LECTURE 5: VECTOR DATA MODEL

Section I - Review

• Conceptualizing Geographic Entities

As GIS operates on the premise that features in the real world can be represented in a digital format - referred to as **ontology**. At the level of an individual geographic entity there are two possible ways of encoding geospatial data into a GIS.

- Vector Graphics – Commonly referred to line drawings or illustrations where an object is represented as a point, line, or polygon.

- Raster Graphics – Where an object is represented as a series of pixels (*e.g.* photo).

Section II – Vector Data Model

• Overview of the Vector Data Model

Vector data models are best suited to geographic entities that have discrete and sharp boundaries. Additionally, vector data is free to define geographic entity at any spacing in a non-uniform manner. Check out Figure 4.8 in Campbell and Shin.

Points or nodes reflect specific locations on a map. Some examples of point geographic data include wells in a county, landmarks, cities on a world map. This is a zero-dimensional feature.

Two points (**vertices**) connected by a line segment define a line. Examples of lines include: roads, streams, pipelines, transmission lines. The starting point for a line is referred to as a node. This is a one-dimensional feature.

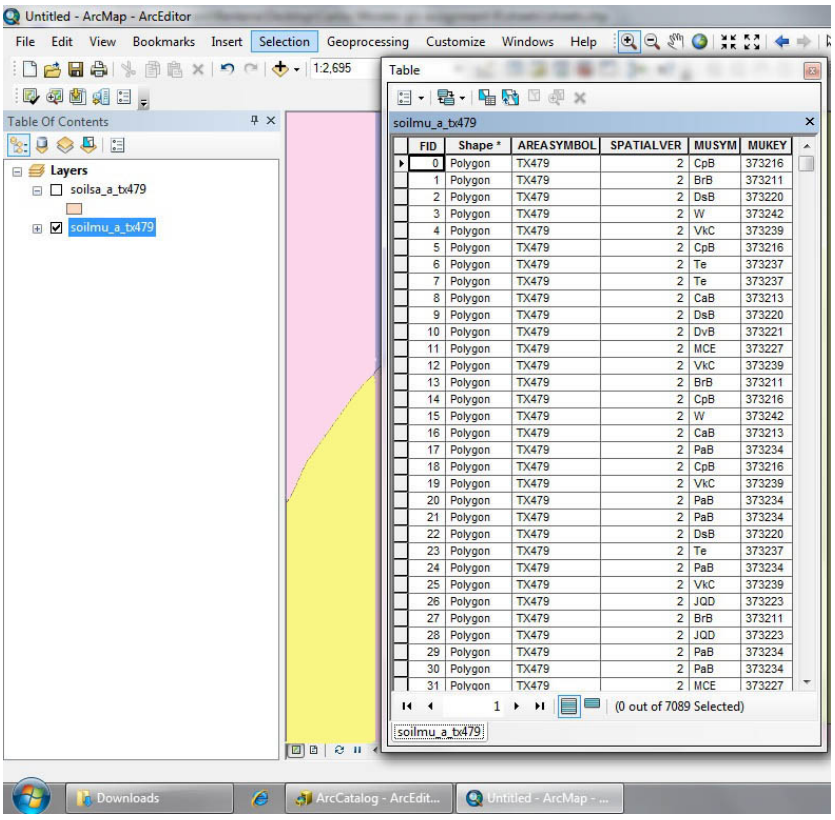
Multiple line segments can be connected to form **polylines**. Polylines that close on themselves form **polygons** and include features such as counties, tax parcels, land use, soil types. This is a two-dimensional feature.

Finally, recall the discussion from last week's lecture of generalization; how features are represented (points, lines or polygons) depends upon the scale of the map. The importance of generalization cannot be unstressed and can profoundly influence how a geographic entity is represented in a GIS. For example, in a small-scale map a city is represented as a point; whereas, in a large-scale map a polygon is used to represent the cities boundaries.

• A Note on Significant Digits?

