BPD
BENDRASIS PROGRAMAVIMO DOKUMENTAS

EUROPOS SĄJUNGA

SOCIALINĖS APSAUGOS IR DARBO MINISTERIJA

NACIONALINĖ ŽEMĖS TARNYBA
PRIE ŽEMĖS ŪKIO MINISTERIJOS

THE BASIS OF AN EU CO-FINANCED PROJECT "QUALIFICATION IMPROVEMENT OF THE LITHUANIAN GEOGRAPHICAL INFORMATION MANAGERS" (NO BPD2004-ESF-2.2.0.-02-05/0143)

Civil servants learning programme

DISTANCE LEARNING OF GEOGRAPHIC INFORMATION INFRASTRUCTURE

Training Material

ELEMENTS OF GEOGRAPHIC INFORMATION SYSTEM

GII-01

Vilnius, 2008

MALASPINA
UNIVERSITY-COLLEGE

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

HNIT - BALTIC

Training material „Elements of Geographic Information System“ (GII-01)

Authors

Module 1, 3 - Brad Maguire

Module 2 - Andrew Miller

Module 4, 5 – dr. Gennady Gienko

Reviewed and edited by

doc. dr. Gintautas Mozgeris (Lithuanian University of Agriculture)

Reviewed by

doc. dr. Aida Mačerinskienė (Vilnius Gediminas Technical University)

From English translated and edited by

Astraneta UAB

Study material for civil servants distance learning programme “GII Distance Learning”, which has been approved by a law “RE Approval of Study Programmes for Civil Servants” (No 27V-72; Apr 17, 2007) by the State Service Department under the Ministry of the Interior has been prepared on the basis of an EU co-financed project “Qualification Improvement of the Lithuanian Geographical Information Managers” (No BPD2004-ESF-2.2.0.-02-05/0143)

Beneficiary of the project National Land Service under the Ministry of Agriculture.

Contractor of the project HNIT-BALTIC UAB, Vilnius Gediminas Technical University

Table of Contents

Introduction

This course is an overview of geographic information systems (GIS) and Geoinformation Science. In this course, we look at the various technologies that make Geoinformation Science possible, including Ground-based Mapping, Global Positioning Systems, and Satellite and Aircraft-based Remote Sensing. We discuss the way these data are organized in GIS, and the various GIS packages that are available, including Commercial and Open Source options.

This course is intended for professionals in a number of different fields who would like an introduction to GIS. Although you may not be directly involved in working with a GIS, having an understanding of the way the GIS works can help you in your day-to-day activities. For example, you may be required to prepare data for entry into a GIS, or you may be analyzing the results that have been created by a GIS. In both cases, understanding how the GIS works will save you time and energy.

This course assumes that you have a basic understanding of the principles of cartography and mapping, and that you are reasonably computer-literate. By the end of this course, you will actually performed introductory GIS analysis, and you'll be well on your way to being "GIS literate."

This course consists of five parts:

1. Introduction to Geoinformation Science;

2. Introduction to Geographic Information Systems (GIS);

3. Geographic Data;

4. Methods of Data Analysis in GIS;

5. Introduction to remote sensing.

1 Introduction to Geoinformation Science

1.1 GIS and Paper Maps

1.1.1 Introduction

For centuries, the main way of displaying and transmitting geographical data was through paper maps. Maps have become part of our culture, permitting geographic literacy in the same way that books permit traditional literacy. Just as children cannot read when they are born, geographic literacy requires training in the basic use of maps. Thus, courses are taught in high school on how to locate one's position on a map, interpret map Symbology and determine elevation using contour lines and spot heights.

In many university-level writing courses, students are taught to use the latest word processing tools to enable them to harness their creative process to the highest degree. Why has this happened? Simply because word processors allow you to write more material in less time, and automated tools, such as spelling and grammar checkers, take much of the drudgery out of writing. In a similar fashion, all map production outside the realm of art is now performed using geographic information systems (GIS). GIS gives you the tools to create maps rapidly and inexpensively. With these tools, you can focus on issues of map design and composition, rather than with the mechanics of map creation.

This course will be, above all, an advanced lesson in geographic literacy. Today, you will begin to learn how to process spatial data using GIS.

1.1.2 Similarities between GIS and Paper Maps

Human beings have one of the most advanced visual systems found in nature. Not only do we have stereoscopic vision, but we also have excellent color vision, and an ability to perceive change that is unique in the animal world. Humans are primarily visual creatures, an ability that comes to us not just through our eyes, but through our brain, which is uniquely designed to process visual information.

Just as we are able to learn a great deal by examining a photograph, a well-designed map or graphical display can convey an enormous amount of information in a single glance. The job of the cartographer is to present spatial information in such a fashion that the user's uptake of information is maximized. This is done using the language of cartography. Cartographers use many different kinds of symbols to present spatial information in a way that is clear and concise. The language of cartography has been developed over centuries through a process of trial and error. We continue to use this language today in GIS.

1.1.3 Differences between GIS and Paper Maps

Geographic information systems have one critical difference from paper maps. On a paper map we have both spatial data and cartographic symbolization melded together into a single entity. The map *is* the database. GIS separates spatial data from cartographic symbolization and inserts the power of the computer between these (Figure 1). This simple separation of spatial data and cartographic symbolization is the key to power of GIS.

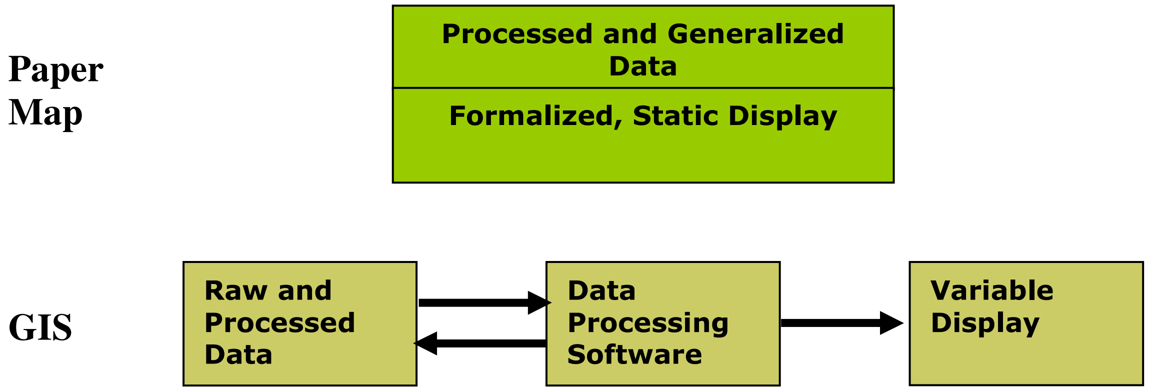


Figure 1. Fundamental differences between Paper Maps and GIS

Because spatial data is now on its own, it can be stored digitally on a computer hard disk, and copied at will. The effect of this is that the original data remains pristine, unaltered by manipulation, and can be retrieved at any time it is required. Contrast that with what happens to a paper map when a map user adds annotations in pencil. Since there is no degradation when data is copied it means that the data in your GIS is exactly the same as the data that was sent to you; copies of your data that you make for other users are as good as your own. As long as you have intact digital media, your data will remain in its original state until you erase or overwrite it.

If you use your GIS to alter your original data through analysis, you can store the data in a new file and leave the original data unaltered. For future analysis, you have the choice of either using the original, unaltered data, or the data that was just processed. One reason that you might want to alter your data is to prepare it for cartographic presentation. Original digital data is of little use cartographically. It must be processed, generalized, coloured, and symbolized to create a useful map. When you only have a paper map, all you get is this modified data. In the best case, the modifications have been slight in the data is relatively intact. Your job, as a user of a paper map, is to interpret the symbols on the map and turn them back into raw spatial data that you use for further analysis. So what you get out of a paper map is at best an interpretation of the original data, and at worst data that is wildly distorted. With the GIS, you always have access to the original digital data in its original form. Of course, if the digital data came from a paper map, then this data is based on a cartographic interpretation.

Spatial data that is stored on computer hard disk need not remain unaltered forever. Suppose, for example, that there was an error in the original data. The data could then be corrected, and any subsequent analysis or cartographic products could be re-created. Of course, in such a case, it would probably be wise to create an archival copy of the original data before updating takes place.

Another variation on this theme happens when spatial data is updated. In such a case, you would very likely want to replace the original data with the new updated data, once the original data have been archived.

Inserting a computer between stored spatial data and the display of that spatial data gives GIS many unique capabilities that are not available on paper maps. We don't necessarily need to produce a map of our spatial data; we could use our computer to produce a three-dimensional representation of the same data.

Are paper maps flat because it's an appropriate model for medium-scale maps, or are paper maps flat because of the limitations of the media? Put another way, if we could produce fully three-dimensional maps very inexpensively and easily, would we still use paper? We already produce special purpose three-dimensional maps for special purposes, such as education and architecture. When we want to look at the Earth as a whole, the globe is the most effective model. The reason that we don't use these more is simply an issue of cost. GIS enables us to easily produce dynamic, three-dimensional views of the Earth's surface, in addition to producing flat maps. This capability allows us to look at the world as we do in daily life; we can use our advanced visual system to obtain an accurate understanding of topography with a speed and depth of understanding that is impossible by interpreting contours from a flat map.

We could even use the power of the computer a little but more, and produce a rotating three-dimensional representation of the data. Furthermore, there's no reason that we couldn't take a data source from the Internet that is being updated in real-time, process that information using the GIS, and display the results in real-time as they are changing (Figure 2).

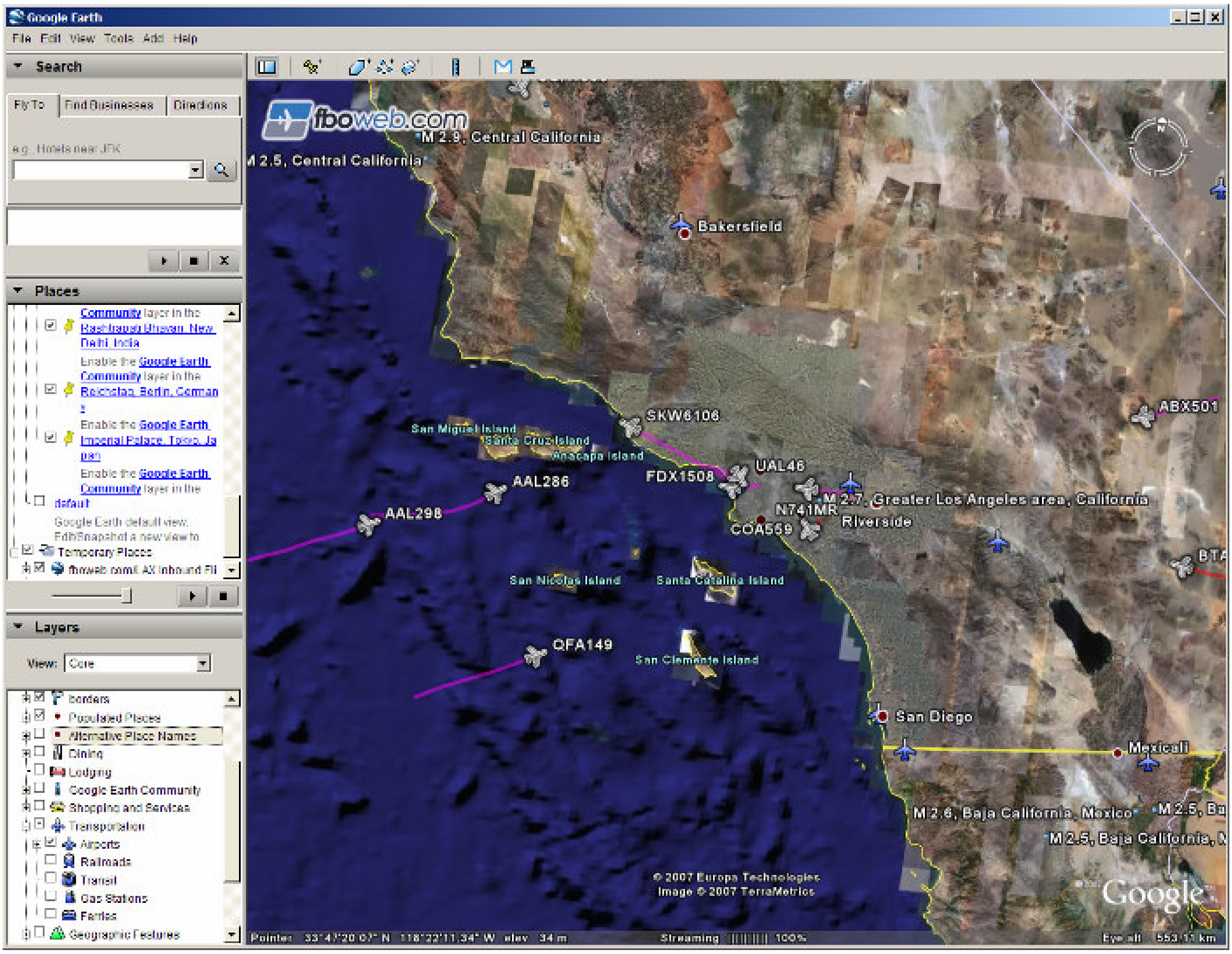


Figure 2. Real-time tracking of aircraft inbound to Los Angeles International Airport (courtesy of Google Earth, fboweb.com)

1.1.4 Map Elements

As we have mentioned already, cartography has developed through centuries of trial and error. Today's maps have a number of features, or map elements, that simplify the interpretation and understanding of the symbolized spatial information that is presented.

The title of the map is a short description of the contents of the map. In as few words as possible, it should express the main topic of the map. In an ideal situation, the meaning of your map is immediately obvious to all viewers, and the map does not require a title. Unfortunately, that is unlikely to happen in reality, particularly for thematic maps. Since a title might be the only thing that a uninitiated viewer of your map understands, it should be clear and free of jargon. New cartographers often have a tendency to use very large text for their map titles, but a properly sized and placed title should be visible, and should not dominate the map.

The map legend indicates the meaning of all symbology used on the map. As with the title, the map legend is supporting documentation that is provided for the user who does not understand the meaning of the map.

The border is a dark line separating the spatial data in the map from the other elements of the map on the page. The neatline is another dark line which outlines all of the map elements, and may be used as a guide for trimming the map.

The *graticule* is a grid of lines that helps a user to determine the coordinates of particular features on the map. Lines may represent latitude and longitude or X and Y in a projected coordinate system. Graticules are generally augmented by a series of ticks around the neatline of the map which show the values associated with the graticule lines.

A map's scale shows the ratio between one unit on the map and a number of units on the ground. For example, a scale of 1:50,000 indicates that one unit on the map is equivalent to 50,000 of the same units on the ground. Scale may also be represented using a scale bar, which shows the scale graphically, or verbally, for example "1 cm is equal to 10 km."

The north arrow indicates the direction of North on medium and large-scale maps. North arrows are commonly omitted on small scale maps such as maps of the world, since the direction of North is different in every place on the map. Generally, north arrows point in the direction of True North, but they may also indicate the location of Grid North (since the graticule lines are not always aligned with the map's neatline), or Magnetic North.

The source indicates where the original data for the map came from, and is an important tool for other cartographers as well as librarians and archivists, who may wish to access or examine the source data.

1.1.5 The GIS User Interface

The user interface for a Geographic Information System is immediately familiar to experienced users of maps, since it incorporates many map elements that are found on paper maps. For example, the GIS user interface shown in Figure 3 features a title, a legend, a border, and X-Y coordinates to indicate where the mouse pointer is on the map, which replaces the graticule. There is no neatline, north arrow, or data source immediately visible to the user, although they can be added to any map is produced by GIS. This is due in part to the fact that the GIS is a tool for manipulating spatial information, rather than a map.

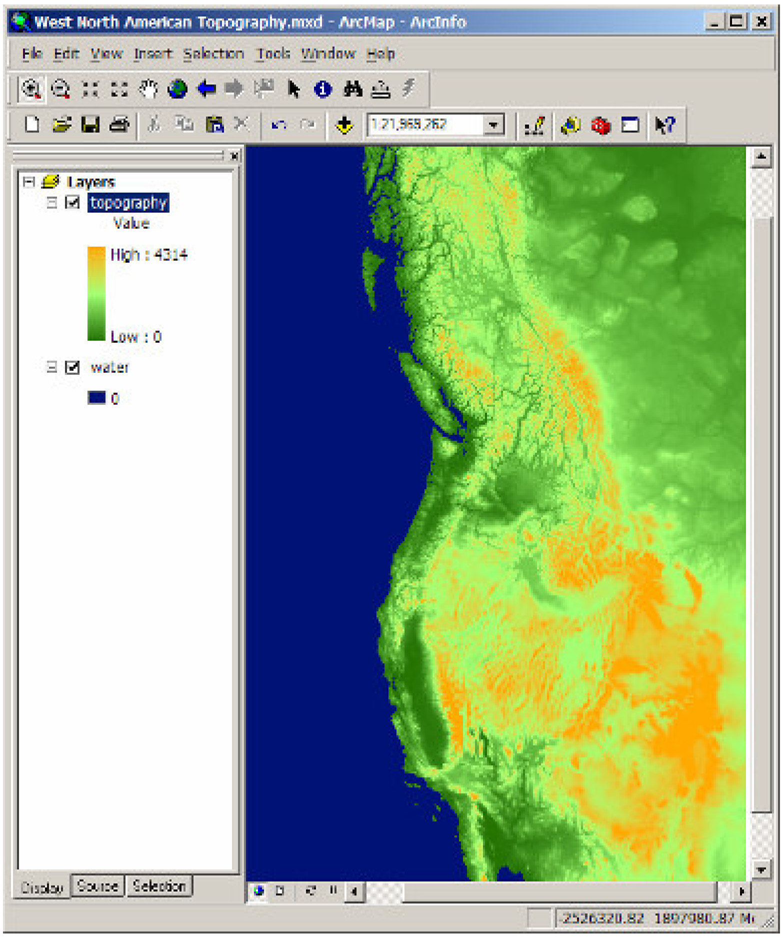


Figure 3. Basic GIS User Interface (ESRI ArcMap 9.1 shown)

The GIS user interface has a large number of buttons, tabs and pulldown menus. As daunting as all these tools are, they pale in comparison with the number of tools that are available, but are not shown. Three toolbars are shown in the figure; an additional 45 are available for functions such as spatial analysis, annotating maps, editing maps, viewing maps in three dimensions, and entering survey data.

1.1.6 Problems with Paper Maps

Geographic information systems were developed in the late 1960s and early 1970s in response to a crisis in the traditional mapping industry. Scientific advances, such as aerial photography, the creation of the first Remote Sensing satellites, and the introduction of computers allowed geographical data to be created at a much faster pace than ever before, making it difficult to process acquired data, and for cartographic production to keep up. In the late 1960s, for example, the Canadian government embarked on a project called the Canada Land Inventory (CLI). The CLI was to map the agricultural capability of all of Canada at a scale of 1:250,000, requiring the creation of over 1000 maps. A project of such magnitude was virtually impossible at the time, and it required the latest in cartographic technology that the 1960s could offer.

