

# **Geographic Information Systems**

**Notes For OC 3902/ OC 2902**

**Dr. James R. Clynch, 2002**

## **I. Overview**

### **A. Definition of a Geographic Information System**

A Geographic Information System is a software program combined with a database that allows the manipulation, analysis and printing/plotting of data with geospatial locations. It grew out of efforts to use computers in map-making, but has grown beyond that application. The database is an important part of the system.

Geographic Information Systems (GIS's) are used by city planners, land management agencies, forest fighters and many others. Many other software systems are effectively GIS's or GIS like that are specialized for one application. They all have two things in common, a data base whose entries are geospatially located and some software to access and use that database.

"Geospatially located" means that explicitly or implicitly the geographic location of each item is known - usually by latitude and longitude. The location may be stored with each item, or the corner coordinates of a regular grid stored along with the step size between grid items. This allows the data to be retrieved, manipulated, and displayed using location as a key element.

This has proved very useful in many applications. Usually this involves bring together different sets of data and "fusing" them for comparison and analysis. This is precisely what is done in map making; the topography, political boundaries, roads, bridges etc. are brought together and printed (displayed). GIS's grew out of early efforts to use computers to aid in map making.

One formally, one can say that a Geographic Information System consists of two elements.

1. A database that has geospatial tags on the entries. This data base normally a relational database in computer science terminology. That is the database can be queried or integrated about the relationship of items it contains.

2. Software that allows the data base to be manipulated and the data displayed. Normally this program performs functions in three categories:

Data Storage and Retrieval

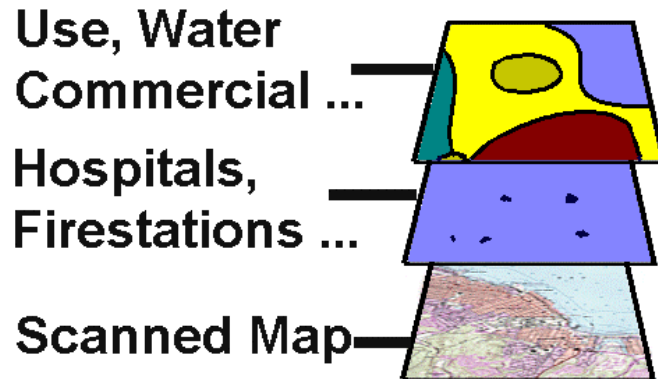
Data Analysis

Display

Many purists emphasize the analysis aspect of a GIS. However we will take a broader view and discuss systems that may lean more heavily on one of the other two function types. The key is having a database that is geospatially tagged.

## B. Layers

In map-making individual kinds of data were sometimes recorded on different sheets. Political boundaries, physical boundaries (rivers, lakes, oceans), and different kinds of infrastructure such as roads were drawn on separate master map sheets. These were then printed one after another on the paper to make the map. Each of these sheets was called a layer. This concept has carried over in the computer databases of GIS's.



## Basic Layer Concept

The above figure shows a schematic of some layers that might be put together for a specialized map, or to do some analysis. There are different types of information in the different layers. And the layers are of different data storage types. The scanned map is just an image - stored in any of several formats. The other layers are unique to GIS systems and have more complicated data inside them.

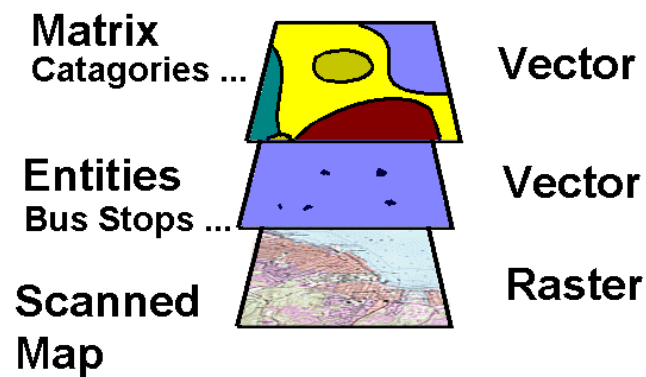
## C. GIS Layer Types

There are two basic types of layers used in GIS. Generically they are called Raster and Vector. Different software companies sometimes use different names, such as the terms Image and Feature used by the makers of the ArcView/ArcInfo product line. This company also calls layers "Themes". (To complicate matters, there are 4 types of data, and one of those is also called Vector. More on this later.)

A GIS data set can contain many layers. And within each layer there can be different data items. There are 4 generic types of items, and each can only go into one of the layer types.

<b>Data Type</b>	<b>Layer Type</b>
Vector	Vector
Matrix	Vector
Text	Vector
Raster / Image	Raster

The image data, which we frequently encounter as GIF's and JPEG's, can only be placed in a raster layer. The other data types are placed inside a vector layer. Normally there is only one image in a raster layer. If more images are present, they are usually each in their own layers. It is common for vector layers to have more than one data item in it. Sometimes this is done because the two items are very closely linked. (One may contain extra information about items in a second for example). Sometimes multiple items are put in a vector layer just because the data base designer wanted to do it that way.



## Layers in a GIS

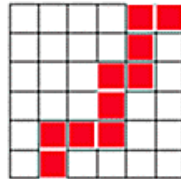
### II. GIS Data Types

In the real world most parameters are continuous in space. The air pressure, heights, etc exist at each point in the real world. These will have to be made into discrete items for computer use. In a GIS the resulting computer data is placed into one of 4 categories or data types. One type is the familiar image (called Raster in GIS), picture on the web, scanned image or digital camera output. Another type, the Vector type, is new to GIS.