Again, the vector data model is best for describing discrete geographic entities that have sharp boundaries such as streets and buildings. However, the vector model is commonly used to define entities that have fuzzy and less discrete boundaries. An example of this is a vector depiction of soil types (Figure 1). Two adjacent soil units are separated by a line that is sharp and distinct with an inferred accuracy that could be measured in a GIS to many digits pass the decimal point (for example, 200312.234324 ft). Do all of these numbers pass the decimal point have real meaning? Can you subdivide these soils along a boundary that is microns thick? No! **You need to write numbers down to the proper level of significant digits.**

Decimal Degrees	##.######°	(Six digits pass decimal point)
Planar Coordinates	##.# m or ft	(one digit pass the decimal point - unless you have a really good GPS unit)



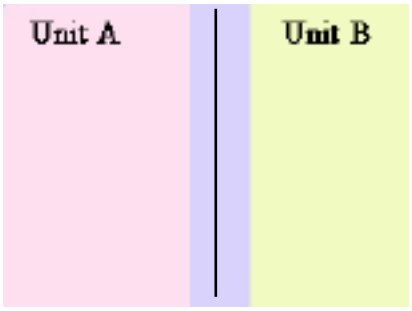


Figure 1 (previous page). Soil units 1 and 2, which in a typical GIS have a sharp line as a boundary but in reality, the boundary between these units is gradational as indicated by the blue zone between units A and B.

• Is the Vector Data Model Always the Best Choice?

You should realize by now that the stark conceptualization associated with the vector data model is not always totally valid. Let's go back to the example of our two soil units from Figure 1. This is a classic example of GIS data taking on a life of its own. We now realize that there is some significant uncertainty in the location of the boundary between these two soil entities (A and B). The question you may have is then why are we using a vector data model to define these entities when we know that these boundaries are highly uncertain. Looking at Figure 1 we can understand that we have two soil units (A and B) with a third transition zone in between the two units where the characteristics of unit A blend into unit B. So perhaps the selection of a vector data model is not appropriate. Soils seem to behave more like continuous phenomena as opposed to a discrete entity. So why not represent these soils in a raster data model? The short answer is that raster data models typically can only encode one attribute value, whereas, vector data models can have numerous attributes. This is why soil vector files exist. The user needs to be mindful that while this data is useful it may not be appropriate for large-scale, highly detailed mapping applications that occur in the vicinity of soil unit boundaries.

Section III - Vector Data Approaches

• Spaghetti Data Approach

There are two data storage approaches to vector data. An ESRI shapefile is non-topology based. Another way of stating this is that a spaghetti model is used to store each entity separately as shown in Figure 2. A way to think of this is to imagine digitizing a pot of spaghetti; each strand is essentially isolated from each other even if there are touching each other. Figure 2 shows the issue that can arise in topological data sets. These issues include: a dangling arc, which overshoots the intended intersection point and undershoots and gaps where line segment to not connect together as they should. To convert this data into a topological correct network problems must be addressed through the editing process.

• Topological Data Approach

An ESRI coverage (and other formats) is topology-based meaning that all the entities are interconnected as shown in Figure 3. Topology is important if you want to be able to use GIS data to do more then just visualize data. For example, advanced query features such as routing and network analysis require that your dataset is topologically sound. Basically. all your streets are connected with each other and are encoded properly. Even tiny gaps or overlaps can completely throw off your system in a topologic data set.

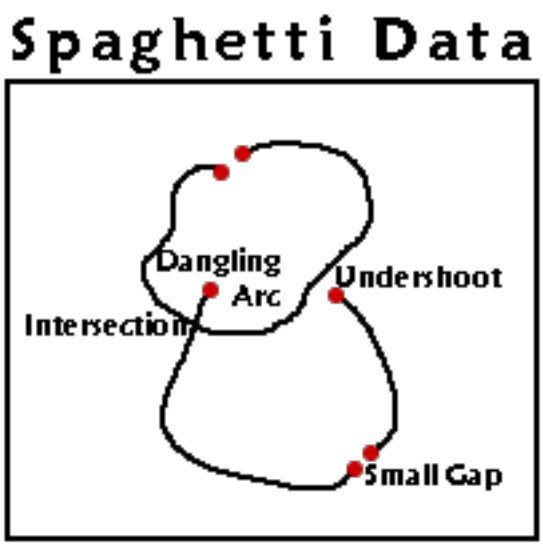


Figure 2. Two-line segments in a non-topological layer.