Burrough and McDonnell (1998) describe some of the problems with paper maps that the designers of the early GISs hoped to overcome. The problems they describe are:

1. The original data must be reduced in volume using classification and generalization

2. It is difficult to present information clearly

3. Data must be broken into map sheets
4. It is difficult to retrieve data from the map for analysis
5. A map is a static, qualitative document

One important point that Burrough and McDonnell miss is that paper is a very poor archival medium. Great lengths are taken by map librarians to ensure that correct humidity and light levels are maintained, and that maps are stored properly. One of the most important reasons for adequate map storage is to reduce the chance or consequences of fire. People who have used maps in the field are all-too-familiar with holes being worn in maps along fold lines, distortions in scale caused by humidity and raindrops, and in extreme cases, maps being turned into a mass of wet pulp during a rainstorm.

Obviously, GISs have some limitations of their own. The most obvious of these relate to issues of portability. Even the most capable of today's handheld computers are bulky and heavy when compared with a paper map. Only if large amounts of information are required in the field does a handheld computer equipped with GIS have a weight advantage over a paper map. On extended field surveys, battery life becomes an issue with handheld computers.

Even the most poorly printed map has enormous advantage over a GIS display. Computer displays are limited in their physical size and in the resolution that they can produce. Although it is possible to "zoom in" using a computer display, it is not possible to have a detailed look at the large area because of resolution limitations. Consider that tablet computers have a display size of roughly 22 cm x 28 cm, at a maximum resolution of 1680 x 1050 pixels. That produces a maximum of 60 pixels per centimetre, which is only 1/8 the resolution that can be produced by in inexpensive printer. Even more important is the fact that a paper map can be unrolled to be a metre or more in each dimension, and so can display very large amounts of information at very high resolution. This combination of display size and resolution gives paper maps a serious advantage over GIS in terms of "situational awareness." It is much easier to get a visual understanding a spatial patterns from a paper map than from a GIS.

One result of the long history of cartography is that paper maps, in particular topographic maps, have a standard set of symbols and colours. Although there are regional variations, it is possible to quickly and accurately read topographic maps that were produced in many different parts of the world without prior training. With GIS, it is up to the operator to choose appropriate cartographic symbolization; only points, lines, and the outlines of areal features are provided.

There are many data standards in use, in part because the technology of GIS is still evolving. Every new technical innovation in data storage requires a reconsideration of the way the GIS data is distributed. Thus, an important skill for GIS technicians to have is to be able to troubleshoot and work their way through issues of data formatting, data manipulation, in cartographic representation of that data.

In recent years, new technologies have been created that promise to bridge the gap between the capabilities of GIS and the portability and ease of use of paper maps. This new technology is called *Digital Paper*. The idea behind digital paper is having a computer display which consumes little or no power, can be read under bright sunlight, is flexible and waterproof, but which can be updated by computer when the need arises. Although the technology is still is in its infancy, colour Digital Paper is now available, and it seems likely that within a decade the advantages of GIS and paper maps will be combined. Imagine pulling a map from a backpack, seeing a red X in your current position (because it will be GPS enabled), then being able to zoom in your current location and then produce a three-dimensional image of the terrain that your hiking in, with your route shown from your starting point.

Figure 4. Digital Paper. Courtesy E-Ink Corporation

Because GIS splits the storage of data from the display of data and inserts the analytical power of the computer in between the two, it is able to overcome many of the limitations that have always been a part of paper maps. So now, not only can we store and display spatial data, but we can also use the computer's analytical abilities to determine the closest route between two points, to perform statistical analysis on spatial data, or to create models of animal habitat, for example.

As we have seen, there are currently two technologies for storing and displaying spatial data. Paper maps are traditional, and have developed through hundreds of years of trial and error to become a highly refined and effective tool. Geographic information systems are a much newer technology, which allow spatial data to be stored, analyzed, and displayed in ways never before possible.

1.2 History of GIS

GIS is a relatively new discipline, first envisioned in the late 1960s, but which has its roots going back thousands of years to the earliest known maps. If we look back in history, we can see a general trend towards increasing sophistication in the portrayal of spatial information, with very few setbacks has civilizations fell or were conquered. Part of the reason for this might have to do with the fact that cartography is strategically and militarily important, so cartographic innovations tend to be adopted rapidly and widely, which means that when a civilization falls, there is always another to continue and build on its cartographic traditions. Even after the fall of Rome, cartographic traditions were preserved by the Arabs, and then were readopted by the Europeans during the Renaissance.

1.2.1 Ancient Cartography

The earliest records of cartography are, not surprisingly, those that have been found on durable media. Materials such as stone, clay, or bone can last for thousands of years, preserving the cartographic record of ancient civilizations. It is quite likely that many more maps were made on less durable materials such as on animal skins, wood, textiles, or even paper. These of course, have been lost to our knowledge, because these materials do not survive very long unless they are carefully preserved.

The earliest maps are somewhat controversial, because at certain level, it becomes difficult to distinguish a "map" from a "drawing." Today, we recognize maps because of certain conventions such as legends, north arrows, and well organized and presented spatial data. However, we don't know what cartographic conventions were used 10,000 years ago, if indeed there were any.

For this reason, when searching for early maps, we look for something that resembles the spatial patterns that we expect to see today. There also needs to be some indication that the spatial pattern described was used as a "model" of the real world. For example, a 16,000-year-old painting in Lascaux Cave, France appears to be a drawing of the night sky. Certainly, we can recognize the stars in the painting, and they match spatial patterns that we see in the night sky. But is this a map or merely an attractive and detailed picture of the night sky? Today, we produce star charts which accurately give the locations of astronomical objects using a coordinate system, which is definitely a type of map, but we may never know whether the Lascaux painting is a map or not, because we do not know whether it was used as a guide to the real sky, or if it is merely a reflection of it.

Maps of terrestrial locations are somewhat easier to include in our list, since they rely on a birds-eye view of the location in question. Since it is more difficult to create an accurate spatial representation of something that you cannot see in its entirety, it is easier to be certain that a map of a terrestrial location was created for a reason. The oldest known example of this type is what appears to be a map carved on a mammoth tusk from Ukraine dating from 10,000 BC. This artefact appears to show a number of buildings organized in a line beside a river (http://www.atamanhotel.com/catalhoyuk/oldest-map.html).

The best documented, and most accepted ancient map is a 7.6 x 6.8 centimetre clay tablet dated from 2500 B.C. which was discovered at an archaeological site 320 kilometres north of Babylon at Ga-Sur, which is near the modern town of Kirkuk, Iraq.

Although maps made of clay, stone, and bone are ideal for archaeologists, they are heavy and difficult or impossible to carry. Over 3000 years ago, the first maps were produced using flexible media, which included papyrus, and vellum, which is made from animal skins. In this document we will simply refer to these as "paper." The earliest known paper map dates from 1300 B.C. and was created in Egypt. It now resides a museum in Turin, Italy, and is known as the Turin papyrus (Figure 5).

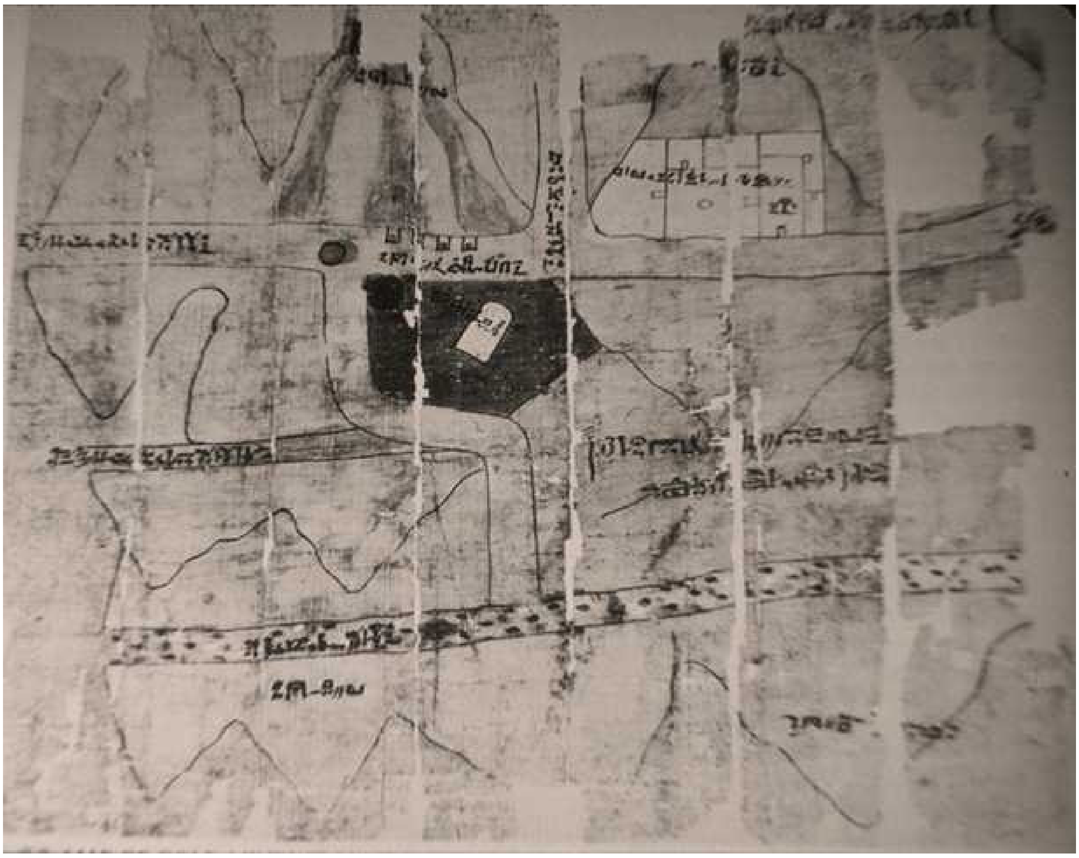


Figure 5. The Turin Papyrus, the oldest existing paper map. (Source: http://www.henry-davis.com/MAPS/Ancient Web Pages/102.html)

It is clear that the ancient Greeks had an advanced system of cartography; and we can see that the Greek conception of the world advanced considerably over the thousand years from Homer to Ptolemy. Eratosthenes, the ancient Greek geographer, who proved that the Earth was round and invented the concept of latitude and longitude. This change in understanding is reflected in Ptolemy's maps which were produced about 400 years later.

Until the Renaissance, Ptolemy's views on geography were considered to be inviolable, and so very little happened for the next 1400 years. Much if the learning in ancient times was preserved by the Arabs while Europe descended into the Dark Ages. One of the first examples that we see of the new thinking in the Renaissance was Waldseemüller's 1507 map, which takes the Ptolemaic map of the world and tacks on the discoveries in the New World.

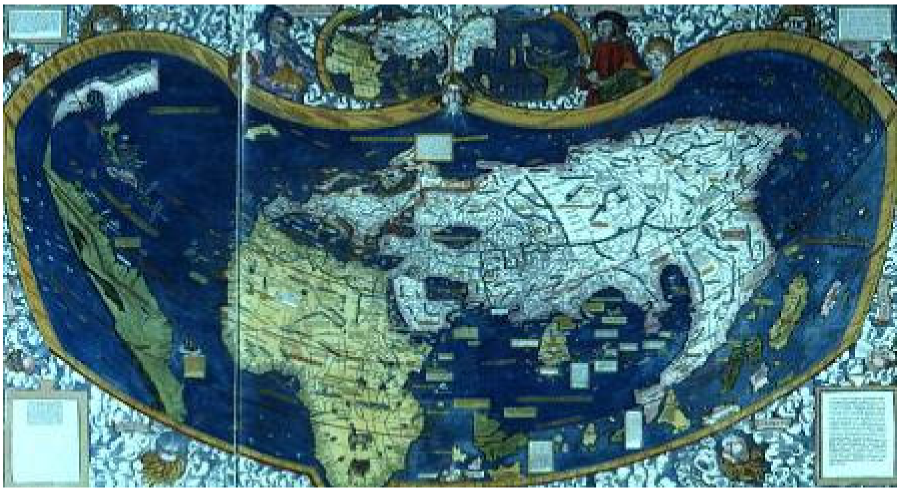


Figure 6. Waldseemüller's 1507 map, showing portions of the New World

1.2.2 Cartography during the Age of Exploration

Throughout the next 250 years, the Age of Exploration provided a constant supply of new material for cartographers. New lands were being discovered as a result of European voyages of discovery, and this kept the cartographers of the time busy. These new discoveries had strategic importance, so it was important that new maps be created as new discoveries became known.

During this time, we see a progression from a map as a work or art to a map as a work of science. We see a steady reduction in the amount of artwork along the map border, coupled with increasingly accurate depictions of the shapes of the continents and their locations. The invention of the Mercator projection by Gerardus Mercator in 1569 made it much easier to chart courses during voyages of exploration, since, on Mercator projections, all compass bearings are straight lines.

1.2.3 General Cartography

It was not until the 1690s that cartography became fully scientific. Under Louis XIV, France became one of the world's first centralized nations, and had a great need for maps of this new modern nation. Four generations of the Cassini family, starting with Jean Dominique Cassini in 1669, helped France become the world leader in cartography.

What the Cassinis did was to reform cartography by making extensive use of triangulation. This process allowed objects to be mapped accurately, by making use of successively smaller triangles. John Dominique's son, Jacques Cassini, completed a first-order triangulation of France, which allowed a national system of topographic maps to be created by another Cassini, Cesar François Cassini de Theury.

Cassini de Theury was able to complete nearly all of the 182 map sheets at a scale of 1:36,400 before his death. Cassini de Theury was also instrumental in tying the triangulations of France to those in Britain. By doing so, it was possible to determine the exact difference in longitude between the Meridien de Paris and Greenwich meridian.

Of course, to obtain the British involvement, Cassini de Theury required local help. Major General William Roy was Cassini de Theury's British counterpart. Roy quickly realized how far ahead of the British the French were, and he advocated the creation of a British topographic mapping organization.

In 1789, the French Revolution broke out, and by 1791, there was serious concern that the French Revolution would spread to England. King George III put the Board of Ordinance (the predecessor today's Ministry of Defence) on high alert and established the Ordnance Survey. The Ordinance Survey (OS) was charged with the mapping of all of southern England, in preparation for a French invasion.

It wasn't until 1801 that the OS produced the first one inch to one mile map. This map of Kent shows roads, buildings, towns, forested areas, and topography using hachuring (Figure 7).

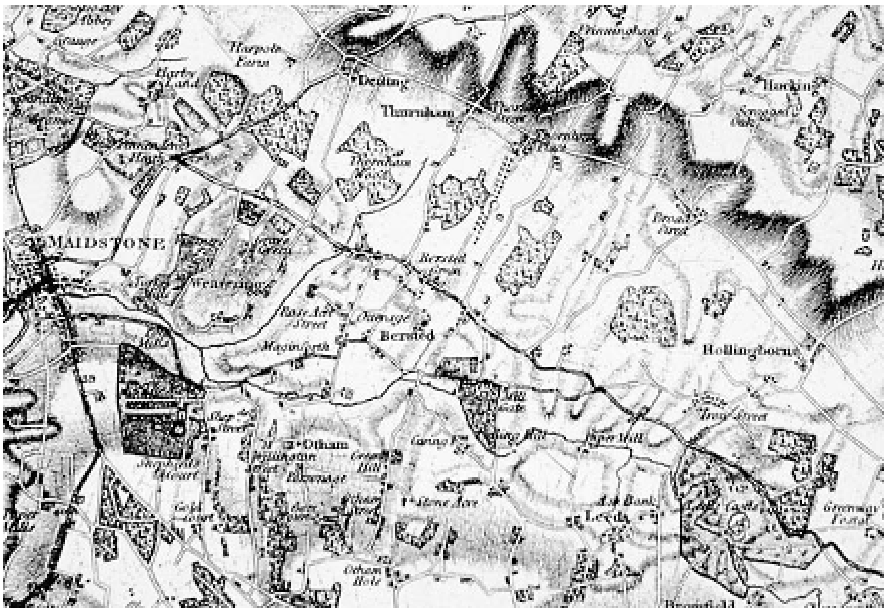


Figure 7. The first topographic map produced by the Ordinance Survey, a 1 inch to 1 mile map of Kent.

With the establishment of the Ordnance Survey, Britain eclipsed France to become the world leader in topographic mapping. British surveyors went on to map large parts of the world including Ireland, North America, and India. This organization became the model for topographic mapping organizations around the world. Soon, many developed nations have their own national mapping programs modeled on the Ordinance Survey.

1.2.4 Thematic Cartography

General maps, such as topographic maps, are designed to be used by everybody. As such, they are deliberately designed to be useful to a wide variety of people, but it is impossible for them to be useful for everyone. There is simply a limit to the amount of information that can be placed on the single piece of paper, and in urbanized areas, topographic maps reach their limit (Figure 8).

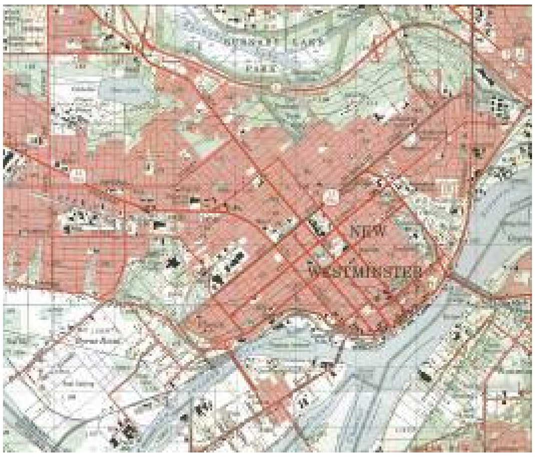


Figure 8. This topographic map of New Westminster, B.C., Canada illustrates the efforts to which cartographers go to clearly depict features in urbanized areas.

Some users require information that is unavailable on general maps. For example, Geologists are interested in surficial geology and the geological features of an area, which are unknown to most users. This information is not included in general maps, because first, it is very dense information and would overwhelm the map, and second, because it is of little concern to most map users.

For this reason, thematic maps have been created. Thematic maps are aimed at a specific group of users, for example geologists, entomologists, or public health officials who have a need for mapped information that is not shown on general maps. Thematic maps may use special symbology to convey their data that is not available on general maps.

One of the first thematic maps was produced by Edmond Halley, the famous English astronomer. In 1686, Halley produced the first is meteorological chart, using special symbols to display information such as barometric pressure and weather fronts on a map.

Thematic maps may display the results of scientific measurements. In 1698, Halley produced another map showing compass declinations across the Atlantic Ocean. Here is a map of something that was very real, but which had never been mapped. Alexander von Humboldt, the German geographer, used a similar technique in 1817 to produce a map showing isotherms in South America. Both of these are examples of how new and formerly unmapped information was beginning to come available as a result of science.