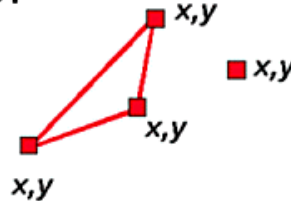
"Vector" here is a label for a specific type of data storage schema within GIS. In the vector format, the individual units are called Entities. These can be complex items containing other items.

The four data types have evolved with the development of GIS. The most primitive is the text type. This is just ordinary ASCII text material. The raster type is familiar as images that show up on the web. The other two types are newer or more specialized.

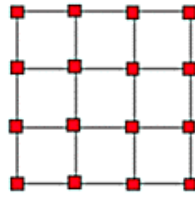
- Raster  
(Images)



- Vector  
(New)



- Matrix  
(Computer  
Matrix like)



- Text

Naval Postgraduate School  
36 35 42.2 N 121 52 28.6 W

## Four Basic GIS Storage Types

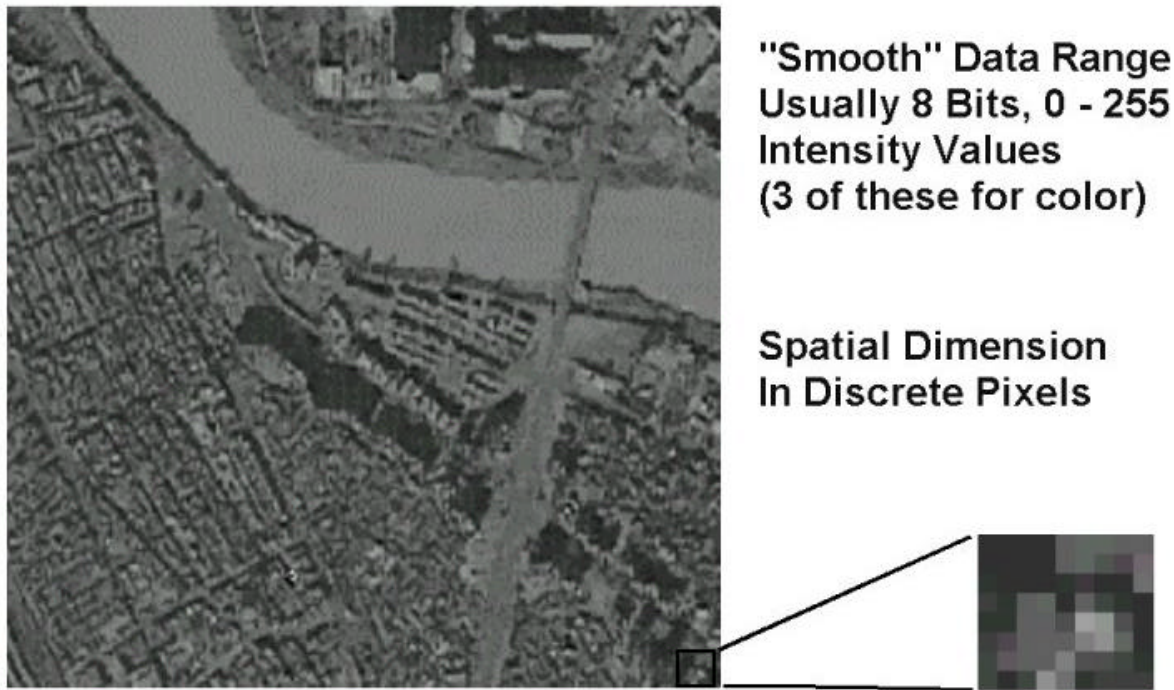
### A. Raster Data

Raster data can be stored in disk files in any of the standard image formats. When it is opened in a program, these are expanded to form a regular grid of cells, call pixels. These are arranged in rows and columns. The image we see is made of these individual pixel elements.

Each pixel has intensity values and nothing else. If the picture is black and white (gray scale) then there is one number per pixel. If it is a color image there are three values. Usually these values are 8 bits long and cover the range 0 to 255. Some image types can have fewer bits per pixel and a corresponding smaller range. These values for images are called radiometric values. In GIS raster images can also be used to store data that is not really an image.

It is important to note that there is no connection between the data in one cell and the adjacent cells in the computer storage. It is only when people see the image that lines and areas are perceived. The viewer

performs the integrating and interpreting function. While there are some "automatic" classification programs that work on image data, these are complex, usually specialized, and work, at best, only most of the time.