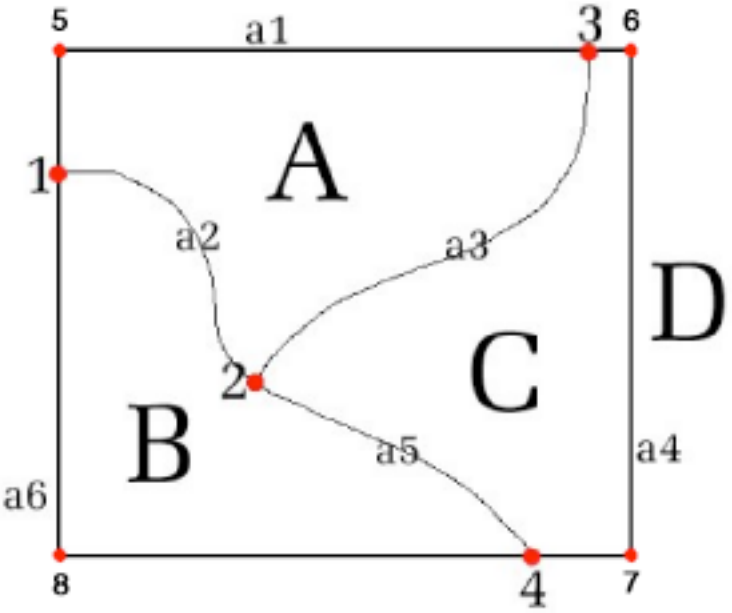


Figure 3. Multiple line segments forming three polygons in a topological correct data set.

A nice discussion of spaghetti versus topological approaches is provided in Campbell and Shin (Chp. 4.2). Also check out DiBiase Section 4.5

ERSI File Formats

Remember that there are two approaches to defining vector data and ESRI has file formats for both of these approaches.

* **Coverage** - Esri's closed, hybrid vector data based on topological relations. This is a legacy format originally designed for ArcGIS Workstation / ArcInfo two decades ago.

* **Shapefile** - without topological relations

Shapefile files consist of a number of files that have spatial data associated with them some of which include:

.shp stores the feature geometry

.shx maintains the spatial index of the feature geometry

and attribute information as is the case with database file (.dbf)

Warning: Do not copy a .shp file in Windows Explorer. For the file to work all associated files are required

US Government File Formats (Topological formats)

* **Digital Line Graph** (DLG) - a USGS format for vector data for topographic maps
* **TIGER** - Topologically Integrated Geographic Encoding and Referencing - from US Census

Other File Formats

* **Geography Markup Language** (GML) - XML based open standard (by OpenGIS) for GIS data exchange
* **Keyhole Markup Language** (KML) - XML based open standard (by OpenGIS) for GIS data exchange. Commonly used in Google Earth.
* **AutoCAD DXF** - Contour elevation plots in AutoCAD DXF format. Commonly used by architects and surveyors (spaghetti and non-georeferenced data)

Section IV - Overview of Attribute Data

• Introduction

GIS has an advantage in that it can be linked to database management system (DBMS). Digital data is stored as files (called tables). A database is a collection of tables. Tables are generally in tabular form. Rows are known as records corresponding to individual geographic entities, typically geographic objects represented in a vector format. Columns reflect attributes associated with each entity, which are referred to as fields or attributes. Typically, each vector entity (point, line, polygon) can have spatial attributes that define the geographic location of the entity. Additionally, any number of non-geographic attributes can be included. Conversely, each record in the **attribute table** is associated with a particular map feature. Table 1 illustrates an attribute table that contains data linked to each geographic entity, in this case Dr. McReynold's scorpions. The data contained in the attribute table provides additional information about the scorpions observed.

• Attribute Data Formats

Note that the spatial data is stored separately from the database of attributes and the two sets of files are linked with a Feature ID number (typically the left most attribute or field in an attribute table). The essence of a **database management system (DBMS**) is that rows correspond to a mapped entity; whereas, the columns define the attributes associated with that entity.