A thematic map combines elements of cartography with those of charts and graphs. In 1837, Henry Drury Harness, a British Army officer working for the Irish Railway Commission began to produce a series of maps showing numerical data. Town populations, for example were shown with circles of variable width, and traffic flows were shown using roads variable width. This work was eventually compiled together with topographic and geological maps into the "Atlas to Accompany the Second Report of the Irish Railway Commissioners."

In 1854, Dr. John Snow was able to use thematic maps as a detective tool in London. By drawing the locations of people who had died from cholera on one map, and the locations of wells on another, Snow was able to determine that a particular well was responsible for the cholera outbreak. This early example of taking to thematic maps and combining them into layers foreshadowed future developments in GIS.

Thematic maps show one or a few themes at most. It is not a long stretch to come up with the idea of taking a series of thematic maps, and stacking them one on top of the other to produce any type of map that we need. Thus, thematic maps were an important innovation that predated GIS and helped to make it possible.

1.2.5 Information Technology

Although simple overlays can facilitate some types of geographical analysis using paper maps (and much later, in the case of Ian McHarg, plastic transparencies), this approach begins to break down when more than a few maps need to be analyzed at a time. First of all, there is the weight and bulk of the paper maps, and secondly, there is a limitation even to the most transparent of plastics.

The solution to this problem is to separate the information from paper and store the information in a more easily accessible fashion. This requires computers. It is not a coincidence that GIS was not born until the computer revolution was producing rapid advances. It is also interesting to note that the concepts surrounding GIS coalesced within a few years of the first commercially available computers being marketed. This is a reflection of the difficulty involved in paper mapping, and the great benefit that early advocates saw in the new technology.

In this section, we will briefly discuss the history of information technology, beginning with the development of computers which made it all possible. Next, we will examine Information Storage and Retrieval Systems, from which GIS developed.

Since mathematics first began, humans had been the primary manipulator of numbers. While humans are surprisingly good at this, they are subject to some limitations, such as being slow, being subject to fatigue, making errors, and lacking numerical accuracy. These limitations have inspired inventors for centuries to use the best technology available to them to create devices to help humans perform mathematics better and faster.

Simple mechanical aids such as the *quipu* of the Incas, or the Chinese abacus have been available for centuries, but it was not until the invention of two years and dials the first calculator could be invented.

In 1900, a Greek sponge divers discovered an ancient astronomical computer off the Greek island of Antikythera. This device, which dates from 150 to 100 B.C., has been a source of great controversy ever since its discovery. One reason for this is because it displays brass gears and dials, which were not thought to have been invented until the 16th century. Recent research lends weight to the theory that this device is an ancient orrery, which calculates the position of the sun and moon, and possibly other planets for any date and time.

In the 18th century, European mechanical clock making became very sophisticated. Advanced gears such as the planetary gear plus improved metallurgy meant that it was increasingly possible to build highly precise mechanical devices. Leonardo da Vinci, coming before this time, had envisioned a mechanical calculating device, but lacked the ability to actually build one. It was not until 1623 that Schickard's mechanical calculator was able to add and subtract six digit numbers.

Gottfried Leibniz demonstrated a mechanical calculator that could determine a square root in 1694, and in 1821, Charles Babbage began designing the difference engine, which was essentially a modern computer that was to of been built using gears and levers for its internal mechanism. Babbage's efforts were abandoned, after a great deal of money had been spent, in part because of the difficulty of doing computing using mechanical technology. It wasn't until 1893 that William Burroughs produced the first successful mechanical desktop calculator.

The 1804 Jacquard Loom used in ingenious system of punch cards to store a set of instructions on the pattern that the loom was to produce. This is the earliest example known of a device that uses an encoded set of instructions. This device was to be the inspiration for a number of later computing devices.

Eighty years later, Herman Hollerith devised a punch card system based on the Jacquard Loom to aid with the tabulation of the 1890 U.S. Census. This system, which used 45 hole punch cards, could record 45 different demographic attributes about a particular person. For example, "married" might be represented by one hole, and "age 30-35" might be represented by another hole. Hollerith's system read cards using electrical signals, which brings us into the era of electromechanical computing In 1890, the U.S. has 63 million citizens; Hollerith was able to complete it one year ahead of schedule By 1900, Hollerith has introduced an automatic punch card feeder to streamline the process even further.

Just before World War II, electromechanical computers were beginning to show their potential. Konrad Zuse created the first programmable computer, the Z1, in 1936. Zuse later impressed the German authorities, and received funding to build the Z2 and Z3 computers. All of Zuse's computers used mechanical relays to do the central processing.

The Harvard Mark I computer was completed in 1944. Like the Zuse computers, these computers used electrical relays for their memory. The Mark I had the ability to multiply two 33 digit numbers every three seconds. The machine took up the entire floor of the building and consumed several tons of ice per day for cooling.

The German Enigma machine, which was used to create ciphers during World War II, created a great deal of interest in the use of computation to break difficult codes. One of the first computers of the next generation, the COLOSSUS, was built to break German Enigma codes. COLOSSUS was an example of the first electronic computers, which made use of vacuum tubes instead of relays. The vacuum tubes provided much higher computing rates, and avoided many of the mechanical problems associated with relay-based computers, such as bugs getting caught in the relays (these were the original computer bugs). Vacuum tubes, however, last less long than relays. This means, that when a vacuum tube burns out, your computer against to produce erroneous results. In addition, computers were frequently down for maintenance, and vacuum tubes were being replaced. The speed of the new electronic computers was demonstrated by the fact that the British were able to intercept and decipher Enigma Codes before the Germans themselves were able to.

One successor to COLOSSUS was ENIAC (Electronic Numerical Integrator and Calculator) which was built by the US Army at the University of Pennsylvania in 1946. ENIAC we used to calculate ballistic trajectories for artillery shells, and could multiply two 10-decimal numbers 300 times per second. Because of its vacuum tube construction (ENIAC had 18,000 of them), the computer had no internal moving parts.

IBM introduced its model 604 computer in 1948. Also still using vacuum tubes, this was the first computer that had boards that can be replaced in the field. The computer was wildly successful, laying a foundation for commercial sales of computers. IBM had planned to sell 75, but eventually sold 5600.

In 1951, the transistor was invented at Bell Labs. Because the transistor replaced vacuum tubes with a solid-state device, this allowed more reliable computers to be constructed. The first transistor-based computer was completed at Bell Labs in 1953 and contained 8000 transistors. The IBM 7090 was the first fully transistorized mainframe computer, introduced in 1958 Transistors also allowed the first "supercomputer" to be built, the Atlas computer at the

University of Manchester in 1962. This computer was the first to use such advanced concepts as virtual memory and paging, and could compute at 200,000 floating-point operations per second.

The trend towards reduced size and increased reliability continued with introduction of the first integrated circuits, which were simultaneously invented by Texas Instruments and Fairchild Semiconductor in 1956. The main business computers of the 1960s, the IBM System/360 and the Digital Equipment Corporation PDP/8 were produced using this technology.

With the introduction of microprocessors, such as the 4004 CPU introduced by Intel in 1971, computers were set to begin entering the mainstream. The first pocket calculator was introduced a year later, as was the 8008 CPU, which was a 200 kHz version of the 4004, containing 3500 transistors. This "computer on a chip" could address 16 kB of memory, and inspired many electronics hobbyists to try to build their own computer.

The Altair 8800 was the first personal computer, and was offered for sale in kit form beginning in 1975. In the first year, more than 2000 sold, despite its steep price. This computer set the stage for the Apple II, which was offered for sale in 1977 as an assembled unit. For $298, you got a complete computer together with the keyboard, which could address 64 kB of memory and had monochrome computer graphics. The Commodore PET, introduced later that year, offered the first colour graphics.

In a 1977, Intel introduced the 8086 chip, which was the basis for the first IBM PC sold1981. Subsequent improvements to those first machines have led to the market for personal computers that we now enjoy today. Most computers today are based on the Pentium 4 chip, which is the seventh generation of the 8086 chip. Of course, today's chips are fallows of times faster in the original 8086, allowing for laptop computers equivalent in power to business computers of a decade ago.

It was advances in computer technology that made GIS possible in the first place, and these continue to make GIS more affordable and available to a wider audience.

1.2.6 Information Storage and Retrieval Systems

Although computers were originally built to perform difficult mathematical calculations, they can also be used to store and retrieve information. This has led to an entire industry based on the storage and retrieval of data, and many different solutions are available.

Although there are Information Storage and Retrieval Systems that are not completely computer based, for example automated Library retrieval systems, we will examine only those systems which completely store the data on a computer. These systems can be classified by the type of data that is stored, for example document retrieval systems, reference retrieval systems, or database management systems.

When computers were first created, the computer memory was considered secondary to the CPU, in which all of the number crunching was performed. The original idea was that the computer memory was a place for instructions and data to be stored before computer analysis. For this reason, computer memories tended to be quite small, barely large enough to store a computer program and some data.

Quite often, even today's vast computer memories are insufficient to store large numbers of documents. For this reason, numerous storage devices exist, from those that can be rapidly accessed, for example cache memory, to those that can only be accessed with a few seconds or minutes delay, such as might be found when a tape robot locates the correct make a tape and inserts it into a tape drive. There is a trade-off between speed of access and volume of storage.

Information Storage and Retrieval Systems are generally optimized to store and retrieve a particular type of data. For example, a system might be specifically designed to access and retrieve textual data, for example a full text library database. Other systems might be optimized to store and retrieve images, audio files, or video files. In each of these cases, a document is treated as an atomic unit, and never broken apart.

Another approach to the handling of data, is to disassemble the data and structure it efficiently. This is the database approach. Databases typically organize data into tables of related data, where a row in a table represents an individual object, and each column in the table represents an attribute of information about that individual object. Although the original object is not stored, pertinent information about that object is extracted, stored, organized, and compressed. This process of structuring data allows it to be retrieved more reliably and more rapidly than for unstructured data.

Unfortunately, spatial data does not neatly fit into the structure of rows and columns. Many GIS vendors have solved this problem by marrying a Database Management System (DBMS) to a proprietary system for storing, manipulating, and retrieving spatial data. These together form the core of Geographic Information Systems.

1.2.7 Geographic Information Systems

Now that we have discussed the enabling technologies for GIS, let's have another look at GIS. We have discussed the advances in thematic cartography that occurred in the 19th century, the developments in computers, and the development of Information Storage and Retrieval Systems, which are related to GIS.

Star and Estes (1990) list three developments that led to the development of GIS. These were:

1. Refinements in Cartographic Technique
2. Rapid Developments in Computers
3. The Quantitative Revolution in Geography

We have already seen how a gradual development in cartographic technique led to the development of topographic maps which featured accurate an consistent scales, followed by the development of single-purpose thematic maps. These developments were supported by advances in mathematics and the sciences of Statistics and Demographics in the 19th century.

Developments in computers have progressed, and continue to progress at an astounding rate. In 1965, Gordon Moore, the chairman of Intel, observed the doubling of the number of transistors on a manufactured die every year since 1961. This has come to be known as "Moore's Law," and rather improbably as remain true ever since 1965. The result has been computers that are now vastly more powerful than those we had even a decade ago.

Finally, GIS owes its development to the change of Geography from a qualitative discipline to a quantitative discipline in the 1960's, which made it possible to describe geographical phenomenon numerically, and attempt to model those phenomena. The Quantitative Revolution did more than to merely quantified geography: it shook up the thinking of many geographers and led them to look at problems in new ways. For example Peter Gould, a professor at Pennsylvania State University, analyzed the flow of mail as it traveled through Europe. Gould was able to determine that, because of linguistic preferences, there were invisible barriers to the flow of mail through Europe. France was found to be surrounded by a barrier, but connected to other Romance language countries. In describing this, Gould stated "We are no longer confined to conventional Earth space when we think of maps."

The Quantitative Revolution not only lead to new ways of looking thinking, but it also lead to new ways of analyzing data. Data that was quantitative could be analyzed by computer. Torsten Hägerstrand of Lund University in Sweden was able to model spatial diffusion using computer simulations as early as 1953. Ian McHarg, although not specifically a Geographer, also was able to look at things in new ways when he used transparent plastic overlays for the first time in his classic "Design With Nature" (1969). Although Dr. John Snow had predated McHarg by over a century, McHarg's book inspired many people, and came to be known as the "McHarg method."

Around the time of the Quantitative Revolution, a fair bit of experimentation was going on in cartography. For example, the University of Washington Department of Geography took an early lead in advance statistical methods and computer cartography between 1958 in 1961. Many later leaders in the GIS world, such as Tobler, Bunge, Berry, and Nystuen began there. In Britain, the Experimental Cartography Unit of the National Environment Research Council began some interesting work on the depiction of spatial data.

Around the same time, the integration of computers and cartographic technologies set off on two divergent paths. The first of these was to use computers to help automate the production of paper maps, which came to be known as Computer Cartography. The second path was that of the Geographic Information System, in which spatial data was stored in a computer. Although GIS started much slower than Computer Cartography, the ability of the GIS to store and manipulate the spatial data, producing maps of any desired type, eventually lead to the triumph of this approach over that of Computer Cartography.

Two early visionaries created the initial concepts behind GIS. The first of these was Howard T. Fisher, who created the first raster-based GIS called SYMAP in 1963. In 1962, Roger Tomlinson proposed the creation of the first vector-based GIS, which eventually became the Canadian Geographic Information System (CGIS). Work began on the CGIS in 1963.

Fisher's SYMAP, which stands for Synagraphic Mapping allowed the entry and display of geographic data. Maps could be printed on a line printer, using the overstriking of characters to allow different shades of darkness to be produced. Although Fisher began work on SYMAP while at Northwestern University in Chicago, the project was completed at the Harvard Laboratory for Computer Graphics, where Fisher moved to in 1965.

The Harvard Lab developed a number of experimental computer mapping products, including

- SYMAP (1963): the first raster GIS, which was simple to use, had limited functionality, and was able to print maps on a line printer using the overstriking of characters
- CALFORM (1960's): SYMAP functionality, but output was to pen plotters. Supported better legends and North arrows
- SYMVU (1969): Allowed perspective views to be created from SYMAP output
- GRID (1967): first raster overlays, allowed for multiple raster input layers
- POLYVRT (1974): this package supported raster and vector data together
- ODYSSEY (1975): Package allowed comprehensive vector overlay and analysis

Tomlinson's Canadian Geographic Information System was the first vector-based GIS, which allowed the Canadian Federal Government to create the Canada Land Inventory (CLI) mapping project. It took more than six years from CGIS to advance from a proposal to a system that was able to produce maps. By 1971, it was operational, and the first CLI maps were printed.

The CGIS contained seven themes initially: Soil capability for agriculture, Recreational capability, Capability for wildlife (ungulates), Capability for wildlife (waterfowl), Forestry capability, Present land use, and Shorelines. The CGIS was truly a pioneering effort. Much of what was built was built from scratch. The CGIS featured many firsts, including:

- The first scanning of maps (they had to build the scanner first)
- The vectorization of scanned maps
- The partitioning spatial database into individual map sheets (tiles)
- The separation of data into themes
- The use of an absolute coordinate system, so that map coordinates were registered to the real world
- The use of variable coordinate resolution, depending on requirements of individual map sheets
- The separation between graphical and attribute data
- The first use of Polygon Overlay
- Spatial topology
-

Many of these innovations were later incorporated into commercial GIS packages.

In 1969, Jack and Laura Dangermond founded the Environmental Systems Research Institute (ESRI) in Redlands, California. ESRI initially ran a consulting company, and the company's experience with the Harvard Laboratory family of products was helpful with later software development efforts. ESRI has become the world leader in commercial GIS development. Also in 1969, M&S Computing was founded, which was later became a major GIS vendor and was renamed Intergraph

In the 1970's, GIS was a very expensive proposition, and only could be attempted by large governments. These systems ran on very large computers and required dozens of support staff to make them work. A number of government agencies were involved in GIS in the 1970s, including the US Bureau of the Census, the Minnesota State Government, and of course, the Canadian Federal Government. In 1977 there were 54 GIS installations operational worldwide.

By the 1980's, continued reductions in the price of computer hardware made GIS more accessible to smaller organizations. Specialized graphics terminals and other peripheral devices such as scanners, digitizers, and pen plotters made GIS more powerful and useful than ever before. Of course, in these early days, GIS was viewed as a system for producing paper maps, rather than as a system for *processing spatial data* that can produce paper maps if they are required.

The reduction in the price of computer hardware lead many firms to get into the GIS marketplace. ESRI and Intergraph by now had over a decade of experience. They were joined by Universal Systems Limited (later CARIS) in 1979, Autodesk in 1983, ERDAS in 1984 (later bought out by Leica Geosystems), and MapInfo in 1986. In the mid-1980's there were many small GIS vendors, each attempting to find a particular market niche. Companies such as PAMAP Technologies, Terrasoft, and TYDAC who could not find a strong niche were joined by many others who went bankrupt as the market consolidated.

By the 1990's, the advent of computer workstations allowed companies to purchase and run GIS software. Personal computers became powerful enough for scaled-down GIS software to run on them, and so programs such as pc-Arc/Info, and later ArcView became available. By 1995, there were 93,000 GIS installations worldwide. Computer peripherals continued to develop, and the introduction of the colour inkjet plotter finally allowed high-quality GIS maps to be produced affordably. These plotters were relatively maintenance-free, and could produce about 50 plots overnight.

In recent years, the World Wide Web has become a destination for much data created using GIS. The Internet is rapidly replacing paper as the preferred medium on which to publish maps, since it can be updated immediately at virtually no cost.

1.3 The Scope of Geoinformation Science

GIS is one part a large group of scientific and technological disciplines known together as Geoinformation Science. Geoinformation Science is the scientific and technological disciplines that have to do with the collection, analysis, and distribution of spatial data. We include in our list such disciplines as Geographic Information Systems, Cartography, Surveying, Satellite Navigation Systems, Remote Sensing, Photogrammetry and Geodesy. All of these disciplines are involved in the creation, manipulation, and presentation of spatial data.

1.3.1 Spatial and Non-Spatial Data

There are two broad classes of data, *spatial data* and *non-spatial data.* Spatial data has some form of spatial information associated with it and non-spatial data does not. Spatial Data may have X-Y coordinates associated with each observation, or it may be associated with a spatial region such as a country or municipality. Of course, GIS focuses mainly on the storage, analysis, and presentation of spatial data.

Let's examine the use of spatial data. Consider, for example, the results of a recent census. You are likely to be presented with a series of tables showing mean income per census unit. If you simply have a list of the Census units in the corresponding mean income, you cannot analyze the data spatially simply because it is not included. If, on the other hand, you have the spatial outlines of the Census units, then it is possible to analyze the data spatially.

In the first example, what we have is *implicitly spatial* data. Although the spatial data was collected with the income data, it was not available to us, so we had analyze the data without considering the spatial distribution. The second example is *explicitly spatial* data. Here we not only have a table of data, but we also have the associated boundaries of the spatial area in which the data was collected.