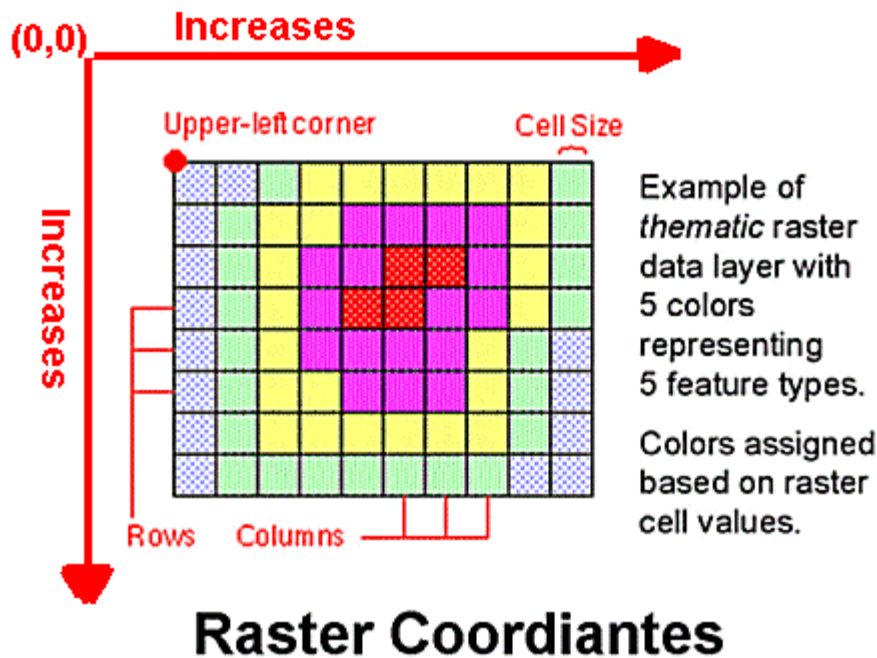
## **Images - Raster Example**

### **Continuous Data Made Into Pixels**

This example of an aerial photograph clearly shows a town with a river and a bridge. However these are evident only after the human mind interprets the data. If you look at the image in detail, it is clear that the image is made up of little blocks, each with a constant value, shade, color or intensity.

This particular photograph has been processed to make the spatial dimensions of each pixel the same in the vertical and horizontal directions. It has also had any tilt in the original taken out as well as some subtle effects due to differences in height of different objects. This process is called orthorectification. It is common for remote sensed images to be orthorectified before they are distributed. (Technically, this is a good approximation to a stereographic projection - but that is a subject for another module.)

It is possible, and common in GIS's, to classify areas into categories. For example in looking at an image of the arctic one could classify areas into open water, ice and water, ice areas, land, etc. Then a raster layer could be generated with a value for each category in each pixel. These might be given colors to assist the viewer in interpreting these data. (All these processes can be done with GIS tools for analyzing data.)

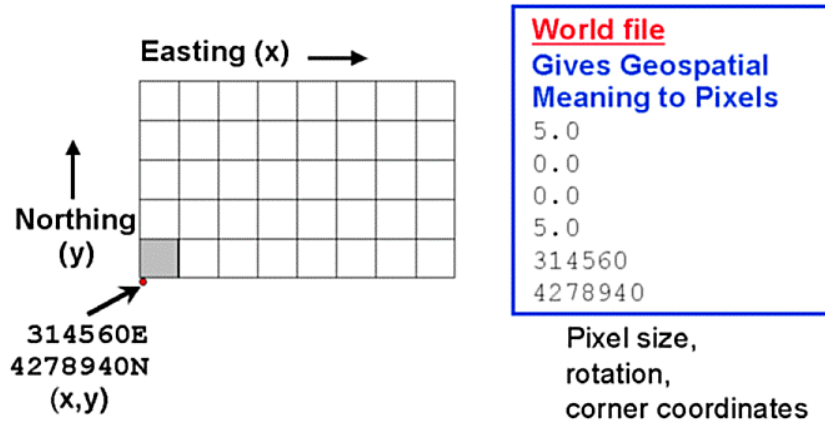


The above is an example of this kind of layer or images. It is often called a Thematic Map or Thematic Image. Because there are only a few types of data or data classes, the image has only a few colors. The color map can be assigned (and changed) with the GIS tools used to make the categories. These could have been made from a raster image or one of the other data types.

This example also shows the orientation of the primitive coordinates for the raster image. The origin is at the upper left. Sometimes these coordinates are used to inquire about pixel values or manipulate the data. Notice that these are just index values. They do not tell anything about ground location, or even the relative location of different cells on a quantitative basis.

In order to assign locations to cells, you need to know the dimension of each cell side in distance units (meters, kilometers etc.), the orientation (which way is north), and the geographic location of at least one cell. This is normally done for raster data with a separate file called a World File.

A world file is an example of Metadata. Metadata is "data about data". It helps the GIS software interpret and use some other data item. Notice that the world file is text. Most metadata is text. Metadata is extremely important in a GIS. On a paper map there is a lot of very important information on the legends. This information exists in a GIS as metadata.



## World File

### Assigns Coordinates to Pixels

### Assumes Map Flat - Has Projection Errors

## B. Matrix Data

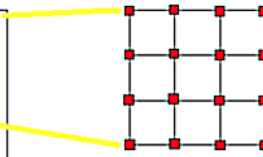
A type of data very similar to Raster is Matrix Data. Matrix data has values at evenly spaced interval in two dimensions. The rows and columns represent spatial locations, but may be spaced in distance (meters) or degrees or other units.

A key difference from raster data is that each location (grid point) may have any number of places for values. While there is one value for a black and white raster images, and three for color, in a matrix there could be 1, or 5 or whatever. This is literally a multidimensional matrix from some computer language such as C, C++, Matlab or Fortran. The values at each location, or in each cell, are often called "tuples" as in 3-tuple or 5-tuple for 3 and 5 components respectively.

A second major difference is that matrix data is stored in a vector layer. It is called by different names by different software. The name grided data is common.

Most GIS systems allow a grid to be converted into a raster image. In this case a new layer is generated. This layer will be a Raster type layer.

15	18	23	32	29	34	38
17	21	27	31	28	33	44
20	25	26	29	27	30	41
20	24	27	33	32	36	40
49	44	63	82	70	62	0
88	54	58	71	0	0	0



**Matrix Data**  
**Usually Values at Point**  
**Multiple Values / Cell Possible - "Tuples"**

### C. Vector Data

The key data type for a geographic information system is Vector data. This data type originated with GIS systems. It has many properties that give the GIS system a lot of its power. Of course, vector data can only be placed in a vector layer.