Note that within an attribute table there are both geographic and non-geographic fields of data. The ID, coordinates, and positional references are geographic in nature.

The attribute table in a vector file there has a great deal of flexibility in the types and number of non-geographic attributes (or fields) that can be stored and several data format types can be used.

Number (integer ## and floating number - with .####)
Text (short and long string)
Date
Binary large objects (images, multimedia, video, etc.)

Answer the following question based on examining the attribute table depicted in Table 1.

Identify all geographic attributes

_____ _____

Identify all non-geographic attributes and note their formats (text, numeric, …).

_____ _____

_____ _____

_____ _____

_____

What type of vector data is this?
(point, line, or polygon) _____

ID	EAST	NORTH	Date	Time	MarkSizeClass	Location	Height	Prey	Flag
1	684057.6096	17099954.35	2/16/06	21:11	3	ground	0	0	ADS
2	684060.4225	17099946.56	2/16/06	21:20	2	grass	5	0	ADS
3	684072.9915	17099941.42	2/16/06	21:27	1	guajillo	70	0	ADS
4	684058.1735	17099957.55	2/16/06	21:31	1	ground	0	0	ADS
5	684074.4782	17099975.01	2/16/06	21:34	3	Opuntia	30	0	ADS
6	684054.7805	17099966.53	2/16/06	21:40	2	grass	5	0	ADS
7	684084.4272	17099969.01	2/16/06	21:47	2	grass	5	0	ADS
8	684080.6031	17099973.83	2/16/06	21:49	2	Blackbru	35	0	ADS
9	684057.6661	17100016.79	2/16/06	22:05	3	guajillo	40	0	ADO
10	684079.1438	17100022.74	2/16/06	22:11	2	ground	0	0	ADO
11	684157.2014	17099984.14	2/16/06	22:28	4	trunk	35	0	AD1
12	684158.02	17100000.6	2/16/06	22:32	1	grass	5	0	AD1
13	684139.5889	17099983.66	2/16/06	22:41	3	guajillo	55	0	AD1
14	684133.1305	17099988.38	2/16/06	22:44	2	grass	5	0	AD1
15	684125.1963	17099985.82	2/16/06	22:47	4	grass	5	0	AD1
16	683846.3468	17099853.59	2/21/06	21:26	2	Blackbru	105	1	ABS
17	683855.6243	17099843.13	2/21/06	21:32	2	deadbranch	65	0	ABS
18	683860.5047	17099849.79	2/21/06	21:32	2	Blackbru	150	0	ABS
19	683833.843	17099862.47	2/21/06	21:40	3	guajillo	70	0	ABS
20	683851.5445	17099900.26	2/21/06	21:48	3	guajillo	85	0	ABO
21	683847.3921	17099935.59	2/21/06	21:54	3	desert olive	100	0	ABO
22	683850.9261	17099924.01	2/21/06	22:03	2	Opuntia	15	0	ABO
23	683892.1292	17099862.28	2/21/06	22:16	1	Blackbru	135	0	AB1
24	683879.5843	17099890.89	2/21/06	22:22	3	tasajillo	25	0	AB1
25	683712.7985	17099851.35	2/21/06	22:48	2	olive	45	0	AAO