Non-Spatial

House #=262
Assessed
Value=$100,00
0
Age=24
House #=343
Assessed
Value=$120,00
0
Age=15
House #=221
Assessed
Value=$82,000
Age=54

Spatial

🏠 **House**
#=262
Address=123
Main St.
Assessed

🏠 **House #=343**
Address=221
Smith St.
Assessed
Value=$120,0
00

🏠 **House #=221**
Address=406
Main St.
Assessed
Value=$82,00
0
Age=54
X=530,254
Y=5,632,704

Figure 9. Spatial and Non-Spatial Data

The distinction between implicitly and explicitly spatial data has more to do with the way the data is analyzed than the characteristics of the data itself. As we can see in the above example, the distinction is simply based on how much of the total data set was made available to us. Even if we have the locations of the houses, we can analyze this data in a spatially implicit fashion.

In most cases, however, unless the agency collecting the data sets out specifically to collect spatial data together with the non-spatial data, we will end up with non-spatial data. Although the agency collecting data may have no need for spatial information, the additional cost of collecting this information is nominal. With a properly configured GPS receiver, it might be as easy as pressing a single button. Consider the alternative, which would be to collect the spatial data separately at a later time. The cost of doing this would be as much or more than the original data collection, because not only do you have to travel the same route as the original data collector, but discrepancies between the spatial data in the non-spatial data have to be reconciled afterward, which is an extremely time consuming and difficult process.

Nearly all data that is collected has some spatial component, whether or not that spatial component is collected and stored with the data. It used to be that collecting the spatial component was extremely difficult, but now, with the advent of GPS, the additional costs are already quite low, and continue to decline.

The ability to integrate data from many different disciplines has resulted in a synergy which increases the capability and usefulness of GIS. The following sections briefly examine the main Geoinformation Science disciplines that create GIS data.

1.3.2 Database Management Systems

Just as Topographic Maps led to the development of the spatial overlay of themes that we use in GIS, Database Management Systems (DBMS) were an important step in helping to organize another part of the data used in GIS, namely the attributes that are associated with the objects identified on each theme.

We divide data stored in a GIS into *graphical data* and *attribute data*. Graphical data requires special routines for their display and analysis. Attribute data are handled by Database Management Systems in the same way that other tabular data are handled in non-GIS applications.

There have been a number of generations of DBMS since the 1960's. The first of these was hierarchical DBMS. A hierarchical DBMS works well when data is organized into a hierarchy. The hierarchy can be viewed as a tree structure, with a common element in the trunk, and more specific examples occurring as one moves towards the branches of the tree. A hierarchical DBMS works well when queries move from the trunk along successively smaller branches, and then back, but such systems are relatively slow if we need to find an item on a parallel branch, since we cannot connect directly between parallel branches, and must follow the links back to the trunk, and then move back out to the parallel branch. Not all data can be organized into a hierarchy, so there are some types of data for which a hierarchical DBMS will not work. In 1968, IBM introduced the Information Management System (IMS), which a hierarchical database that ran on an IBM System/360

One serious limitation of hierarchical DBMS is that it is only able to answer questions that are related to the hierarchy. For example, in an early implementation of hierarchical DBMS at Kew Gardens in London, the database management system was used to classify the botanical collection. The system worked very well for all queries related to the Kingdom, Phylum, Class... of particular botanical samples, until one day the Director wanted to know which botanicals were found in Mexico. At that point it was discovered that the system could not respond to a query for which it was not designed.

A more sophisticated DBMS system is network DBMS, in which the data are organized into layers, and an item in one layer may be linked to all of the items in the layer below or above. In a network DBMS, objects are linked together in a series of parent-child relationships. Each parent can have many children, and each child can have many parents. The concept for the network DBMS dates from the mid-1960s, when the Integrated Data Store (IDS) was developed by Charles Bachman at General Electric. In 1971, the Conference on Data Systems Languages (CODASYL) Database Task Group formalized the model for a network database. A recent example of a network DBMS is IDMS, from Computer Associates International Inc.

In 1970, E.F. Codd defined in entirely new class of DBMS in his paper "A Relational Model of Data for Large Shared Data Banks." This was the relational database in which data are divided into rectangular tables of data in rows and columns. Relational DBMS has a number of advantages over the older classes of DBMS:

1. Groups of objects or organized into tables. Each column in the table is used to store a particular attribute about an item; each row represents all of the data about that item. Thus, a row of data is a collection of variables related to a particular object; this is referred to as a *record*.

2. Tables can be constructed so that each represents a particular type of object, for example points, lines, or polygons.

3. Tables can be *normalized*, to eliminate the duplication of data.

4. If each object is given a common and unique key, this means that it can be identified in each table, and tables can be connected together using a common key in the procedure called a *join*.

5. All data storage, manipulation, and retrieval operations can be described in an English-like language called Structured Query Language (also known as Standard Query Language, SQL, or "Sequel")

One of the first relational databases was System R which was produced in 1973 by IBM. This early relational database eventually gave rise to SQL/DS and DB2. Oracle corporation also came from this line

Also in 1973, another effort to create a relational database by Eugene Wong and Michael Stonebreaker at the UC Berkeley led to a second line of databases. Wong and Stonebreaker created INGRES, which eventually led to Sybase, Informix, and NonStop SQL. Wong and Stonebreaker later create Relational Technology to commercialize Ingres.

In the 1980s, Object-Oriented DBMS (OODBMS) was developed to provide a more natural representation of how objects interact in the real world. In a OODBMS, objects or organized by hierarchy, whereby objects in the highest level contain objects at the lower levels using a process called encapsulation. For example, the object "District" contains a number of objects of type "address." Both the district object and the address object have their own attributes, and we can determine that an address lies within the district because the encapsulation is stored with the object.

1.3.3 Cartography

Cartography is the oldest Geoscientific discipline, and most refined. It has been the mandate of government agencies to produce maps and charts for general use for centuries. Because maps and charts are useful to a large proportion of the population, it makes sense for charting and mapping to be a government activity. In general, we use the term *map* for terrestrial applications, while we use the term *chart* for marine and aeronautical applications.

Private companies they also produce maps to support their own operations. For example, forestry companies will often produce maps showing information about the trees in forest stands. Despite the fact that these maps are expensive to produce, they are one of the costs of doing business in the forestry industry. The risks involved in cutting the wrong trees are significant and greatly outweigh the costs of producing accurate maps.

1.3.4 Surveying

It is the job of the surveyor to produce *precise* and *accurate* spatial data showing the outlines of buildings or parcels of land. Because of the importance of delineating land parcels exactly, the complex legal issues that ensue should a boundary prove to be wrong, surveyors are licensed professionals.

Surveyors use of specialized equipment to precisely determine *distances* and *angles*. This *radial coordinate system* allows the surveyor to measure relative to his or her instrument. One of the ways that is surveyor is able to measure very precisely is by removing errors in a *traverse*. In a traverse, the surveyor takes successive measurements such that they form a complete circle and the end point is the same as the start point. By measuring the amount of error that has accumulated by the end of the traverse, the surveyor is able to proportionately correct all of the errors that have been made during the traverse.

Survey information is an important source of data in GIS because it is both *precise* and *accurate*. However, the radial coordinate system used in surveys is incompatible with the *Cartesian coordinate system* used in GIS. In order to convert between these two systems, and the GIS technician can use *Coordinate Geometry (COGO)* software. With COGO software, the surveyors notes are entered, and the radial measurements are converted into a Cartesian coordinate system, in effect digitizing the measurements and putting them into GIS.

1.3.5 Satellite Navigation Systems

The first modern Satellite Navigation System was the NAVSTAR Global Positioning System (GPS), which was created by the U.S. military in the 1970s and 1980s for a number of military uses. Civilian uses for the GPS system soon followed, and then eclipsed the military use of the system.

From the initiation of the GPS program in 1973, it took 20 years and $12 billion before the GPS system was declared operational in December, 1993. The GPS system was used militarily even before completion of the system. In 1990, during the Gulf War, the U.S. military initially issued 1000 GPS units. By the end of the war, over 9000 units were in use, most of these civilian units.

Although the United States permits civilian use of the Global Positioning System, they ultimately retain control of the system, and can now selectively degrade the quality of the navigation signal to users in particular parts of the world. So, although the system works very well for civilian applications, it could potentially be made useless at any time due to a political or military emergency. With an increasing number of applications relying on GPS, such reliability issues have led governments around the world to question the long-term utility of the system for which they have no control.

The European response to this has been Galileo, a separate, independent Satellite Navigation System that is controlled by the EU. Galileo, will offer the same benefits as the GPS system, but will have increased accuracy over the current series of GPS satellites, because the Galileo system will feature more accurate atomic clocks. The first test satellite for Galileo (GIOVE-A) has been launched, and a second launch is in the works. Galileo is expected to be operational by 2010.

In addition to the GPS and Galileo systems, the Russians have a Satellite Navigation System known as GLONASS, which will be returning to full operation in the next few years, and the Chinese have been experimenting with similar systems as well. Although creating and launching a Satellite Navigation System is expensive (GPS cost US $12 billion to develop) the vast number of uses for such systems makes them increasingly attractive to those countries that have the resources to make such a system happen.

1.3.6 Remote Sensing

Remote sensing is a topic that it first appears daunting to many people. It is true that the discipline deals with satellites and complex topics such as spectral Imaging scanners and image classification, but for the GIS user, remote sensing images are a rich source of data that are used routinely.

Just about everybody has experience with remote sensing, although they may not know it. A camera is an example of a simple remote sensor. Most remote sensing systems are nothing more than glorified digital cameras which are able to store or transmit photographs as they are collected.

Remote Sensing systems can handheld, or can be mounted in any vehicle. Vehicles which carry remote sensing systems (referred to as *platforms*) may include satellites, aircraft, trucks and cars, ships or submarines. Early remote sensing consisted of human, and later photographic observation from balloons, such as was used during the American Civil War. This continued with the introduction of the aeroplane, with pilots making mostly oblique air photos. It was not until World War II that the taking of aerial photographs, and their subsequent analysis using the new science of photogrammetry became popular.

In 1972, the first Earth Resources Technology Satellite (ERTS-1) was launched. ERTS-1 (later renamed Landsat-1) was the first remote sensing satellite, and allowed the development of the field of remote sensing, one of the cornerstones of Geoinformation Science, together with DBMS and GIS. Landsat offered a Multispectral Scanner, which enabled images to be transmitted to the ground in one of five spectral bands: Red, Green, Near Infrared (2 bands), and Thermal Infrared.

Throughout the 1980's, the United States continued to lead the field, with the launch of Landsat 4 and Landsat 5, which offered the Thematic Mapper (TM), which was a much improved version of the Multispectral Scanner.

In the late 1980's and early 1990's the remote sensing environment became much more competitive, with the launch of increasingly capable and higher resolution satellites by a number of new players. In 1986, SPOT-1 was launched by France and in 1988, IRS-1 was launched by India. SPOT 1,2,3, and 4 offered 20m multispectral images and 10m panchromatic images, and SPOT-5 offers 10m multispectral images and 5m panchromatic images. IRS-1 offers 23m resolution for visual bands, and a 70m resolution for infrared, and 5.8m panchromatic resolution.

RADAR satellites came on the scene when the European Remote Sensing Satellite (ERS-1) was launched in 1991, and Canada's Radarsat was launched in 1995. ERS-1 is no longer operating, however ERS-2 offers 15.8 x 20m resolution RADAR images (variable resolution). Radarsat also offers variable resolution, with pixel sizes ranging in size from 8m to 100m.

During the 2000's, national remote sensing programs have been joined by commercial programs. India has made notable progress with its series of IRS satellites, with a total of six in orbit, and additional satellites being readied for launch. Commercial remote sensing satellites became available in 2000, with the launch of Space Imaging's IKONOS satellite, which offered 4m multispectral and 1m panchromatic capabilities. Digital Globe's QuickBird followed in 2001, and offers 2.8m multispectral and 70cm panchromatic images.

1.3.7 Photogrammetry

The discipline of Photogrammetry involves making precise measurements from aerial photographs. This may involve the use of a stereoplotter, which can determine the precise X, Y, and Z coordinates of objects found in stereo pairs of aerial photographs. This is done by precisely locating benchmarks or other points with known coordinates, and setting up a stereo model, which is an accurate representation of ground coordinates. Photogrammetrists are also involved with the creation of orthophotographs, which are mosaiced airphotos that have been corrected for relief displacement.

Within GIS, aerial photographs, orthophotographs, and remote sensing images may be imported into GIS and georegistered in order to place them correctly on the surface of the Earth. Once georegistered, remote sensing data can be used as a base layer, on top of which other data can be displayed, as a data source, from which new maps can be created, or as a data source for updating existing maps.

1.3.8 Geodesy

Geodesy is another discipline within Geoinformation Science. It is geodesy that is concerned with the shape and size of the Earth, and this discipline is key to ensuring that map projections are correct.

Summary

Geoinformation Science consists of a number of disciplines that are involved with the collection, processing and analysis of spatial information. Many of these tasks are related to GIS, because it is in GIS that the data is processed and analyzed.

Module Self-Study Questions

- In what ways is a GIS superior to paper maps, and what ways are paper maps still more effective than GIS?
- Which map elements can be found in a GIS user interface?
- Explain how thematic mapping was necessary before GIS could be developed.
- How can increased spatial and spectral resolution in remote sensing images improve data collection in GIS?
- Describe some ways that GPS is making spatial data more commonly available.

Required Readings

- Wilford, John Noble (1981). The Mapmakers. New York: Knopf. Prologue and Chapters 1, 5, 8 Index of Cartographic Images illustrating maps from the Ancient Period: 6,200 B.C. to 400 A.D. (http://www.henry-davis.com/MAPS/AncientWebPages/AncientL.html)
- PhysicalGeography.net: Fundamentals of Physical Geography. (http://www.physicalgeography.net/fundamentals/2e.html) Chapter 2: Maps, Remote Sensing and GIS section E, Introduction to Remote Sensing
- Smithsonian National Air and Space Museum (1998). GPS: A New Constellation (http://www.nasm.si.edu/gps/)

ESRI Virtual Campus Module

- Learning ArcGIS 9 Module 1: Getting Started with ArcGIS Desktop

Assignments

- Lab 1: Viewing data in ArcGIS
- Lab 2: Working with Multiple Layers

References

- Aber, James S. Brief History of Maps and Cartography (http://academic.emporia.edu/aberjame/map/h_map/h_map.htm) (Feb. 24, 2006)
- BBC News Online: World: Europe (2000). Lascaux caves reveal earliest star map (http://news.bbc.co.uk/2/low/europe/873365.stm) Feb. 27, 2007.
- Burrough, Peter & Rachael McDonnell (1998), Principles of Geographical Information Systems (2nd Ed.). Oxford: Oxford University Press, p. 5.
- Cartographic Images (http://www.henry-davis.com/MAPS) (Feb. 24, 2007)
- Chang, Kang-tsung (2006). Introduction to Geographic Information Systems. New York: McGraw Hill
- Chrisman, Nicholas R. History of the Harvard Laboratory for Computer Graphics: a Poster Exhibit (http://isites.harvard.edu/fs/docs/icb.topic39008.files/ History_LCG.pdf) (Mar. 11, 2007)Crone, G.R. (1968). Maps and Their Makers, 4th Ed. London: Hutchinson University Library.
- Digital Cameras (http://en.wikipedia.org/wiki/Digital_camera) (Feb. 24, 2007)
- Environmental Systems Research Institute. History of ESRI (http://www.esri.com/company/about/history.html) (Mar. 11, 2007)Ferri, Filippo (2000). GeoFile 2000-1: Preliminary Bedrock Geology between Lay and Wrede Ranges, North Central British Columbia (NTS 94C/12/8E; 94D/9,16) (http://www.empr.gov.bc.ca/mining/Geolsurv/Publications/GeoFiles/Gf2000-1/toc.htm) (Feb. 27, 2007)
- From One Revolution to Another (An Introduction to the Ordnance Survey) (http://www.ordnancesurvey.co.uk/oswebsite/media/features/introos/index.html)
- GIS Development History (http://www.gisdevelopment.net/history) (Mar. 11, 2007)Gregorvich, Andrew. Ancient Inventions of Ukraine (http://www.infoukes.com/history/inventions/) Feb. 29, 2007.
- History of Computing Project (http://www.thocp.net/timeline/1874.htm) (Mar. 10, 2007)Index of Cartographic Images illustrating maps from the Ancient Period: 6,200 B.C. to 400 A.D. (http://www.henry-davis.com/MAPS/Ancient%20Web%20Pages/AncientL.html) (Feb 24, 2006)
- Infoplease. Information Storage and Retrieval (http://www.infoplease.com/ce6/sci/A0825197.html) (Mar. 11, 2007)Klinkenberg, Brian. History of GIS (http://www.geog.ubc.ca/courses/klink/gis.notes/ncgia/u23.html - SEC23.5) (Mar. 11, 2007)
- Minerals (1974). (http://www.lib.utexas.edu/maps/europe/spain_mineral_1974.jpg) (Feb. 27, 2007)
- Moore, Gordon E. (2001). The continuing silicon technology evolution inside the PC platform. Intel Dev http://www.intel.com/update/archive/issue2/feature.htm eloper Update Magazine http://www.intel.com/update/archive/issue2/feature.htm (Oct. 19, 2002).
- Star and Estes (1990). Geographic Information Systems: an introduction. Prentice Hall. 303 p.
- Taubes, Gary and Kleppner Daniel. The Global Positioning System: The Role of Atomic Clocks (http://www.beyonddiscovery.org/ content/view.page.asp?I=463) (November 11, 2006)

- Using Oracle Data Mining to Analyze Sequence Data (Source: http://www.oracle.com/technology/obe/obe10gdb/bidw/blast/blast.htm) (Feb. 27, 2007)
- Wikipedia. Antikythera Mechanism (http://en.wikipedia.org/wiki/Antikythera_mechanism) (Mar 10. 2007)Wilford, John Noble (1981). The Mapmakers. New York: KnopfWorld-Wide Media Exchange (http://wwmx.org/) (Feb. 24, 2007)

Terms Used

- Canadian Geographic Information System (CGIS)
- Cartesian
- Coordinate Geometry (COGO) Software
- Coordinate Systems
- Data
- Database Management Systems (DBMS)
- Digital Paper
- Digital Storage
- Explcitly Spatial
- General Cartography
- Geodesy
- Geographic Information Systems (GIS)
- Georegistration
- Geoinformation Science
- Global Positioning System (GPS)
- Graticule
- Landsat
- Non-Spatial
- Ordnance Survey
- Relational DBMS
- Remote Sensing
- Remote Sensing Platforms
- Satellite Navigation Systems
- Spatial
- Spectral Imaging Scanners
- SYMAP
- Thematic Cartography

2 Introduction to Geographic Information Systems (GIS)

2.1 Definitions of GIS

2.1.1 Introduction

One of the most difficult things for new GIS students, as well as for experienced GIS practitioners, is to come up with a concise definition of what exactly GIS is. After all, GIS is used for many different applications in dozens of different industries.