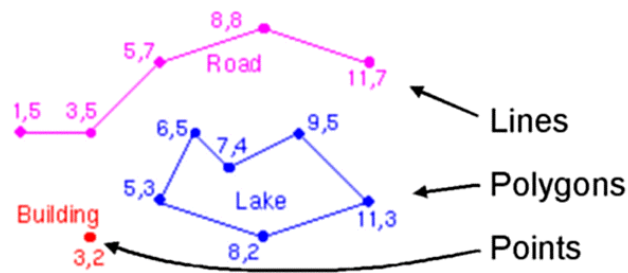
The basic properties of vector data are:

1. Primitive Entity types
2. Topology Level
3. Attribute Tables (also called Feature Data).

It is the presence of attribute tables, which attach extra information to entities, which makes vector data sets so powerful.

The lowest level items in a vector data set are the primitive entities. These are geometric quantities, points, lines or polygons (areas). (Some books differentiate areas from lines and points, but that will not be done here.)





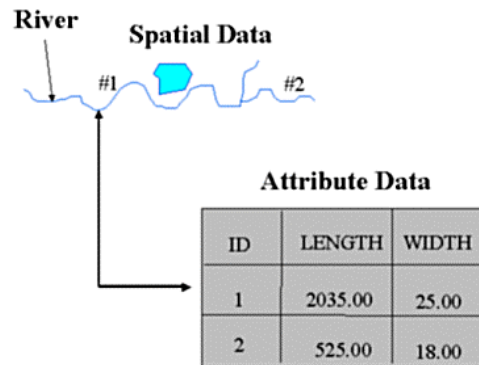
## Vector Layer Primitives

The pixels at a point or along a line are not stored. The actual image of a line will exist only when the display engine of the GIS software renders the vector data. The stored data of the primitive is just the location of the points or the nodes - points where the edge changes direction - of the lines and polygon boundaries.

This storage method means that the storage requirements are usually a lot smaller than for a raster layer. In addition, the scale of the images can be zoomed in and out and the rendered images do not get blocky. (At extreme zoom, the discrete nodes may show up though.) The entire inside of a polygon can be colored some solid color, or covered with a pattern with only a few bytes of storage. The data stored tells the display engine what to do; the data does not contain the end results.

The spatial relationship between entities in a vector layer can be complex. Two polygons may share a common boundary, and that boundary may exist as a separate line entity. These relationships can be pre-computed and the results stored. These relationships are called the topology of the data. There are 4 levels of complexity that the topology can have. At the lowest level, often called spaghetti topology, there is no spatial relation data stored. This is called level 0. At the highest level, level 3 or full topology, relationships between points, lines, edges and faces are all included. There are two intermediate levels. The major purpose of these extra topology data items is to assist in the spatial analysis of the data - to allow more complex questions to be asked of the GIS about the data. Most high-end programs contain level-3 topology.

Notice that we have not said what these points, lines and polygon represent. That is the task of the Attribute Tables. These are structured tables of data that contain information about the items in the vector data set. They might indicate that the first entry is a point, represents a building, that it is a school, and that a detailed diagram is contained in some other file. For an airport it might contain information on the runway, their length, direction, and surface. Vector data is sometime called "Smart Data". It is the power of attributes and of the GIS to analyze attributes that gave rise to this label.



## ATTRIBUTES

### Additional Information Tied To Vector Entity

The attributes tables can be displayed and edited in a GIS. They can also be processed by powerful analysis commands. The table for an airport might look like the following. Here the data describing the Washington Dulles airport is shown.

1. Trans Airport Points(vmt) - SEOULM	nam	UKN
2. Trans Road Lines (vmt) - UKN	nam description	NO entry present
	rst	1
	rst description	Hard/Paved
	rtt	13
	rtt description	Primary Route
	use description	0
	use	Unknown
	wtc	1
	wtc description	All Weather

Attribute Table

**Attribute Table**  
**Contains Information About Features**  
**Can Be Used In Relational Functions**

With smart data, one can ask a GIS to extract all points that meet some criteria and make a new vector data set. This might be something like "extract all points within 2 kilometers of both a currently active railroad and an airport runway of 12000 feet or more". This would be useful for a logistics planner. An urban planner might ask about points in a watershed with large slopes little vegetation and less than 50 cubic feet per second of sewer capacity. (Good to know in a hard rain to predict floods.)

#### **D. Text Data**

The last data type is standard text. In the early days of GIS development a lot of data existed as text files. Each line represented one item and could contain several details. These still exists and can be imported into GIS's. In many cases where data is written by computer programs to be imported into a GIS it is written to a text file. Then some commands are used to bring the data into the GIS as a grid or table.

As noted in the discussion of raster data, metadata is usually text data. It is often, although not always, in a separate file from the data that it describes. One example of text data is to include data about an airfield such as runway length, tower frequencies etc. There are many other common uses.

### **III. Continuous vs. Discrete Data**

Now that the four types have been briefly described, a short note on converting the real world continuous data to discrete computer data is in order. For the raster format, this is usually simple. A picture is taken digitally or scanned or some matrix is converted to a raster. However the power of a GIS is in the vector data sets.

There are four generic ways that continuous data, often called a field of data, is converted to vector entities.

1. Gridding
2. Making into a TIN
3. Contours
4. Profiles

This list has the most common method on top and the least used on the bottom.

