the raster to use in the output feature class. Take a look at Figure 11. The outputted polygons will be the same size as the inputted cells. It is important to note that where a cell had no value, no polygon is created. Additionally, recognize that the cells which are adjacent that have the same value are combined to create a larger contiguous polygon.

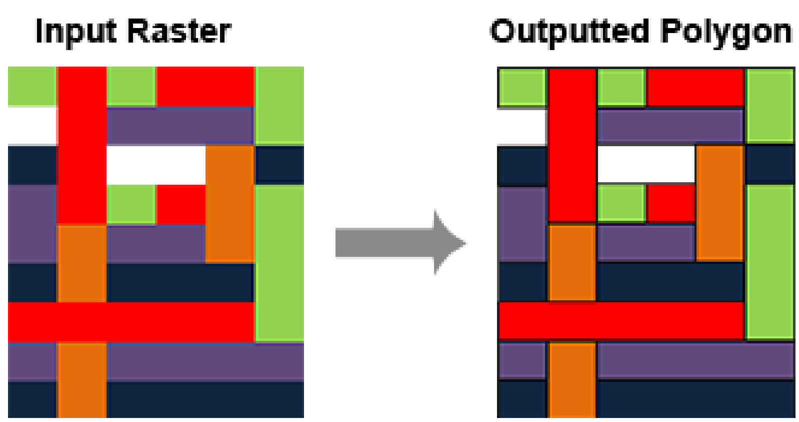


Figure 11 Raster Converted to a Polygon

SUMMARY

This lesson covered the topic of raster data analysis – density surfaces. You learned why and when density maps would be created with examples to help you understand. Point density, line density, and kernel density were illustrated to show you the difference of each and how they are used. This lesson also explained converting between vector and raster data with images to show the processes. The conversion from polygon to raster was also covered including the three methods which are maximum area, maximum combined area, and cell center. The raster to vector conversion process was also explained with examples to further your knowledge on how this process occurs.

ASSIGNMENTS

1. Lab: Raster Data Analysis – Density Surfaces

2. Quiz: Using Advanced Attribute and Spatial Queries for Data Exploration