Like the Hydra of Greek mythology, a mythical monster having one body and nine heads, GIS has a single body (spatial data storage) and many heads (applications), each of which is a different way to make use of stored spatial data. In a GIS, we separate the storage of spatial data from the display of the data, and in-between we insert the analytical power of the computer. Because the computer can make use of different software, many different things can be done with the spatial data; the only limits are imposed by the nature of the spatial data, the sophistication of the computer software, and the speed of the computer on which the software runs.

As soon as we master one application, and chop one head off the monster, two new applications become available, just like the Hydra, to take its place. It's not surprising that GIS is a dynamic field of study, where there is always something new to learn.

How do you describe a nine-headed monster? It is difficult to describe each of the nine heads in detail, so you are forced to describe the whole in very general terms, or to focus on a particular head that you find most interesting. This is the nature of the definitions for GIS. Some definitions are very specific, and others are very general. Short of writing a book on the topic, it is impossible to explain all the details.

2.1.2 Five Definitions of GIS

Geography, and GIS by extension, is an inherently multidisciplinary field of study. There was no single "inventor" of GIS; many individuals and teams recognized that there was a problem with the way that mapping was being performed in the early 1960s, and set about fixing it. These teams came from disparate fields such as Geography, Computer Science, Data Processing, and Information Technology.

The problem with maps was simply one of data volume. The science of the 1960s was bringing in more data than ever before, at a pace that was unthinkable even 10 years before. It's not surprising that, with computer information systems being developed for other types of data, geographical data would not be far behind. It was much more difficult, however, to store and manipulate geographical data than it was to store and retrieve or other types of simple data, such as numbers and text.

It was not until the early 1990s that the ability to store and retrieve spatial data became commercially available to the mass market. Even now, geographical data are difficult to manipulate. As the technology matures, new paradigms will emerge for handling spatial data, and with them will emerge new definitions as well.

The following five definitions of GIS all help to explain what GIS is, yet none succeed at being detailed while at the same time providing a good overall picture.

- The Map View: GIS is a storage and display device for electronic maps.
- The Functional View: GIS is a set of computer *hardware* components.
- The Geoprocessing View: GIS is a *software* toolbox.
- The Database View: GIS is software for storing and retrieving spatial *data*.
- The Systems View: GIS is a *system* composed of hardware, software, people, data, procedures, and applications

The Map View

The Map View of GIS is the most popular amongst new users of GIS. This view is popular because it allows people to build upon what they already know (maps) and to use this to understand a new technology. In the Map View, the GIS can be considered to be a very fancy electronic storage cabinet for maps. In the storage cabinet, there are many interesting maps that can be examined. Not only are these maps viewable as static documents, but they can also be displayed dynamically, showing three-dimensional or time-series views of the data.

GIS offers a number of sophisticated ways of viewing maps in the Map View. Users are able to zoom in, zoom out, and pan to view different portions of the map. Different thematic maps (layers) can be turned on or turned off, and these can be displayed using a variety of cartographic symbols. In addition, the unsymbolized data can be queried directly to determine the data values that were used to create the symbolized maps.

When Web-enabled, the Map View becomes an electronic atlas. An example of this is the Electronic Atlas of Arizona (http://atlas.library.arizona.edu/map.html). The Map View is a good definition for GIS because it describes the overall picture in a limited number of words. It can explain to the uninitiated what GIS is, and allow them to get a basic understanding of GIS. Unfortunately, the Map View does nothing to explain what goes on within a GIS to make the storage and display of electronic maps possible.

The Functional View

In the Functional View, the GIS is basically a computer. This view, which is favoured by computer technicians, views a GIS as a series of computer components. In the Functional View, the GIS has components to perform the following tasks:

- Inputting data (keyboards, digitizers, scanners, data collection computers...)
- Processing data (a computer having sufficient memory and processing speed, software)
- Outputting data (computer monitors, printers, plotters...)
-

With the Functional View, you gain an understanding of the machinery on which a GIS runs. However, this view is also incomplete. An interesting thing to note about this view is that the GIS does not have an operator; there is an empty seat waiting to be occupied by the person who holds this view. There is no concept of an organization surrounding the GIS or rules and procedures, and the role of computer software and spatial data are minimized.

The Geoprocessing View

The Geoprocessing View of GIS looks at the computer as a software toolbox. Just as the Functional View is overly focused on computer hardware, the Geoprocessing View is overly focused on the software.

In the Geoprocessing View, the GIS is a repository of unprocessed spatial data, awaiting analysis. Through the clever use of software tools and linkages between these tools, the GIS operator can create a process to convert the unprocessed spatial data into a desired result. This view focuses on the steps that are necessary to transform raw data into useful information.

While the Geoprocessing View is an effective description of the way that some experienced GIS operators work, and it describes the way that many consultants use GIS, in reality, few GIS users attain this level of sophistication. For them, the Map View of GIS is sufficient. As with the Map View, this view ignores the role of the Organization, rules and procedures, and the GIS operator.

The Database View

Professional Information Technology people, in particular DBMS Experts, who use many different forms of data storage and retrieval tools, tend to view GIS is simply another form of IT tool. As with the Geoprocessing View, the Database View is software-centric, however, in the Database View, it is the storage and retrieval of spatial data that is the main focus, not the processing of those data.

Two examples of the Database View are espoused in the following quotes:

> A database system in which most of the data are spatially indexed, and upon which a set of procedures operated in order to answer queries about spatial entities in the database (Smith et al., 1987)
>
> Any manual or computer based set of procedures used to store and manipulate geographically referenced data (Aronoff, 1989)

Because the Database View focuses on the storage of spatial data, it places a strong emphasis on the quality of the data. Above all, data are something that should be reused, if possible. Poorly structured or poorly documented data are difficult to reuse and should be avoided. Thus, people holding this view expend a great deal of energy keeping a "clean" database. In this view, GIS data should have the following characteristics:

- They should be simple to use, and easy to understand
- They should be easy to use with other data sets
- They should be effectively compiled and validated
- They should be clearly documented for content, intended uses, and purposes (Source: What is GIS?)
-

This view of GIS lends itself quite naturally to the concept of the Data Warehouse, which is simply a very large collection of spatial data. One problem with this concept is that a great deal of effort is spent on the validation and correction of data to meticulous standards. It is much more cost-effective to ensure that the data are collected correctly in the first place, than to go back and attempt to make corrections where errors are found, and risk undetected errors not being found at all.

The Systems View of GIS

So far, we have seen one holistic view of GIS (the Map View), and three detailed views of GIS (the Functional, Geoprocessing, and Database views). The Systems View of GIS is a holistic, but sophisticated look at GIS. In the Systems View, a GIS has the following components:

- Hardware
- Software
- People
- Data
- Procedures
- Applications
-

The Systems View (Figure 1) incorporates the Functional, Geoprocessing, and Database views of GIS, and recognizes the importance of computer hardware, software, and spatial data. However, this view also recognizes that GIS never exists in a vacuum.

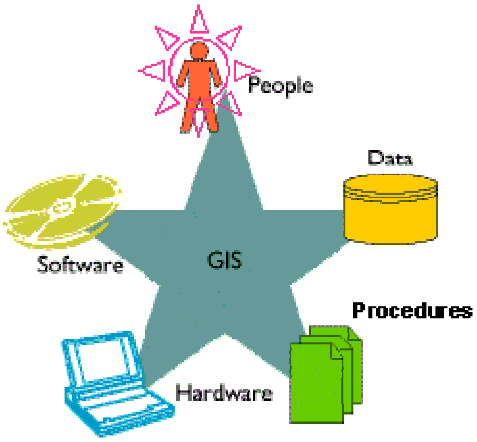


Figure 1. The Systems View of GIS. Source: http://www.esri.com/library/whitepapers/pdfs/healthserv.pdfA GIS is an expensive proposition, whether it involves purchasing commercial software, or modifying open source software. There are significant costs and resources required for the acquisition and maintenance of computer hardware and spatial data. For these reasons, a GIS will rarely exist without an *application*. There has to be some important problem to be solved, for which agency is willing to spend large amounts of money and effort in investing in GIS as a potential solution. It is important that the application never be forgotten, because the degree to which the application is addressed is one of the few concrete measures of the effectiveness of a GIS project. Because GIS is useful in so many different areas, it is easy to lose track of the application, and waste resources solving problems that were not in the original mandate, while ignoring the application.

From the application flows a set of *procedures*, which are the means by which organizations (i.e., groups of *people*) solve the problems inherent in the application. From the application will flow a list of software requirements that can fulfill the procedures. In addition, the procedures and software requirements will help to define a list of skills that will be required by project staff.

2.2 The Role of GIS Within Geographic Information Science

GIS is only one of many different subdisciplines within the Geographic Information Science. Other fields, such as the Surveying, Satellite Navigation Systems, Photogrammetry, and Remote Sensing all have vital roles in the collection and processing of geographic information.

Geographic Information Systems have emerged as the central component of a number of geospatial technologies that together are referred to as "Geographic Information Science" (GIScience). All of the technologies within GIScience have to do with the collection, processing, or output of spatial information. GIScience technologies include:

- Surveying
- Satellite Navigation Systems
- Photogrammetry
- Remote Sensing, and
- Geographic Information Systems, which incorporate
- Database Management Systems, and
- Cartography

Geographic Information Systems act as a central repository for spatial information, and the tool in which those data are processed and displayed. All of the data created by Surveying, Satellite Navigation Systems, Remote Sensing, and Photogrammetry consist of spatially discrete or continuous data.

Discrete data are stored using a vector data model, in which features are represented as points, lines, or polygons. These discrete features have abrupt boundaries, and do not blend into one another. A city, for example, might be represented by a polygon, since it has a discrete legal boundary.

Not all data have abrupt boundaries, however. For those data that do not have abrupt boundaries and very continuously over a surface, we use the raster data model. The raster data model represents spatially continuous data as a grid of rectangular cells that completely cover a study area. Each cell is assigned a unique value, which can be used to represent an elevation or a reflectance value, among other things. Continuous data can also be represented as a Triangulated Irregular Network (TINs), which offer variable resolution, and are superior at representing complex terrain. These will be discussed in a later section, when we examine the vector data structure in detail.

Because vector and raster data models are based on the underlying characteristics of spatial data, it is possible to convert virtually any type of spatial data into one or the other of these models. A number of conversion tools are available to convert external file formats into an internal data format used in the GIS.

Survey data, for example, consist of a series of points, lines, and areas, representing benchmarks, tie points, or instrument positions (point), bearings or traverse lines (line), or building outlines (polygon). Because of the way that surveys work, surveyors work in terms of radial coordinates, consisting of bearings and distances, rather than the Cartesian coordinate system that is used in map projections and GIS. Thus, there is a special conversion process to input a Surveyor's notes into GIS called Coordinate Geometry (COGO). Survey data are represented in a GIS using the vector data model.

Data from Satellite Navigation Systems, such as the Global Positioning System (GPS), are represented in GIS using the vector data model, as well. Unlike survey data, GPS uses Cartesian coordinates, so it is relatively easy to input the points, lines, and polygons collected with a GPS receiver into a GIS.

Photogrammetry literally means the measurement of photographs. It is the job of the photogrammetrist to identify features such as buildings, roads, and forest polygons that are visible on aerial photographs. To accomplish this, the photogrammetrist makes use of a device called a stereoplotter to precisely follow the outlines and centerlines of objects on registered aerial photographs. The identified objects are stored digitally as points, lines, or polygons, and these data can also be converted into the internal GIS vector format.

Photogrammetrists may also use digital stereoplotters to create orthophotographs, which are airphoto mosaics that have been corrected to remove distortions caused by camera tilt and relief displacement. Orthophotographs are planimetrically correct, and may be overlaid with map data, or may be used directly as a data source.

Other types of data collection systems used within Geographic Information Science produce continuous data. This is the case for Remote Sensing. Remote Sensing returns information in the form of rasters of colour information. These raster data can be incorporated into a GIS by converting the data format produced by the satellite into the internal GIS raster data format.

Because of the ability of GIS to store all types of spatial information, GIS has become the centre of Geographic Information Science. Within GIS, all spatial data can be combined and analyzed to solve geographical problems. The following section describes how GIS combines data from the different sub-fields of Geographic Information Science.

2.2.1 Case Studies: The Role of GIS in GIScience

In Module 1, we discussed the scope of GIScience, and examined the different fields of study that combine with GIS to make up this dynamic field. In this section, we will examine a number of case studies which illustrate how fields of study such as Database Management Systems, Cartography, Surveying, Satellite Navigation Systems, Remote Sensing, and Photogrammetry all work together with GIS to create a unified field with powerful capabilities.

The Cartographic Editing System

The Cartographic Editing System (CES) is the GIS application created by the Topographic Mapping Division of Energy, Mines, and Resources Canada to support the publication of 1:50,000 and 1:250,000 topographic maps. This system focuses on the ability of GIS to store and present data effectively to create an entirely digital alternative to the previous system, which combined digital and analog elements.

Digital stereoplotters are used to collect data from aerial photographs. By identifying points with known X, Y, and Z coordinates (Ground Control Points) on the airphotos, it is possible to correct for the scale, rotation, and tilt of the aerial photographs to create a stereo model. The features entered by the stereoplotter operator become immediately available in digital form, and they are then used to update the features in the digital database.

The CES was designed to meet the following criteria:

- Accept digital data files in various formats from many sources including the National Topographic Data Base
- Produce quality text and symbology for point, line, and polygon features
- Allow onscreen graphical editing of digital map files and final compilation of maps
- Allow for the automation of manual procedures
- Have high throughput, to allow large amounts of data to be edited, and
- Be able to send its output to Calcomp and Scitex plotters, for production of screen-ready plate negatives
-

This CES, being an entirely digital system, has reduced equipment and labour costs for map production. Computers replaced old pieces of analog equipment, which were difficult and expensive to maintain. In addition, the ability to accept digital files made it possible for the first time to have subcontractors involved in the map production process, which has reduced staffing requirements.

In this example, we see that the storage and cartographic abilities of GIS were critical for this application. The ability to automate manual procedures through programming was an additional capability provided by the GIS that was not possible with the old digital/analog system. Relatively little use was made of the GIS analytical capabilities, however, because the data are being edited and structured, these data could potentially be made available to other users of GIS (Donner, 1992)

Property Fabric Identification System

The Property Fabric Identification System (PFIS) is an application to store survey data for all surveys that have been conducted by the Legal Surveys Division of the Canada Center for Surveying in Energy, Mines, and Resources Canada. The Legal Surveys Division is responsible for all legal surveys in Canada's 2300 Indian reserves, National Parks, and all public and private lands in the Yukon and Northwest Territories, which are administered by the Canadian Federal Government.

Legal records at the Legal Surveys Division go back over 100 years, and include more than 70,000 survey records, as well as related documents such as airphotos and maps. The volumes of data have made it very difficult to effectively access paper-based survey records. Carkner and Egesborg state "Conventional methods of accessing and managing records can no longer meet the day-to-day information needs of internal users and the various other outside agencies and individuals in the public and private sector" (Carkner & Egesborg, 1992, p. 272).

The new system combines a Relational Database Management System and a GIS to allow all the records to be stored and accessed efficiently. All projects are geographically tagged, and can be retrieved via map or textual query. Because the system contains survey data, including legal descriptions of property boundaries, this system contains only original survey data, and no derived or adjusted data. The only provision for the adjustment of data are the conversion of source data points from one map projection to another to allow them to be viewed together in the GIS.

When new data are entered into the PFIS, it is compared to the existing survey data in the system. This allows erroneous measurements to be quickly identified and corrected. New data, such as the locations of survey and control monuments, lot plans, and legal descriptions are entered continually.

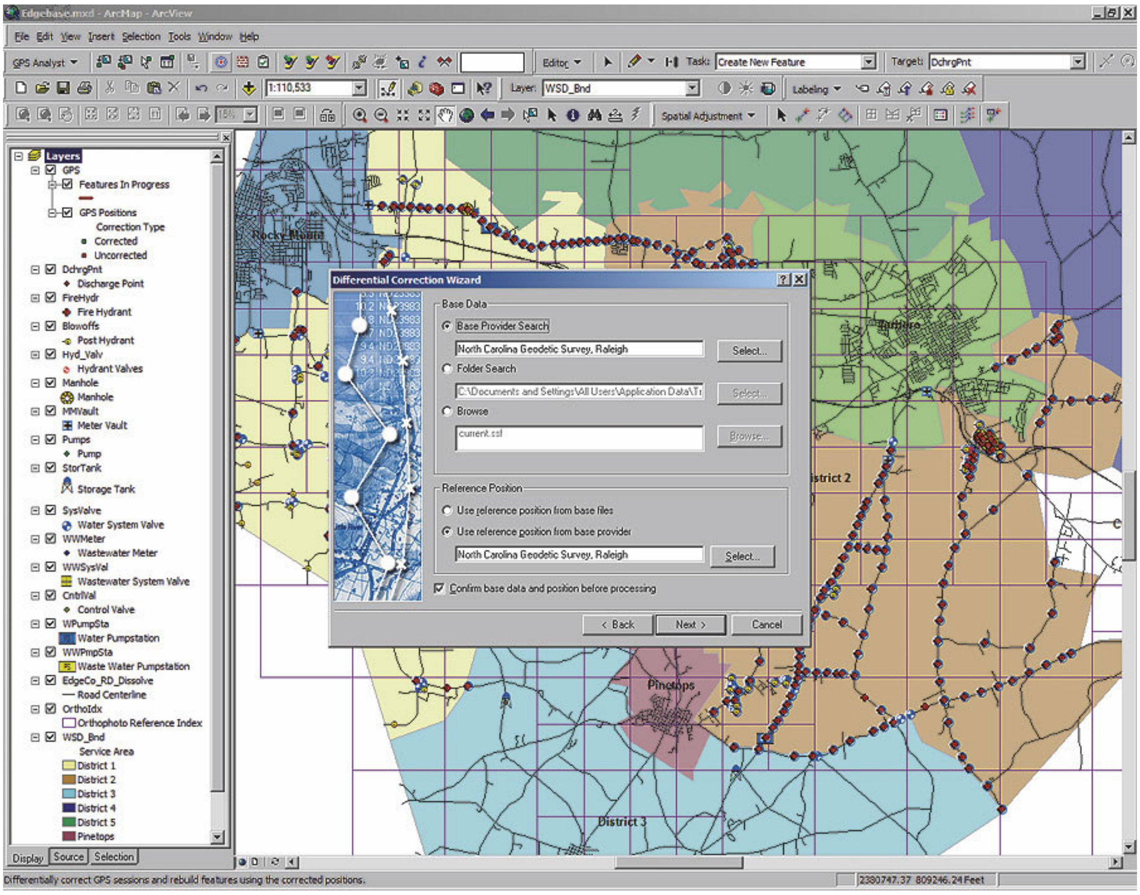
In the PFIS, the GIS is being used once again primarily as a tool for data storage and retrieval. The ability of the GIS to convert survey notes into graphical representations of what was surveyed using COGO is employed in this system, but little use is made of the analytical capabilities of GIS in this application. The system imports survey data, and employs a Relational Database Management System to allow for the effective storage and retrieval of that data. The GIS enables the data to be displayed in conjunction with other surveys of the same area. As with previous example, a large amount of well-structured digital data is being collected, which can be used in future for GIS analysis (Carkner and Egesborg, 1992).