The generation of a grid is the most common. This is just the process of creating a set of matrix data. In an intermediate output file, it might be in text format and become "matrix" format inside the GIS. The most common field used in GIS is height. There are many sources of height data, which is often called a Digital Elevation Model (DEM). Agencies and firms as grids or matrix type data almost always distribute these.

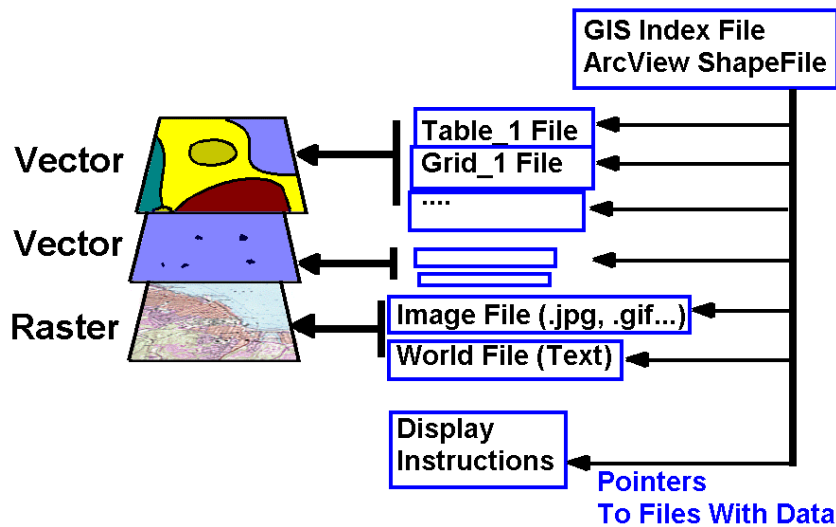
When doing analysis on height data, the early GIS developers took over a method used by classical surveyors. They often had data only at key points in the field. The measurements were recorded at high points, low points and critical points of slope change. This was modeled by hand as a series of surfaces made up by connecting the points. Each surface was flat, but could be sloped. Therefore it had a center elevation and a slope vector (direction and magnitude). The simplest figure was used, which is a triangle. This then became as

a Triangular Irregular Network (TIN). When GIS's were developed TINs were implemented. Specific GIS commands or subroutines were developed to deal with TINs. If these are not a built in commands for a particular GIS, it is likely that several people that include these. These commands will convert matrix to TINs and back, and compare, analyze and display TINs.

The two other forms are less common, but special extensions are also available for them. A contour is a line of constant value (formally called a isoline). These lines are generated from digitized topographic maps or DEMs or other matrix data. They are tagged with a special attribute to indicate they are contours. The fourth special entity used is a profile. This may be along a straight line or not (along a road for example). The common thread about all these types is that they are methods of representing a continuous value that have special processing routines available inside GISs.

#### IV. Data Storage in a GIS

If you make a data set for a project with a Geographic Information System it might contain several layers, and the vector layers might contain several data sets of different kinds. Yet if you look at the file that is "the data set", it usually will be quite small. This is because this file does not contain the actual data, just pointers to the files that have the data.



The above diagram shows how one particular family of GIS product, the ESRI ArcView program, does this. The index file is called a "shape file" in their terminology. What is stored in what kind of file is called the data model for the GIS program. The data files might not be the original input files, but reformatted and re-organized files. In some cases, images for example, it might be the original data. In many cases the data is also split and reorganized into file structure that the particular program uses.

This means that data from one program usually will not fit into a program from another manufacturer. However most of the major software makers include or sell "plugins" or "extensions" that load other data formats. There are many users groups for the most popular products with user-supplied extensions. There is even a military users group for the larger programs. These have extensions to load most the Defense Department data formats.

## **V. GIS Analysis Functions**

As mentioned in the introduction, the functions that can be performed in a GIS system are often divided into three groups:

Data Storage and Retrieval,  
Data Analysis, and  
Display.

Historically the data analysis functions have been viewed as the heart of a GIS system. However all three are really important. There are cases where a GIS is employed just to obtain the data base storage and retrieval functions and many cases where the display functions are most important. This is especially true with the "lite" GIS systems now available from several vendors for the PC platform.

Here a list of some of the functions and function types in each category will be given. There can be some debate about what is the proper location of some functions, but the intent is just to give an overview of the types of functions. Also some functions may be missing or expanded in different GIS systems.

### **A. Data Storage and Retrieval Functions**

These functions include the generic retrieval and writing of data to the database. The input (and sometime output) can be combined with other functions. For example with large data bases the input may be done only on records that meet some criteria. This could be geographic area, or some attribute, possibly in some combination or logical function. For example one could ask for all the records in the state of California that are north of 35 N. The state is probably present as an attribute while the location is part of the geospatial tags.

An important data storage function is data base maintenance. This includes the data validation and quality checking that must be done with any database to keep it useful. Sometimes this is done with a program or script that checks consistency of parameters. Often a human is in the loop. Manual editing is also possible.

Finally, the data may be converted at or after entry. While this might be considered an analysis function, it will be listed here. A grid (matrix data) might be converted to a raster image in a raster layer for example. Or scanned images of a map might be re-projected into another projection or another datum. (The raster data re-projection functions are usually present only on high-end GIS systems.)