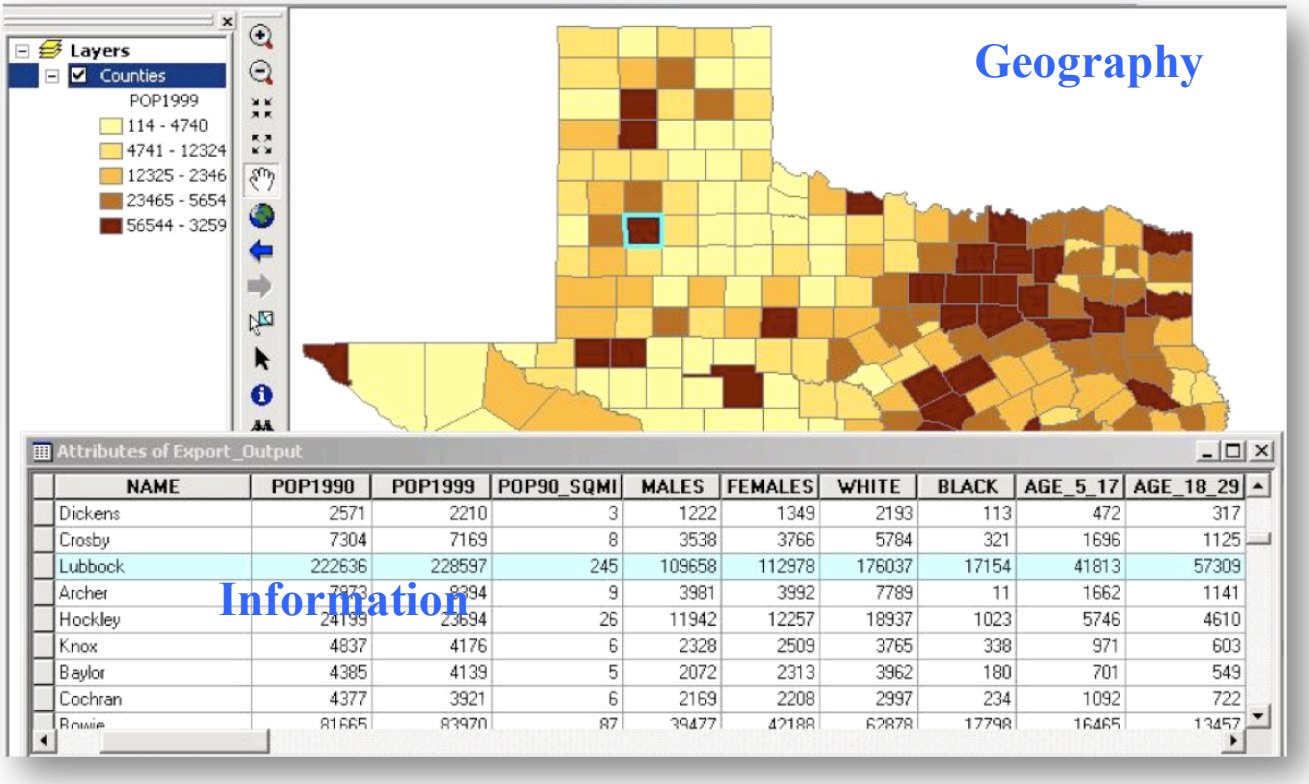


Figure 4. Texas county map layer with attribute table shown. Lubbock county is selected; shown in blue.

In other words, a GIS links features in the real world with information about those features as you can see in Figure 4. By selecting a county extensive information becomes available through an **attribute table**. The map display can be thought of as the tip of the iceberg. Most of the data is hidden below the surface in the attribute table. Additionally, you can select multiple counties based on the value of any attribute. For example, you can set up a query that will select only counties with a population greater than 200,000 allowing one to visualize all of the urban area throughout the state. GIS allows the use to connect geographic data with any type of attribute information.

Question: What type of vector data is utilized to depict the counties in Figure 1?

Points or Lines or Polygons

• Levels of Attribute Data

Additionally, geographers have devised a system for classification of attribute data beyond just a basic description that you observe in an attribute table. Basically, all attributes can be either non-numeric or numeric in nature. Furthermore, there are two subtypes within each of these classifications resulting in a total of **four fundamental levels of data**.

Non-numeric data that is unranked, lacks hierarchy, is referred to as **nominal** level attribute data

An example of nominal data is different types of land cover around Laredo indicated below:

Rangeland
Urban
Agricultural

There is no implied ranking in the above list and therefore this data is classified on a nominal basis.

If non-numeric data is ranked or has a hierarchy then this reflects **ordinal** level attribute data. An example of ordinal data is the different types of roads, which has an implied hierarchy as indicated below:

Most Traffic — Interstate
US Highway
State Highway
County Road (Paved & Unpaved)
Least Traffic — Trail

The functionality of a geographic entity is directly related to the hierarchy established by the ordinal classification scheme. For example, if you wanted to evacuate a region you would more likely direct the public to take the interstate that can accommodate more traffic than a small county road.