Water and Wastewater Infrastructure Mapping

In the previous two examples, we have seen how data from aerial photographs, and from survey records have been imported into a GIS. One other GIScience field of study that can provide input to GIS is Satellite Navigation Systems, such as GPS.

Edgecombe County, North Carolina, U.S.A., undertook a project to replace its entire water and wastewater distribution system. To do this, it required an accurate map of the location of all infrastructure in the water and wastewater systems, including pressurized lines, gravity lines, manholes, control valves, storage tanks, fire hydrants, meter vaults, and pump stations. The county covers 1307 km^2, and has approximately 650 linear kilometres of water and wastewater lines. The mapping had to be completed in two months.

To accomplish this task, the county hired a consultant who used handheld computers equipped with GPS and mapping software to record the locations of all features in the water and wastewater system. More than one dozen attributes were collected for each feature mapped, and a number of features such as fire hydrants were "discovered"; they had never been mapped before. At the end of each day, the handheld computers were connected to desktop computers at the consultant's main office, and the data were seamlessly downloaded, differentially corrected, incorporated into the spatial database showing the water and wastewater system (Figure2).



Task: Create New Feature

Target: DchrgPnt

1:110,533

Layer: WSD_Bnd

Base Data

● Base Provider Search

North Carolina Geodetic Survey, Raleigh

○ Folder Search

C:\Documents and Settings\All Users\Application Data\Tr

○ Browse

current.ssf

Reference Position

○ Use reference position from base files

● Use reference position from base provider

North Carolina Geodetic Survey, Raleigh

☑ Confirm base data and position before processing

Differentially correct GPS sessions and rebuild features using the corrected positions.

2380747.37 809246.24 Feet

Figure 2.Raw GPS data is automatically downloaded to a desktop computer for differential correction. Source: http://www.thewootencompany.com/edgecombe_casestudy.pdf

Tests of the GPS-based data collection system showed that, after differential correction, mapped features or accurate to within 1 m of their actual location. The data were verified by comparing the location of features that were visible on a georegistered orthophoto with those reported by the handheld computer. The speed of this method of data collection allowed the mapping process to be completed within the two-month timeframe. Using this system, the consultant was able to build a topologically correct network of water and wastewater lines, which could then be used to support the replacement of the existing system with the new one.

In this example, we see yet another way in which GIScience activities support the use of GIS. In all cases, the GIS has been used as a central repository for spatial data, but little use has been made of the data inside the GIS (Fuller, 2005).

Lyme Disease Prediction Mapping

In Westchester County, New York, U.S.A., there have been a large number of cases of Lyme Disease. Lyme Disease is the most common vector-borne disease in the United States, and is characterized by a wide range of symptoms, initially including skin rashes, fever, headache, and fatigue, and then proceeding to arthritis-like symptoms, heart problems, and problems with the nervous system. Most cases can be treated with antibiotics, at least in the initial stages.

Westchester County is located near forests where the tick *Ixodes scapularis* lives. The tick is the vector for the bacterium *Borrelia burgdorferi*, which causes the disease in humans, small vertebrates, and white-tailed deer.

Researchers began with maps of land cover classes and infection rates in the county, and used a statistical package to identify which land cover classes were most highly correlated with infection of humans. Once this was established, Landsat TM satellite imagery was classified to identify land cover classes, and was imported into the GIS.

Only those forested areas containing Lyme Disease that are adjacent to residential areas are of concern in this analysis. These areas can be targeted for disease eradication efforts. To identify these regions, a 3 x 3 "moving window" is used to identify when pixels of Lyme Disease infected forest are adjacent to pixels representing residential areas. The analyzed satellite imagery was then used to produce a map of all of Westchester County, showing the risk of Lyme Disease in residential neighborhoods throughout the county (**Klaida! Nerastas nuorodos šaltinis.** 3).

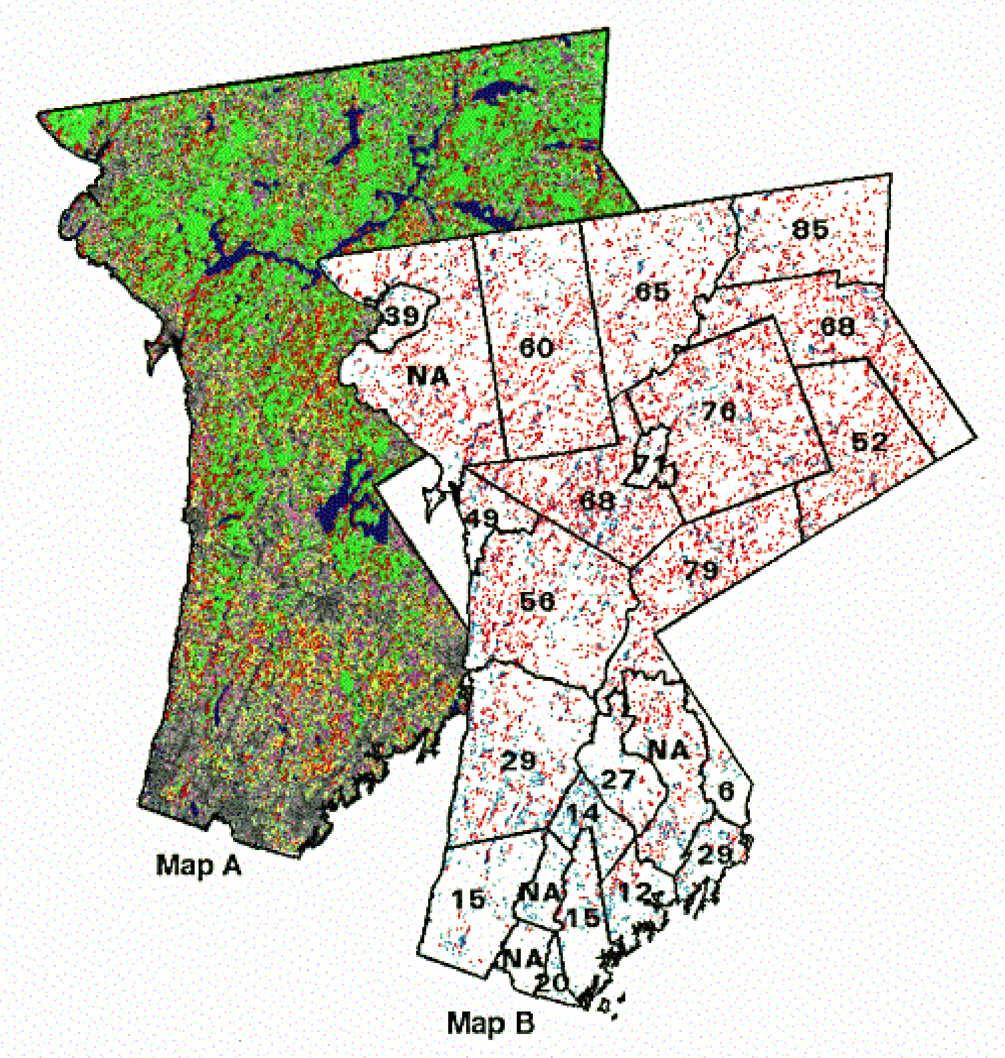


Figure 3.Maps of Westchester County showing Ixodes scapularis tick infection. Map A shows the forest lands with the highest infection rates; Map B shows residential lands that are in close proximity to the infected forest lands. Source:http://geo.arc.nasa.gov/sge/jskiles/fliers/gif_folder/image19/image19a.gif

In this example, we see GIS being used with specialized packages for statistical data processing, and classified satellite imagery being used as a data source for GIS. The analytical capabilities of the GIS are used to determine where Lyme Disease is most likely to affect people, and the GIS is then used to produce maps showing these areas (Clark, 1997).

2.2.2 Collection of Spatial Data

Now that we have examined a number of applications that made use of GIS and other disciplines within GIScience, let's examine the components that make up a GIS in detail. The rest of this module is based on the Systems View of GIS, since it is the most comprehensive definition of GIS that we have examined.

Although many of the other technologies within Geographic Information Science have the ability to collect data that can be processed with a GIS, it is also possible to collect data directly into a GIS, by reading files of digital data, or by scanning or digitizing paper maps. In all cases, the data that is entered into the GIS has to be appropriate for the project that the GIS is being used on.

Know your client

In order to determine what is appropriate, it is important to have a good understanding of the institutional environment in which the GIS exists. Is important to know your client, and maybe even your client's client. In other words, having a complete understanding of why a GIS is being built, and who's paying for the GIS to be built are all critical to the success of a GIS project.

Too often, GIS Technicians, and even GIS Managers become so involved with the day-to-day challenges of running a GIS, that they forget their client's requirements. It is easy to become obsessed with a particular technical solution, and forget that other solutions may equally satisfy the client's needs. In the first case, the GIS manager and his team may exhaust themselves on a very difficult and technically demanding solution, causing cost and time overruns. In the second case, another solution may be readily available, which meets all of the needs of the client, and can be accomplished inexpensively and quickly. Quite often, the only difference between blindly proceeding down the first path, and altering course to pursue the second, is an understanding of the needs of the client, and what solutions will be acceptable and which will not.

Collect only what you need

Understanding a client's requirements and the environment in which a GIS exists is very important in helping to determine which data should be included in the GIS. DeMers (2005) states that a common practice in new GIS projects is to collect all the data that can be collected. This is most certainly a recipe for failure. Although such a policy may seem to make sense at the time, it is important to remember that every piece of data have to be stored, backed up, examined from time to time, updated, and documented, even if it is never analyzed a single time.

It is crucial to remember that those people who hold the *map view* of GIS may not be aware of these issues, and may simply view a GIS as a very large map cabinet, in which unimportant maps can simply be left in an unused drawer without causing a great deal of difficulty. It is important to know what data are important for the current roles the GIS is expected to fill, as well as for future roles. So there is some leeway in the data that should be collected for GIS. Those data that are required to meet the current goals obviously need to be included, and those data that are likely to help with future applications of the GIS should also be included. Those data that meet neither of these requirements should not be incorporated into a GIS (although it might be worthwhile to note the content and location of these data in case some unusual and unforseen application becomes a priority in future).

Seven rules

DeMers (2005) suggests seven rules for the collection of GIS data. These are:

1. Determine why you were building the GIS
2. Define your goals as specifically as possible before selecting layers

3. Avoid the use of exotic sources of data when conventional sources are available
4. Use the best, most accurate data necessary for your task
5. Remember the law of diminishing returns when deciding on data accuracy levels. Data that are more accurate than your minimum requirements may make it more difficult to answer the questions for which you built your GIS
6. When ancillary data are available from data sources, include these as separate layers, and
7. Each layer should be as thematically specific as possible. There is overhead associated with having to maintain unnecessary data in a GIS (DeMers, 2005, p. 126-127).

Accuracy of Digitized Maps

When digitizing a map to enter it into a GIS, the scale of the map determines the accuracy of the data that can be extracted from it.[1] A general rule of thumb in cartography is that object positions are accurate to within 0.5 mm on a paper map. The more area that is represented on a map (i.e. the smaller the scale), the more area on the ground that is covered by 0.5 mm on the map.

For example, at a scale of 1:1,000,000, 0.5 mm at map scale represents 500 m, and at a scale of 1:50,000, 0.5 mm represents 25 m. In other words, the accuracy of a location on the ground depends on the scale of that map. Table 1 summarizes resolutions and polygon sizes for maps printed at different scales

Map Scale	Minimum Accurate Measurement (0.5mm at map scale)	Minimum Accurate Polygon Area
1:10,000	5 m	0.0025 Ha
1:50,000	25 m	0.063 Ha
1:100,000	50 m	0.25 Ha
1:250,000	125 m	1.56 Ha
1:1,000,000	500 m	25 Ha
1:5,000,000	2500 m	625 Ha

Table 1. Minimum Resolutions and Polygon Areas for Map Data at Different Scales

Minimum Resolution

Depending on the type of analysis that you wish to do with your GIS, you will want to set a minimum standard for spatial data resolution. How small do you go? Modern GIS store their coordinates using double-precision accuracy, which can accurately locate features to 0.004 µm. So, in theory, a GIS allows the location of a virus to be accurately recorded on a map of the entire Earth!

If GIS technology is not the limiting factor, then how do we determine an appropriate resolution for our GIS? Fortunately, discoveries in the field of Information Theory can help us. Shannon (1948) discovered that for a voice signal to be effectively transmitted, the smallest unit of that voice signal had to be sampled twice. In other words, in a one-dimensional system, the smallest piece of information that can be transmitted must be at least double the size of the unit of measurement. In a two-dimensional system such as a map, the smallest object to be mapped

[1] Technically, in a GIS, the scale at which the data were collected (measurement scale) is independent of the scale at which the map was printed (cartographic scale). On paper maps, however, the compilation and printing of the map means that the cartographic scale overrides the measurement scale.

should be two units wide in one dimension and two units wide in the other dimension, so the smallest object should be four square units in size (DeMers, 2005). This explains a long-standing rule-of-thumb in GIS, which is that the resolution of the data file should be half the size of the smallest object to be mapped.

If, for example, the smallest object that you wish to store in the GIS were 20 m across, then a good guideline would be to set your minimum spatial data resolution to be 10 m. This would ensure that the feature was mapped. Of course, if this object has to be represented with some fidelity, then it might be desirable to consider a resolution of less than 10 m.

Knowing your minimum standard for spatial data resolution immediately allows you to include or exclude spatial data sources based on their minimum resolution. In the above example, if our minimum spatial data resolution had to be 10 m, then a Landsat Thematic Mapper (TM) image having a resolution of 30 m would clearly be inappropriate, but a SPOT panchromatic image having a resolution of 10 m would work. A paper map at a scale of 1:20,000 would meet the minimum spatial data resolution, since 0.5 mm on the map would represent 10 m on the ground. In addition, the minimum standard for spatial data resolution allows us to determine what sort of data collection techniques will be appropriate for the needs of our GIS.

Data Collection Methodology

It is important to ensure that data are collected from maps having an appropriate scale, but that is only the beginning of the process of data collection. Data may be collected using either vector or raster data formats.

Vector Data Collection

In the section, we discuss issues in the context of digitizing of paper maps, but these issues are equally important for maps that are scanned, or for data that are digitized from orthophotographs. The process of entering data from a paper map of appropriate scale is called digitizing. Digitizing requires a specialized piece of equipment called a digitizing tablet in order to allow an operator to enter lines from the map into the GIS.

The resolution of the data source for GIS controls the minimum spacing that is advisable between adjacent points. However, it is rarely advisable to use this minimum possible resolution for a line, unless the line is extremely complex. For example, if digitizing the meanders in a river, it might be necessary to digitize points with the separation equal to the resolution of the data set. For all other lines, however, using this data density is not only unnecessary, but it is counterproductive. If the line is known to be straight, then it is optimally identified with two points, the start point, and the endpoint. Any other points in the line only reduce the accuracy of this line. DeMers (2005) gives the example of an outer polygon on a United States Geological Survey (USGS) quadrangle topographic map being digitized with 2000 points, despite the fact that it is a straight line. Not only does this reduce the geometric perfection of a two-point line, but it also creates unnecessary data volumes. If this practice occurred throughout a GIS database, the GIS would be unnecessarily large, and would perform very slowly. Of course, if other lines intersect the polygon, then it is necessary to split that line, but all segments between intersecting points should have only two points in them.

This brings us to the issue of point-to-point digitizing versus stream digitizing. Digitizing software offers the operator two choices. With point-to-point digitizing, the user enters every single point in a line. This allows the user to digitize with high precision on complex lines, and low precision on straight lines. Stream digitizing is different in that the computer enters the points, and all the user does is guide the digitizer cursor over the map. A number of settings control the minimum distance between points that are entered during stream digitizing. In the above example, mentioned by DeMers, it is likely that the operator used stream digitizing to digitize the edge of the USGS quadrangle map, rather than the more appropriate technique of point-to-point digitizing. Stream digitizing is appropriate when entering large numbers of lines with curved boundaries. In such a case, the operator simply has to trace the curved lines, and the computer takes over the job of pressing the enter button for thousands of points.

Another important consideration is the size of the polygon that will be entered into the GIS. Once again, the minimum resolution of the data layer comes into play. Remember that in Section 0, we discussed that the minimum feature size should be twice that of the resolution. This guideline gives us an idea of which polygons should be included or ignored during the digitizing process. However, other considerations, such as the area of the polygon should come into play as well. Take, for example, the polygon representing a river on a topographic map at a scale of 1:20,000. Although this polygon may be thinner than a polygon that we would normally ignore, it is a relatively important feature on a topographic map, and because it is so long, its area is very great. In such a case, it makes sense to include the river. This is an example of the type of decision where a GIS operator has to apply his or her experience to the problem at hand, rather than blindly following digitizing rules that have been set out for a project.

Raster Data Collection

Unlike the collection of vector data, which requires many rules and procedures to ensure the collection of quality data, raster data tends to come to the GIS operator in a more readily usable form. The reason for this is simply because vector data are being collected from scratch whereas raster data have typically been processed to a high degree.

DeMers (2005) points out that the digital numbers in remote sensing images are rarely appropriate for direct input into GIS. Only if these images are being used as a raster backdrop for vector data can digital numbers be used without classification. In all other cases, the digital numbers need to be converted to some sort of classification, whether this is performed by a Remote Sensing Specialist, or by a GIS technician.

A Remote Sensing Specialist would produce a supervised or unsupervised classification of the image using the advanced tools found in image processing systems. This preprocessing eliminates a great number of the decisions that need to be made for vector data. Once the classification is complete, a raster containing the classifications (nominal or ordinal data) can be provided for input into the GIS.

Many GIS offer rudimentary tools for the classification of remotely sensed images. Even when satellite imagery is not classified in an image processing system, the data have been preprocessed by the agency that manages the satellite, and many corrections have been made in order to make the satellite data ready for public consumption. For example, errors produced by problems with the satellite sensors and satellite motion are typically removed, to produce a scene that is geometrically and radiometrically correct. Only three jobs remain for the GIS operator: the reprojection of the satellite image into the coordinate system used by the GIS, the selection of appropriate bands for analysis, and the analysis and delineation of features shown on the satellite image.

Data Collection Standards

It is clear that many choices need to be made when setting out to collect data for a GIS. Questions of data resolution, features to be collected, methods for collection, and ways of encoding data all must be answered before data collection commences. What happens, when you have not one, but dozens or perhaps hundreds of people collecting data for the same project? In such a case, it is very important that all of the people are collecting data to the same standard to ensure data quality.