### **B. Data Analysis**

These functions are quite varied. There are many common threads present in most GIS's though. The base or primitive functions usually involve the geospatial nature of the data tags. These usually include:

Measurement Functions - distance, length, area, volume,  
Connectivity or Topology Functions - connected, overlap, crosses, adjoins, etc.  
Neighborhood or Proximity

The measurement functions are fairly simple. What is the distance (shortest distance) between A and B (such as in Mapquest)? Or what is the length of a given path/route between A and B. Also what is the area of some feature or combination of features? The topology functions are often used to qualify some selection of points/areas/features to make a new data set. For example finding all the points/features within some other item (California for example) that meets other criteria. Clumping or clustering is also in this category. Find all the points that touch or are adjacent to California that are above sea level are an example. Neighborhood allows a "within x distance of" type criteria to be added. This is also useful in clumping.

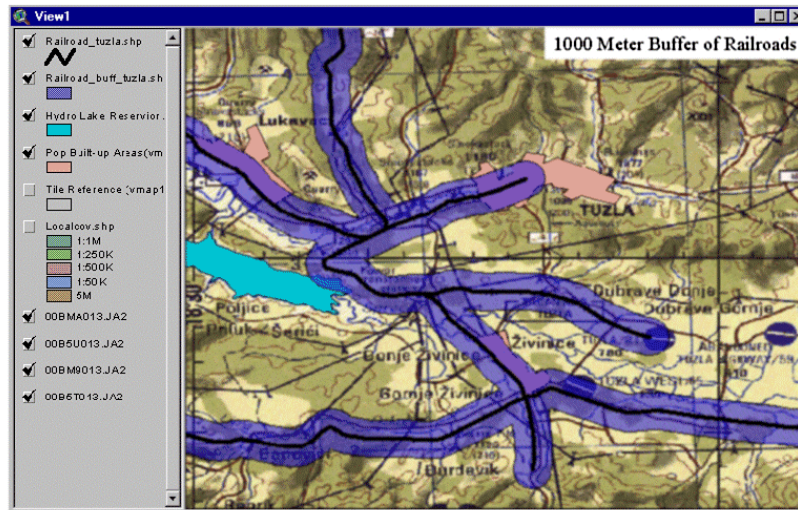
Some common general functions for spatial analysis are:

Clumping	joining small pieces (sometimes not contiguous)
Buffering	Defining an area within x-distance of some criteria
Registration	Alignment of data set / layers. Includes techniques for Edge Matching Identification of Mis-alignment – sliver areas Inter-Layer alignment – Conflation
Network Analysis	Finding (best) Routes – with or without cost function Inter-visibility areas
Classification	Areas by ranges of some value or attribute
Attribute Analysis	Usually Logical functions AND, OR, XOR, Greater_Than, Less_Than, Equal
Continuous Surface Analysis	Slope (derivatives) Generate Distributions Decisions based on distributions Statistics – decisions based on statistics

It is also possible to have scripts or programs that perform computations on values. These user-defined scripts are usually included inside the GIS as plugins. This makes the function available at the click of a button. The older scripting languages were specific to each GIS program. Newer GIS's have moved to Visual Basic as a scripting language. There are also user groups and web sites that have many scripts available. Many were prepared by large government agencies such as the US Geological Survey or state resources agencies. There are specialized defense users groups for the more common GIS's.

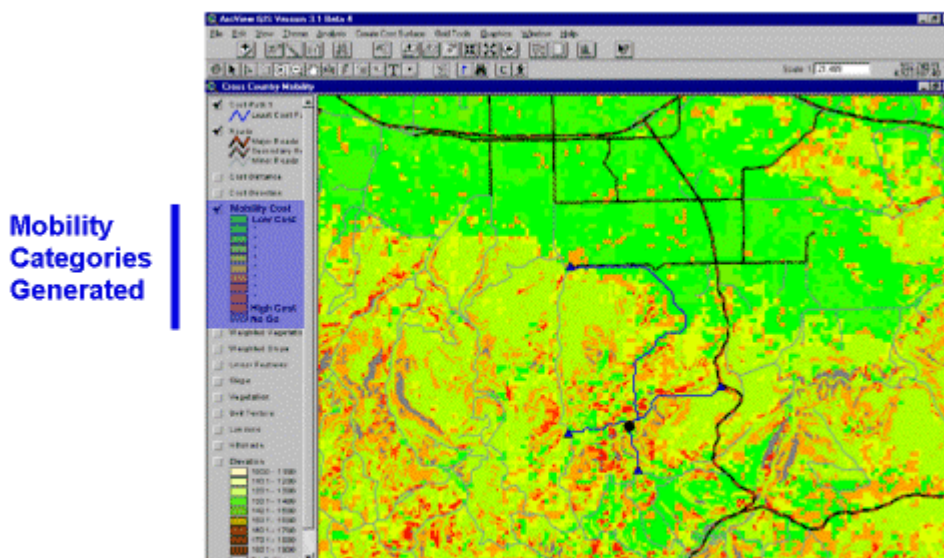
The output of an analysis is usually another data set within a vector layer or a new layer. A new layer is the most general for future operations, especially in research. For predefined operations in a production environment keeping data in one layer may be useful for storage.

Some examples may make things clearer. Below is a case where a Buffer has been created 1000 meters about all the railroads. This polygon (area) feature likely would be put in the same layer as the railroads, or a new layer containing both this feature and the railroads created. This could be useful in defining a right of way, areas to be considered for location of a heavy industrial factory etc.