For numeric attribute data, classification is based on whether the number is tied to an absolute measure scale. If numeric data is not based on an absolute scale this represents **interval** level attribute data. A classic example of an interval scale is the Celsius temperature - 100^{o} C is not double the temperature as 50^{o} C. Another example of interval

data is year. The year 2000 is not double of the year 1000 as our calendar is based on an arbitrary, non-absolute, system for measuring time.

If numeric data is based on an absolute scale this represents **ratio** level attribute data. For example, age is ratio because 10 years old is double the age as 5 years old. Note that temperature expressed in Kelvin's is considered ratio level data because the Kelvin scale is based on a temperature of absolute zero. Note that all mathematical operations can be performed on ratio level data

Also, note that ratio and interval data are frequently grouped into ordinal level categories for **thematic mapping** (Lecture 10 and Assignment 4). Why would you want to do this? You cannot make a map with hundreds of colors but can make one with five or six that conveys your message/point. The human eye has limits on the number of colors it can discern on a map and it is a good idea not to exceed 11 colors. Too many colors will make your eyes "busy" and it will be difficult to interpret the map's meaning.

Now let's go through Dr. McReynolds attribute table (Table 1) again and classify the attributes in this table both based on data format and level of data. Do not confuse data format with level of data - data format is how the data is represented in the attribute table; whereas, level of data indicates the functionality and utility of how data can be used in GIS applications.

For the following attributes indicate the level of data represented (nominal, ordinal, interval, ratio)

MarkSizeClass _____

Location _____

Height _____

Prey _____

Readings

GIS Commons webpage; Chapters 1, 4.
Campbell, J and Shin, M., 2011, Essentials of Geographic Information Systems. Chp. 4.1.
DiBiase, D., 2014, Nature of Geographic Information Systems. Sections 3.7 to 3.10; 4.3 to 4.5.

Terms

Node	Point	Polyline	Polygon
Vertices	Ontology	Spaghetti Approach	Intersection
Small Gap	Undershoot	Topologic Approach	Dangling Arc
Coverage	Shapefile	TIGER	DLG
Attribute Table	DBMS	Ratio	Interval
Nominal	Ordinal	Thematic Mapping	

Concepts

What is the difference between the file format type and data format type in an attribute table?
Understand how many significant digits are valid for features in the geographic and planar coordinate systems
Why would a user select a vector data model for a non-discrete feature like soils or land cover?
Understand the different levels associated with attributes in a DBMS
What are the advantages of using topological datasets? Any disadvantages?
Know about the common file formats used for vector data

HOMEWORK

1. In your own words describe the difference between spaghetti and topological vector data.

2. Comment on the following statement. In some cases, a vector depiction of an entity can provide a false sense of accuracy regarding the extent of an object.

3. Despite the obvious limitation of a shapefile, this format to this day remains one of the most popular vector file types. Since ESRI is the GIS software giant this partially explains the popularity of this legacy file format. Can you think of another reason for the long lasting appeal of shapefiles?

4. Vector data consists of all but the following

(a) Points (b) Lines (c) Polylines (d) Pixels (e) Polygons

5. Which of the following expresses the correct number of significant digits associated with a latitude measurement.

(a) 27.45^{o} (b) 27.4512^{o} (c) 27.451254^{o} (d) 27.45125467^{o}

Examining the data associated with Exercise #3 and answer the following review questions.

6. Fill out the data below based on examining only your vector data

Layer Name	**Data Model**	**If Vector Type (Pt, Line, Polygon)**	**If Vector Data Approach**

For Data Approach Select either Spaghetti or Topological

7. Examining the data packet provided and answer the following review questions. Format (text versus number, if number is it integer or decimal) and level (ordinal, nominal, ratio, interval)

	Format	Level
Stream Layer (LENGTHKM – Attribute)	_____	_____
Water_Bodies Layer (GNIS_Name – Attribute)	_____	_____
Roads Layer (RTTYP – Attribute)	_____	_____
States_Layer (REGION – Attribute)	_____	_____
Hydrologic_Units_Layer (LOADDATE – Attribute)	_____	_____

8. In the attribute table sort by clicking on top of the RTTYP (Road Type) column. You want to find the number of line segments used to represent interstate (I).