National and international organizations have defined a number of different standards for the collection of GIS data. Unfortunately, many of the standards are incompatible with each another, although all serve the important purpose of ensuring that data coming from a particular organization is of equal quality, and can be processed using the same methods. Even if data coming from two national organizations are collected using different standards, it is still relatively easy to convert the data from one national standard to another, as long as the data are consistent[2]. National standards ensure that the following four aspects of data collection are consistent:

- Data Capture Methodology
- Data Representation
- Feature and Attribute Storage
- Relationships Between Objects (Geographic Data Files, 2007)
-

Table2 following describes some national and international data collection standards, and the organizations that created them.

BS 7666	Spatial data sets for geographic referencing (UK)
CEN TC 287	European norms for geographic information (Comité Européen de Normalisation)
DIGEST	Digital Geographic Information Exchange Standards (NATO)
DNF	Digital National Framework (UK)
GDF	Geographic Data File (Europe)
ISO 6709	Standard representation of latitude, longitude and altitude (International Organization for Standardization)
ISO 8211	Specification for a data descriptive file for information interchange (International Organization for Standardization)
ISO 15046	Geographic information (International Organization for Standardization)
NTDB	National Topographic Database (Natural Resources Canada)
TRIM	Terrain Resource Information Mapping (British Columbia Ministry of Sustainable Resource Management)

Table 2 National and International Data Collection Standards

2.2.3 Storage of Spatial Data

There are two important aspects to the storage of spatial data in a GIS. The first, called physical storage, concerns the way that digital data are stored in the computer. The second, called logical storage, concerns the way that spatial data are encoded into computer data files.

Physical storage is predominantly within the realm of expertise of the Information Technology (IT) expert, however, important decisions sometimes need to be made by the managers of GIS installations about what form of physical storage should be used. Because, at this level, GIS data are no different from any other kind of digital data (except that the files tend to be larger), the IT expert can handle most questions about physical storage. The role of the GIS manager is to work with the IT expert to ensure that GIS data are always available to GIS technicians,

[2] This assumes, of course, that the equivalent features were collected at an equivalent scale.

and that the integrity of the data is not compromised. For this reason, it is important for the GIS manager to understand issues of data security, in particular the backing up of data files, and control over access to those files by appropriate personnel.

Many different options for secure storage of GIS data are available, and the choice of system depends upon the requirements of the GIS project. Depending on the importance of the GIS project, and the risk associated with the loss of data, a GIS manager may opt from a simple weekly backup of the personal computer hard disk to a thumb drive, to multiple, automatic redundant backups distributed to a number of secure sites. The most advanced systems, such as RAID (Redundant Array of Inexpensive Disks) drives, are able to restore lost data instantaneously and automatically, usually without the GIS manager or operator noticing. Off-site storage may be accomplished by the use of online storage systems, or by the physical transportation of backup tapes to off-site locations.

Physical storage media, such as computer tapes, flash ROM drives (thumb drives), CDs, and DVDs, each have particular advantages and disadvantages, including length of media life and resistance of media to accidental destruction.

Logical storage of GIS data is also the focus of the GIS manager, who is supported by software vendors. Software vendors are constantly striving to ensure that their GIS retrieve data as rapidly as possible. This is accomplished through the use of efficient data retrieval algorithms. The use of the right algorithm has the ability to increase data retrieval speeds by hundreds of times. Software vendors offer consultancy services, to help GIS managers implement their GIS in the most efficient way possible using the software provided by the vendor.

Within the GIS, several data models are available to store different types of spatial data. Three models are commonly in use at the present time. These are the *vector model*, in which discrete data are represented as lines, points, or polygons, the *raster model*, in which continuous data are represented as a series of cells in a grid, and the *object model*, in which data are represented using objects which are modeled on real-world features. In the object model, a lake might be known with an object type called "lake" which has parameters that include pH, turbidity, oxygen level, and temperature. This object would also include a method for how the object is to be represented in the GIS (i.e., vector or raster representation). So the object-oriented model overrides the raster and vector models, and provides additional information about the objects that are represented in the GIS.

For the GIS Manager, important decisions about how GIS data are structured need to be made early in the design of a GIS, and these decisions will have long-lasting implications for the success of the GIS. The reason that these decisions are so important is because it is relatively easy for poorly structured GIS data to overwhelm the capabilities of even the fastest computers.

Consider, for example, a large GIS that contains spatial data for an entire country. Most decisions that are made with such GIS are for a subset of the entire country, perhaps a province, or a small region. In such a system, it makes sense to divide the GIS into a series of "tiles," which can be loaded and unloaded as required. For regional analysis, the data from only a few tiles is required. If the entire country must be analyzed, the GIS can perform the analysis on a tile-by-tile basis, until all of the tiles have been analyzed to produce the final result. This is sufficient for many, but not all types of analysis. For some types spatial analysis, it may be necessary to load large amounts data into computer memory.

Every feature that is loaded takes up space in computer memory. At some point, the computer runs out of RAM, and begins to use "virtual memory," which allows the memory in RAM to be swapped with records on the hard disk. Unfortunately, hard disks are thousands of times slower than RAM. This means that when virtual memory is used, the speed of the system begins to decline. An ideal GIS has sufficient amounts of RAM for most analyses, and uses virtual memory only for the most complex of analyses, ensuring that the GIS operates relatively efficiently.

The use of tiles also allows data to be edited in a piecemeal fashion. An individual GIS technician can "check out" a particular tile for editing, and when complete, the tile can be "checked in." The check-in process involves a supervisor verifying the data that has been changed to ensure that no errors are inadvertently introduced into the GIS database. This process also divides the GIS database into portions, which ensures, that in a worst-case scenario, the entire GIS database does not become corrupted.

Of course, there is overhead associated with the process of dividing the GIS data set into tiles, and allowing users to check-in and check-out data. For example, a database of user names, passwords, and security clearances needs to be maintained. Software used to manage a check-in and check-out process, ensuring that only one user is able to check out a particular set of data at one time, and ensuring that if the user corrupts the data, a backup copy can be retrieved. Finally, when the data are checked in, the system must perform an edge-matching task to ensure that the lines entered by the technician are adjusted to join correctly with lines on adjacent tiles.

The checkout and check-in process also allows data security to be imposed. Only those users with the correct credentials are able to check out data and modify it. Other users may be able to view and analyze the entire data set as a unit, but this activity may be restricted to GIS managers.

Another way of segmenting a large GIS database is on a theme-by-theme basis. This is another way of improving GIS performance and ensuring data security. In a complex GIS, there may be hundreds of separate themes, and downloading all of them would quickly overwhelm an operator's workstation. By downloading only the themes of interest, the speed of editing and updating data can be increased, while at the same time ensuring that data remains secure. Because only certain themes may be checked out and edited by particular technicians, sensitive data can be restricted to people with the correct security clearances.

The combination of tile and theme-based distribution and editing of data allows large GIS to operate relatively effectively, performing large queries, and storing enormous volumes of data in multiple themes and tiles. Data security is ensured by granting only users with the correct credentials the right to access or edit data.

2.2.4 Analysis of Spatial Data

In general, the analysis of spatial data in a GIS consists of the creation of new layers based on existing layers, or the combination of layers using some sort of overlay procedure. However, the specifics of spatial analysis depend on the data model that is used internally by the GIS.

In the vector model, points, lines, and polygons must be combined during overlays. This involves having the software determine the mathematical intersection between these feature types, and producing the appropriate result. Overlays are based on the combination of layers and the application of Boolean logic (union, intersection, erase etc.) to produce the desired result. New layers can be created from existing layers through the selection of the subset of features, the merging of a set of features, or the application of a geometrical function, such as a buffer, to a set of features.

In the raster model, new layers can be created through the selection of particular pixel values, or the application of filters, masks, or mathematical models such as distance functions.

Multiple layers can be combined mathematically using algebraic notation (e.g., layer3 = layer2 + layer1).

The analysis of the vector and raster data will be discussed more thoroughly in the Module 4 of this course.

2.2.5 Creation of Output Products

Although maps were the first output products generated by GIS, many other types of output products have been created since. The flexibility of the GIS allows data to be manipulated in many different ways using computer programs, to create multiple output products.

Paper maps produced by GIS were once crude approximations of the fine work that could be produced using photomechanical techniques such as scribing or engraving. For example, the first maps produced by SYMAP were produced on line printers by the overstriking of multiple characters. Later, pen plotters allowed vector graphics to be drawn on paper with reasonable accuracy. Unfortunately, although these plotters were a vast improvement over SYMAP, pens rapidly ran out of ink, maps could only be printed with the limited number of colours (typically eight, although pens could be changed if necessary), the shading of areas was a very poor, and the plotters were prone to mechanical failure. The advent of color electrostatic plotters, and later high quality inkjet plotters, enabled professional quality, full-color maps with area fills to be produced rapidly. The algorithms within modern GIS can now take advantage of all of the capabilities provided by modern plotters, and these systems can now produce better maps (employing transparency, for example) than could be produced using photomechanical techniques.

Modern plotters are very convenient and offer extremely high resolutions, but they are limited in the ways in which the data can be presented. Computer monitors, on the other hand, despite the fact that they have very low resolution, allow the display of dynamic data. This allows time-series data or three-dimensional data to be displayed in place of a conventional static paper map. Digital Terrain Models (DTMs) are three-dimensional representations of the Earth's surface, which can be rotated to be viewed from any angle, or simulated views can be created from any point on the model. It is even possible to fly through the scene to obtain a better understanding of the data.

Often, map output is not required at all. It may be more valuable to obtain a series of numbers to describe land use classes by area, for example. In cases such as these, GIS data analysis can be presented in the form of tables and charts.

2.3 Components of GIS

2.3.1 Direct GIS Input Devices

Direct GIS input devices are used to directly import map or spatial data into a GIS.

Digitizers

Digitizing tablets range from table-top units that are about 40 x 50 cm in size, to freestanding units that may be 2 x 2.5 m in size (Figure 4). Attached to these units are computer mouse-like devices called pucks, which are used to follow the lines on the map. Pucks typically have four to 16 buttons, which allow the operator to control the digitizing process. The puck uses magnetic induction to determine its location relative to wires in the underlying tablet. Accuracy values are typically about 0.001 inches (0.025 mm).

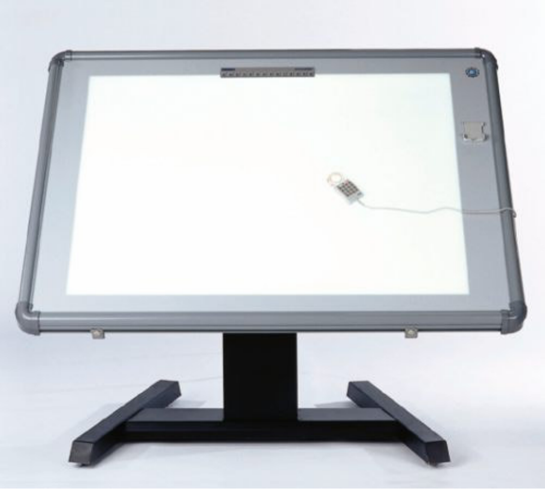


Figure 4 Pedestal-Mounted Digitizing Tablet showing 16 button puck in centre of tablet. Source: http://www.gtcocalcomp.com/photos/PHatsl500.jpg

Scanners

A scanner is a device that reads paper maps and converts them into very high-resolution bitmap files (typically in TIFF format). The files may then be used directly in the GIS as a raster backdrop, or a scanned image may be vectorized using post-processing software.

Scanners are typically divided into two groups. Flatbed scanners require the map to be laid out face down on a glass table. A scanning head containing a charge-coupled device (CCD) then moves back and forth across the face of the map, recording data. Drum scanners require the map to be attached face out to a drum, which rotates. The scanner head, which also contains a CCD, moves back-and-forth as the map rotates on the drum, to obtain data for the entire map.

Post-processing software can be used to clean up the scanned image, select lines to be vectorized, and build topology for the vectorized lines. Post-processing software may be semiautomatic, requiring a user to identify lines of interest on the scanned map, or fully automatic, converting all lines on the map or within an area of interest for later editing.

2.3.2 Indirect GIS Input Devices

Indirect GIS input devices involve the use of some form of an intermediate computer between the input device and the computer on which the GIS runs. Of course, as computers become increasingly inexpensive, all input devices are becoming indirect.

Remote Sensors

A remote sensor is any device that can obtain data about some object without coming into physical contact with that object. Your eyes are an example of a remote sensing system. In this section, we will discuss remote sensing systems that make use of electromagnetic radiation (i.e., light). Remote sensing systems may be classified as either *passive* or *active*. Passive Remote Sensing systems make use of an external source for electromagnetic radiation, which is usually the sun. Active Remote Sensing systems emit their own source of light, which is reflected off the target object to a sensor.

Passive Remote Sensors

Cameras, and most remote sensing systems, are examples of *passive* remote sensing systems. Like a digital camera, these systems require a source of energy (the Sun), a target (the object being photographed), a method of transmission (light), and a sensor (the charge-coupled device that receives the data).

Spectral imaging scanners can be classified as *multispectral*, where a few distinct spectral bands (e.g., red, green, blue, near infrared, thermal infrared) are collected, *hyperspectral*, where a few hundred spectral bands are collected, or *ultraspectral*, where thousands of bands are collected (Covey, 1999). Being able to view images in different spectral bands enables different objects to be identified. Just as you are able to identify plants by their green colour, remote sensing analysts are able to identify different objects by their relative reflectances in different spectral bands. The more spectral bands that are used, the more precisely the colour of a particular object can be examined, which leads to higher classification accuracies.

Spectral imaging scanners are also available to see outside the range of light that is received by the human eye. In particular, scanners can receive light in the near infrared and thermal infrared bands. The near infrared band is particularly useful for distinguishing between vegetation and bare soil, for example.

Active Remote Sensors

It is also possible to use RADAR sensors for remote sensing. RADAR is an example of an *active* remote sensing system. With RADAR remote sensing, microwaves are emitted from a satellite or aircraft antenna towards a target, reflect off the target, and are received once again by the antenna. The main disadvantage of an active system is that it requires a large amount of power to produce the microwaves, but it has the advantages of being able to work day or night. Microwaves also pass unhindered through cloud cover, making a RADAR system all-weather.

Another active remote sensing system is LiDAR (Light Detection and Ranging), in which the sensor determines the return time of LASER beam as it is reflected off a target. LiDAR is primarily used for collecting elevation information from aircraft, although LiDAR has been used to map the topography of Mars, as well. Airborne LiDAR has the ability to identify the elevation of different parts of the forest canopy, since some of the LASER pulses will reflect off the top of the canopy, some will reflect off the understory, and some will reflect off the ground. In urban areas, the superior resolution of the LASER pulses (typically 10 centimetres horizontal resolution) allows individual buildings to be modelled.

Because, unlike images from a digital camera, spectral imaging scanners do not collect an entire image at once, images must be pre-processed in order to make them geometrically correct. Images may be processed using an image processing system to identify individual features, such as particular crop types, based on the spectral information in the image.

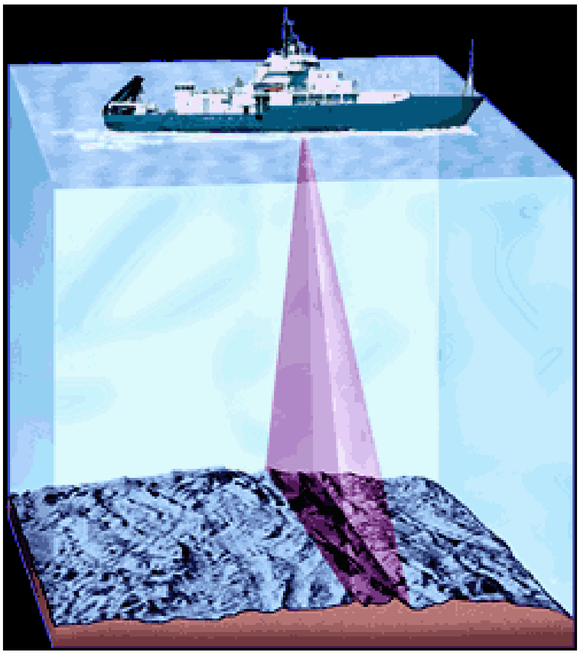
Multibeam Echosounders

Multibeam echosounders are advanced versions of depth sounders used in many recreational vessels, and are used in marine mapping applications. Whereas a recreational depth sounder uses a single acoustic beam to determine the depth of water, a multibeam echosounder uses multiple beams to map the ocean depths not only directly beneath the ship, but also to each side.

Two varieties of multibeam echosounder exist. The first transmits sound waves from a transducer, which are received by an array of microphones that are towed behind the ship. The second transmits sound waves from a transducer, and receives the reflected waves using an array of microphones that are attached to the hull of the ship.

In both cases, a number of corrections need to be made to convert the received signals into accurate depths. The first correction is related to the angle at which each pulse of sound is transmitted. Those that are transmitted straight down determine an accurate depth, but as the angle increases, the depths must be multiplied by the cosine of the deflection angle from the vertical. The next correction has to do with the attitude of the ship. The ship will generally be pitching and rolling while mapping the seafloor, so the angles, and hence the distances returned by the multibeam echosounder must be corrected. Next, corrections must be made for tidal fluctuations in order to obtain standardized depths. These corrections ensure that the mapped seafloor is correct.

The Konigsberg-Simrad EM300 is an example of a multibeam echosounder. This unit has both the transducer and the receiver microphones mounted on the ship's hull. The unit emits 135 individual beams at 30 kHz, each 1° by 1° in size. The unit is able to map from 10 to 5000 m in depth. When mapping, the ship travels at 7-10 knots, and the beams determine the depth to the bottom in a line perpendicular to the ship's track (Figure 55). A GPS unit accurately determines the ship's location, and the depth measurements are then given an X and Y location, based on their position relative to the GPS receiver. This allows large swaths of seafloor to the mapped rapidly (Figure 66).