## Proximity Analysis (Buffers)

One common analysis tool generates a new variable and layer by classifying or categorizing a variable into bins. In the example below the difficulty in moving heavy equipment is computed by a script using digital elevation and other data. A new layer is generated and then is categorized according to some predefined bins. The bins color is assigned color codes. This kind of layer is called a Thematic Layer. The computation of the mobility function could be done with the general built in GIS commands or a script.

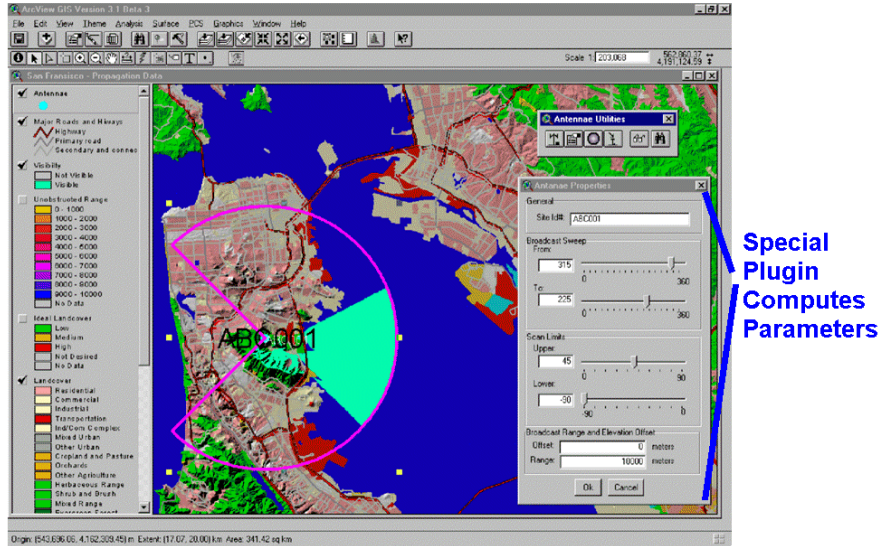


## Analysis Computed Mobility Cost Categorized, Made Thematic Layer

As an example of a more complex computation, a radar visibility and detection problem is shown. A special script was used to input parameters from the user. These included the site location (in this case given



just as a site identification number – GIS looks up the coordinates) and some basic radar parameters. The parameters of the radar might also be pre-stored in a vector entity. The topography is found from a digital elevation model. The output is an entity that shows the area that could be seen by the radar and the area that is seen when it is pointing at a specific azimuth.



# Radar Visibility Computation Script Plugin Example

Clearly if the properties of air defense systems were pre-stored, then areas of vulnerability for specific aircraft could be computed and displayed in a similar way. These “threat domes” are included in some defense plugins. This type of feature is included in many aircraft mission-planning tools for the military. These mission-planning tools are “GIS Like” applications.

## VI. GIS Display Functions

The analysis might be useful as is or stored. The most common case is for it to be displayed for a human operator. In essence a GIS is a generalized data fusion tool used by the operator to convert data into useful information for a human user.

The display functions define how an output sheet will be generated. It can include the functions:

- Ordering of Layers
- Layer Visibility flags
- Element layout on a sheet – positioning multiple ‘displays’ on one sheet
- Adding Titles and Text to Output Sheet
- Adding Meta-data

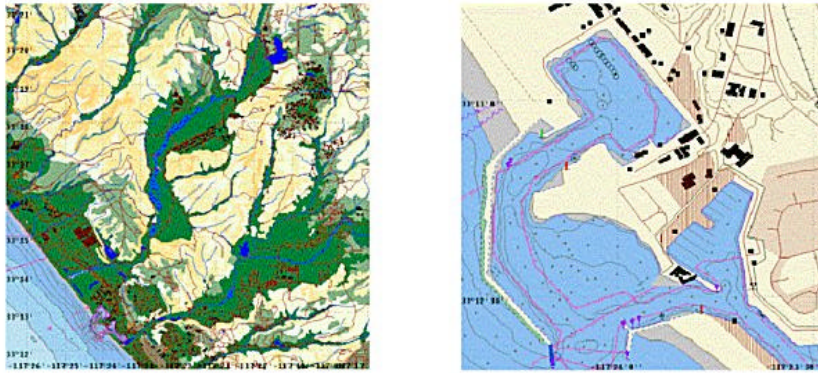


## Color Re-Coding Projections

Vector layers have no fixed projection within the database. The entities are tagged with location. At display time some projection is selected and performed on the fly for the data elements. For raster data, the projection is usually fixed. In high end GIS's a new layer can be generated in a new projection, but this is not usually done at output time. In a mixed raster / vector display care must be taken both to have the same projection but also that the layers are aligned.

Often a layout arrangement is saved in a special file. Then in a production mode this can be called up and used with little operator effort.

One common function of the display engine is "smart display". When dealing with vector entities the actual display data is generated on the fly. Therefore a road could show up at different zoom levels. However when zoomed out, there might be too many items to plot for a readable display.



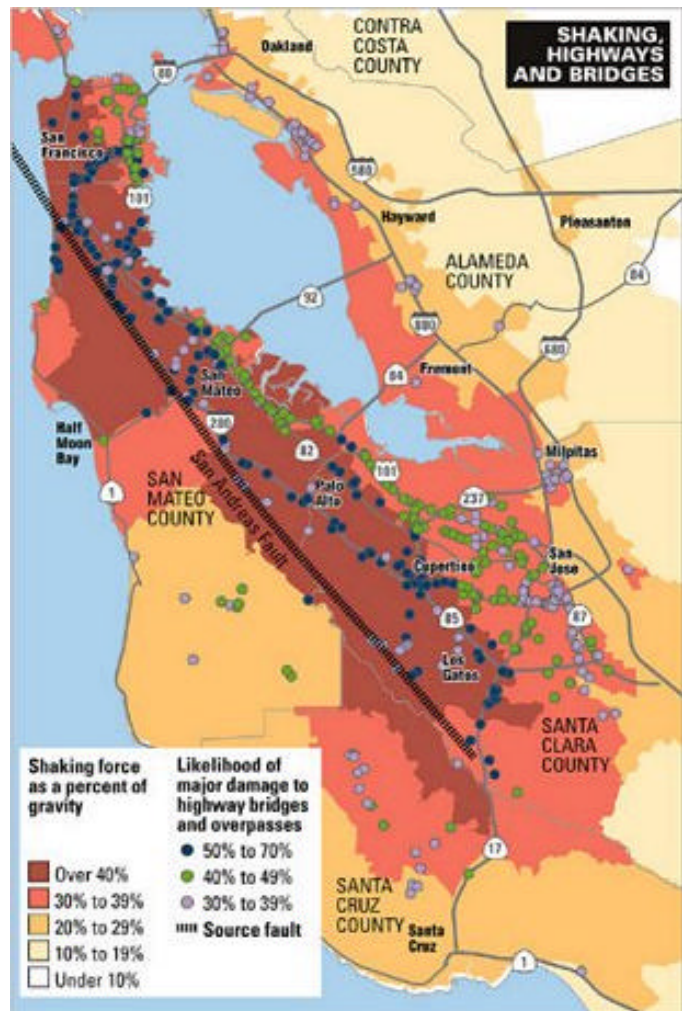
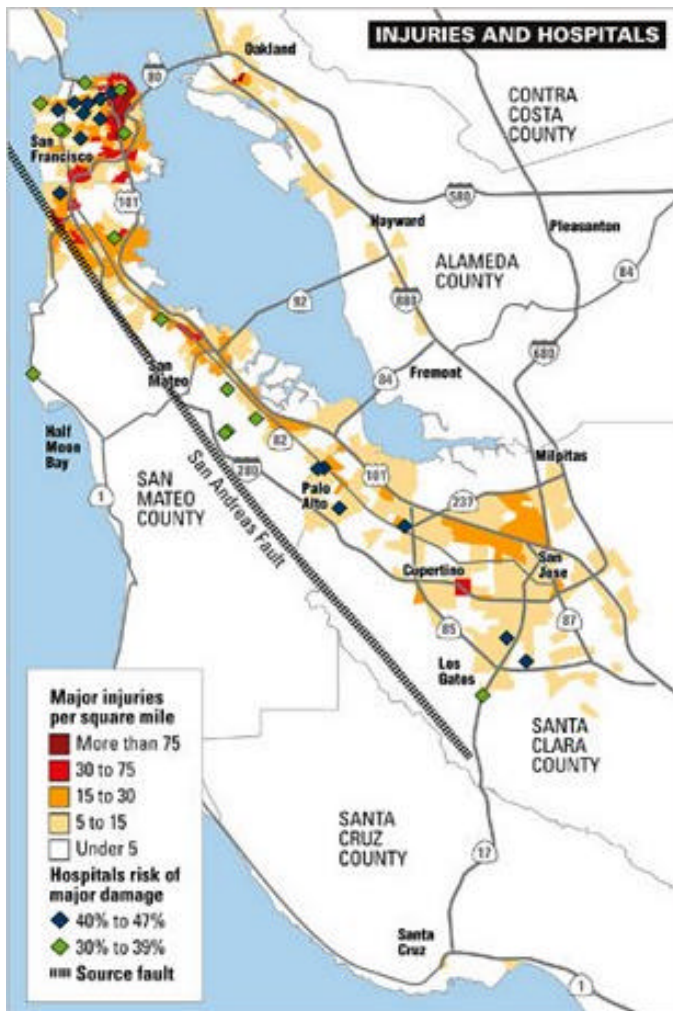
## Vector - Smart Data Zoom

Here the same area is shown at two extremes of zoom. All the data from the right panel is available to both images, but the display engine has omitted detail on the large area map. Some coding of the data might be needed to define importance of data. Common web map servers work this way. More roads show up on a small area like downtown San Francisco than on a map covering the US. (These map servers are GIS programs that take data from the web request, generate the map, convert it to a raster image, and imbed it into an outgoing web page.)

## VII. An Example - Preparing for the Next Big One

In the San Francisco Bay area it is not a question of if there will be another major earthquake, but when. To plan for this scientist make models of the fault lines. They run scenario's of various kinds of fault motions and predict the ground motion. This depends on the distance from the quake, its magnitude and the kind of soil in a region. Hard rock moves less, loose soil more, and fill in marshy areas can turn to a fluid. Take the output

of this engineering computation and put it in a GIS. Then import data from the US census bureau on population density, urban data on build types and a GIS can compute the likely numbers of injury and deaths.



Hospitals might also be added to see how the capacity to handle the injuries compares to the load. The same engineering data could then be applied to the database of bridges to see if the people could get to the hospitals.

Notice that the science computation was done in a dedicated program. However the spatial analysis was all done directly inside a Geographic Information System. Once the computation was designed and the data acquired, the computation was quite rapid. Many of the data sources such as the census data are readily available on the web. Cities and counties often have infrastructure data in a GIS format or some format (Computer Aided Design - CAD for example) that can be directly imported into large GIS programs. This example shows the power of a Geographic Information System.