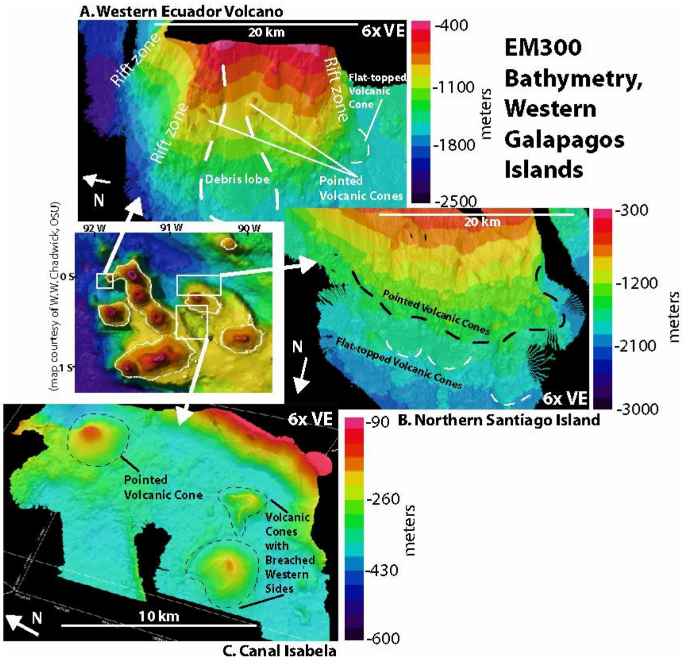


Figure 5 (Left). Fan-shaped acoustic swath formed by multiple sonar beams being used to scan the ocean floor. Source: http://www.whoi.edu/instruments/viewInstrument.do?id=15267

Figure 6 (Right). Ocean Floor mapping by multibeam echosounder, Galapagos Islands. Source: http://www.whoi.edu/instruments/viewInstrument.do?id=15267

In addition to depth, the strength of the returned beam is used to determine the hardness of the seafloor, which provides important information about its composition (Ocean Instruments).

Survey Instruments

In this section, we distinguish between traditional survey instruments, which use polar coordinates to determine the precise location of objects in a local coordinate system, and Global Positioning System Receivers, which directly determine the location of objects in a projected Cartesian coordinate system.

Survey instruments have been greatly advanced by the introduction of inexpensive lasers. Older survey instruments, such as theodolites, were able to determine angles accurately, but surveyors had to physically measure distances using a surveyor's chain, which was, in essence a tape measure made of metal that could be stretched from the instrument to the target. By combining the angles measured from the instrument with the distance measured from the chain, the surveyor was able to triangulate (or trilaterate) to determine the relative locations of all objects from any known survey marker.

With new LASER-based instruments, both the angle, and a distance can be measured by a single instrument. These instruments, called *total stations* allow surveys to be completed much more rapidly and precisely than ever before. These stations allow measurements to be stored in digital form within the instrument itself, and the data can then be downloaded to a computer at the end of each survey.

The measurements made using a polar coordinate system need to be converted into a projected Cartesian coordinate system in order to be compatible with data in a GIS. To accomplish this, Coordinate Geometry (COGO) is a technique used to convert the surveyor's notes from one coordinate system to the other.

Satellite Navigation Systems

The Global Positioning System (GPS) is a satellite radio navigation system that allows the position of the receiver to be determined within metres or centimetres. GPS uses a "constellation" of 24 satellites having precisely known positions to allow a GPS receiver to calculate its position. Each satellite measures time precisely using an atomic clock, and broadcasts a time-synchronized signal. The receiver receives the signal from four or more satellites, and calculates the time differential to each satellite, which allows the distance to each satellite to be determined within a few metres. Since the satellite locations are known precisely, the receiver can then calculate its location.

Inexpensive handheld GPS receivers require translation software to convert from the receiver's internal data format to a GIS-compatible data format. More expensive mapping computers combine the capabilities of GPS and GIS into a single handheld unit.

Data Collection Computers

Data Collection Computers are GPS-enabled handheld computers, which allow field data collectors to enter data directly in digital form. The capabilities of the handheld computers allow some processing to take place in the field; increasingly sophisticated software is beginning to transfer many GIS functions into the handheld computer (Figure 77).

Handheld computers are ruggedized versions of Personal Digital Assistants (PDAs) that are used by many office employees to keep track of appointments, phone numbers etc. Like PDAs, these systems typically run Windows CE, although other operating systems for PDAs, such as PalmOS, Symbian, or Linux may be used. The ruggedization process for these machines typically involves the addition of armor plating, rubber bump guards, and making machines water resistant or waterproof.

Using such units, it's possible not only to collect data, but also to edit it in the field. This is much simpler than finding an error once back in the office and then having to return to the data collection site to recollect data.

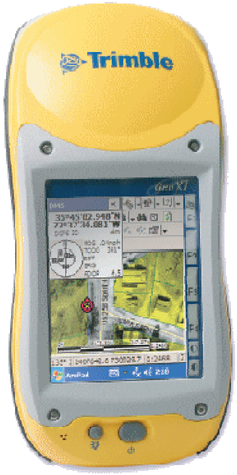


Figure 7Trimble GeoXT GPS-Equipped Data Collection Computer. Source: http://www.thewootencompany.com/edgecombe_casestudy.pdf

At the end of each data collection session, the handheld computer is connected to a more powerful computer (either a laptop, or a desktop computer at the office) and the data are downloaded and backed up. Because the data were collected digitally, and has been corrected in the field, it can in theory be directly merged with the GIS database without any human intervention.

Practical considerations however mean that at the very least, the data should be checked for consistency, completeness, and accuracy in the office. Even when working under optimal conditions, it is easy to make mistakes when performing fieldwork. Physical exertion, cold weather, and rain only contribute to the number of errors that will be made.

2.3.3 Computers for GIS

Although the personal computer is currently the most common choice for a computer on which to run a GIS, that was not always the case, and it is unlikely to be the case in the future.

As you may recall, the first GIS, such as the Canadian Geographic Information System (CGIS) ran on mainframe computers. This phase lasted for about 10 years. Later GIS were able to run on minicomputers, such as those made by Prime and Digital Equipment Corporation (PDP 11 and VAX)

By the 1980s, powerful computer workstations, such as those manufactured by Sun Microsystems came to dominate the GIS market. Workstations, which were designed to maximize the processing of scientific data, typically ran a variant of the Unix operating system, such as Solaris, AIX, or HP-UX. At the time, personal computers had relatively little power, and workstations were built with large amounts of memory, fast disk drives, multiple processors, and very fast data transfer between memory and CPU. Understandably, the high-end computer hardware used on these machines also made them five or more times more expensive than personal computers. However, despite the costs, such machines made GIS possible for many smaller companies.

In the 1990s, Personal Computers had advanced to such a state that many of the capabilities found in workstations were now available directly on the CPU. Pentium 3 and Pentium 4 CPUs were becoming extraordinarily powerful, and could be produced relatively cheaply. This led to a new era of PCs that could easily run GIS software. GIS software vendors responded to this trend by porting their software to PCs to make it available to a much wider audience. The introduction of ArcGIS 8 in 1999 meant that the PC has finally taken over from the workstation as the main platform for Geographic Information Systems.

The newest personal computers now support multiple CPUs on a single chip. For example, the new Intel Core Duo processor has two Pentium-class processors on a single chip. These chips are now available on laptop computers, allowing a laptop to run a full GIS in the field.

With multiple CPUs on a single chip, each processor is able to run its own application. This is not to be confused with parallel processing, whereby a single computing task is divided between multiple chips. Parallel processing requires special programming techniques to allow computing task to be broken down effectively into subtasks that can be handled by each processor. For large computing projects, such as some of the heavily analytical procedures that are required by GeoComputation, it is desirable to break down a task so that it can be attacked by multiple processors or multiple computers.

Clusters of computers can be built by connecting a large number of personal computers together, such as in a (Linux) Beowulf cluster, or individual computers can be distributed across the Internet, and portions of work can be assigned to each computer for later reintegration. This is the model followed by the successful Berkeley Open Infrastructure for Network Computing (BOINC), which now runs many large computing initiatives, such as the BBC Climate Change Model, or SETI@home (Search for Extraterrestrial Intelligence at Home). Through the computing resources provided by thousands of volunteers who allow BOINC to run on their computers when they are not being used for other purposes, these projects can churn through billions of calculations each day. Within a few months of the release of SETI@home, the networked computers of thousands of users worldwide became the world's largest supercomputer.

2.3.4 Output Devices

Computer Monitors and Graphics Cards

The most common GIS output device is the computer monitor. Computer monitors have the ability to display dynamic data in almost any colour. Depending on the amount of use a computer monitor will receive, a number of different options are available.

Until recently, the cathode ray tube (CRT) monitor was the only choice available. CRT monitors were available in every size from 4 inches to 21 or more inches diagonal size. Most CRT monitors used in office environments were 14 or 15 inches in size. Larger monitors were built for people who did extensive work with computer graphics, and were available in up a 21-inch size. Modern CRTs featured (relatively) flat screens, high brightness, and easy viewability from any angle. Unfortunately, CRTs tend to be very large, are very heavy, and take up a great deal of desk space.

LCD monitors have recently become widely available, and feature large display sizes that do not take up as much desk space because the monitors are nearly flat. The main disadvantage to LCD monitors is that, depending on the technology used, they are not viewable from a wide variety of angles. LCD monitors are only visible from 60 to 85 degrees off-centre, depending on the type of monitor. LCD monitors are available in the same sizes as CRT monitors. One important factor to consider when purchasing LCD monitors is that they feature a native (maximum) resolution, at which they work best. For technical work, ensure that the native resolution is as high as possible. Although the resolution can be stepped down (with reduced quality), increasing the resolution beyond the native resolution is not possible.

Most modern computer graphics cards support two monitors. A "dual head" workstation is often quite useful for technicians performing production GIS work, since it minimizes the amount of window manipulation that needs to be performed to see everything that is going on. Some graphics card manufacturers now produce graphics cards that support three monitors, which allow a main monitor to be used in the centre, and subsidiary monitors to be placed on each side, for a "wraparound" effect. Placed side-by-side, the user has the experience of a work area that is 3840 pixels wide (3 x 1280) by 1024 pixels high.

Plotters

Current inkjet plotters offer very high resolutions and quality colour reproduction for static displays. These plotters have completely replaced the pen plotters of yesterday, and offer reliable, high-resolution, low maintenance colour plotting. These plotters usually make use of cyan, magenta, yellow, and black ink cartridges, which can be combined to produce any color in the spectrum. In newer models, the print head has been separated from the ink reservoir, to allow the plotter to print many plots before the reservoir needs to be replaced. The plotter is able to print at 2400 x 1200 dots per inch.

Plots are produced on large rolls of paper (typically 36 or 48 inches wide) that are wide enough to accommodate ISO A0, B0, or C0 plots.

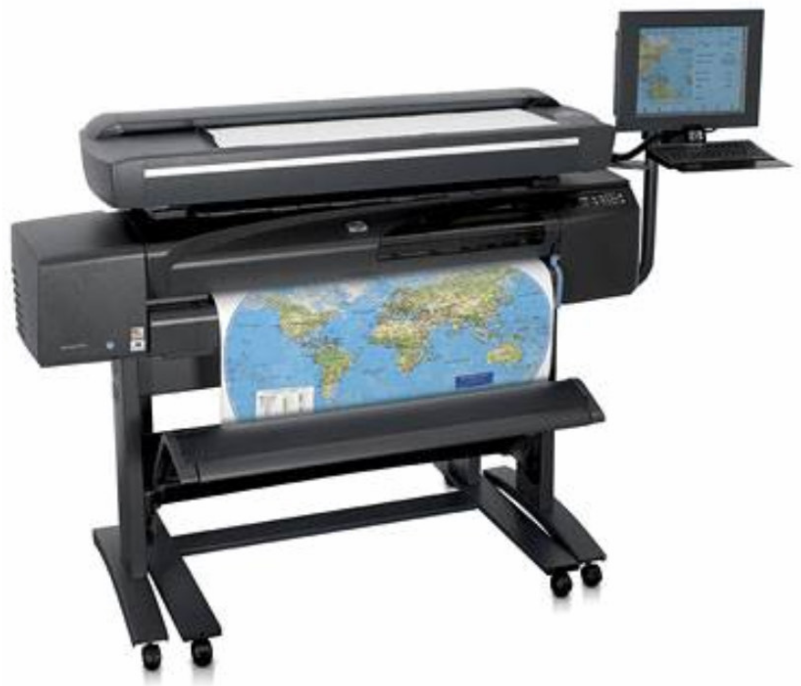


Figure 8Hewlett-Packard 820mfp Scanner/Plotter Combination. Source: http://h71016.www7.hp.com/ctoBases.asp?oi=E9CED&BEID=19701&SBLID=&ProductLineId=503&FamilyId =2367&LowBaseId=7088&LowPrice=$1,288.00#Three-Dimensional Printers

Using some of the principles developed for inkjet printers, Z Corporation has developed a line of fully three-dimensional printers that can be used to rapidly create full-colour models from CAD (Computer Aided Design) systems and GIS. These printers work by using a modified inkjet print head to spray layers of adhesive onto thin layers of plaster. By building up a model with successive layers, it is possible to create a complete color 3-D model (Figure 99).

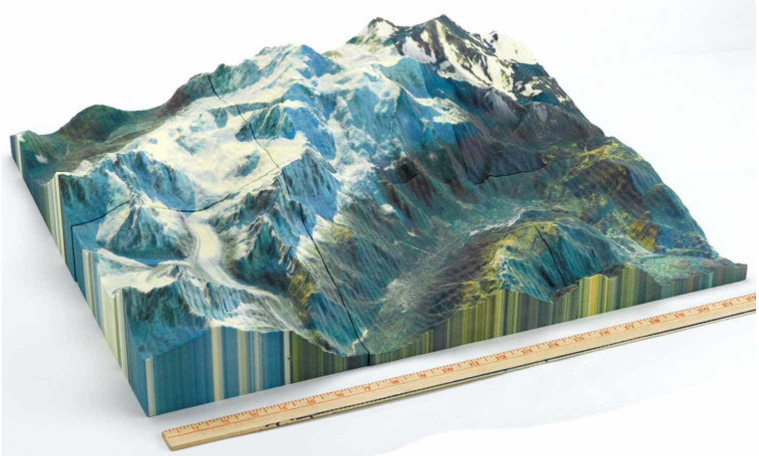


Figure 9. Three-Dimensional Model of a Digital Terrain Model. Source: http://www.ems-usa.com/sample_parts_architectural.html

2.3.5 Data Communications

The Internet has rapidly taken over the job of communicating data; it is hard to remember that only a few years ago all data had to be written to some sort of media and had to be couriered to its destination. Magnetic tape of some form was usually the medium of choice, but later, compact discs (CDs) and digital video discs (DVDs) could also be used. People in remote locations, who still do not have access to the Internet, must still rely on these time-tested methods.

However, everybody with access to the Internet now exchanges files using e-mail (for small files) or File Transfer Protocol (FTP) for large files. Only for the largest files, greater than several hundred megabytes in size, does it make any sense to write them to physical media and courier them if a broadband Internet connection is available.

2.3.6 People

People are the key component of GIS, without which, nothing else would work. People are also the most complex component in the GIS, arguably the component most prone to failure, and the only component in the system that is self-correcting and able to repair other components.

People operate both as individuals and within groups. In this section we will discuss the three roles played by individuals in a GIS, and the role of groups in shaping the environments within which GIS operate. Individuals may function as the Managers of GIS systems, as GIS Analysts, as GIS Technicians, or as users of GIS data. Organizations work to build GIS software and to ensure that different GIS are able to work together.

Roles of GIS Workers

The role of the GIS Manager is to set objectives for the GIS project, choose technologies used in the GIS, ensure that the project proceeds smoothly and on budget, and to make sure that the GIS meets the objectives that have been set. Although the GIS Manager may not be familiar with all of the details of the project, it is his or her role to focus on project-wide issues, such as how the GIS should be integrated into the larger corporation, how the sponsors of the project can be supported, and how to ensure that the GIS data are backed up in case of disaster. The GIS Manager has a role in the hiring and assignment of staff, and works to ensure that team members do their best and are as productive as possible. The manager often takes on the role of cheerleader and visionary, guiding the project and making sure that the successes of the team are brought to the attention of the people who sponsored the project.

For the GIS Manager, who must guide the project and be accountable for its success, some important skills include a basic understanding of Accounting, the ability to convey information effectively through public speaking and in print, interpersonal skills, a strong understanding of the role of GIS in the organization, and an understanding of the role of the organization itself. Technical skills are also important, but more in a general sense then in being able to solve particular programming or GIS analysis problems.

The GIS Analyst has the role of ensuring that the GIS solves the objectives set by the GIS Manager the most efficient and cost-effective way possible. The GIS Analyst typically has a number of years experience applying the GIS to different types of problems in different environments. This person should be an experienced programmer, so that they are able to automate processes to make the GIS more efficient. The GIS Analyst has a full understanding of all the details of the GIS operation, but does not generally focus on the managerial and administrative aspects of how the GIS fits into the larger organization.

The key skill of the GIS Analyst is the ability to shape the GIS to solve problems by programming new applications. Obviously, to be able to program the GIS effectively, the GIS Analyst must have a strong understanding of all the GIS functionality as well. As the senior technical member of the GIS team, the GIS Analyst should have well-developed troubleshooting skills. The GIS Analyst need not be a specialist in all subdisciplines of GIS, such as Digital Terrain Modeling, Geocoding, or Raster Analysis; individual GIS Technicians should be hired so that each is a specialist in one or more subdisciplines that are required for the GIS project.

GIS Technicians perform the day-to-day activities involved in the operation of the GIS. The activities may include data entry, data conversion from different file formats, data correction, data analysis, and map production. Typically, these people are not heavily involved in programming the GIS, but in their day-to-day duties, they may make use of programs developed by the GIS Analyst. GIS Technicians need to have a general understanding of how GIS works, as well as one or more specialties in which they are intimately familiar.

GIS Users make use of data stored in the GIS for scientific, technical, and policy analysis. These people may request data in the form of maps, they may work directly on the GIS, or they may use an Internet Portal to access GIS data from their desktop computers.

GIS Users should have a basic understanding of what GIS can do, and how to operate basic office software. It is the job of the GIS Analyst and GIS Manager to ensure that the GIS User is able to access GIS data effectively, using customized User Interface tools.

Certification

Now that GIS is maturing as a discipline, with increased standardization between different GIS products, a movement to certify GIS Professionals is now steadily gaining ground. This is somewhat of a controversial topic, because GIS technology is still evolving, and there is relatively little agreement on what technical skills are most important. To take an extreme example, an Open Source GIS Programmer might be very skilled and knowledgeable about one aspect of GIS, and yet be of little use in a facility that makes use of a particular brand of Commercial GIS.

Unlike licensing requirements for engineers or surveyors, the current certification programs are entirely voluntary. Certification helps to ensure that important skills are maintained, that ethical standards are upheld, and that standards of professional practice continued to be met. However, GIS Certification lacks the legal power to ban individuals from practising GIS if they should violate professional principles.

In the United States, the American Society of Photogrammetry and Remote Sensing (http://www.asprs.org) and the GIS Certification Institute (http://www.gisci.org) now both offer certification programs. The ASPRS Certified Mapping Scientist, GIS/LIS determines an applicant's knowledge of GIS, Photogrammetry, Remote Sensing, Earth Science, and Physics through an exam and an evaluation of the applicant's professional experience. The GIS Certification Institute assesses an applicant's knowledge through a review of his or her application portfolio. GISCI argues that the discipline of GIS is not yet mature enough for there to be a standard body of knowledge shared by all GIS Professionals. Ongoing credits must be obtained to ensure that GISCI-certified Professionals continue to make contributions to the field of GIS.

Organizations

In a field as broad as GIS, there are many activities that need to be coordinated. Not surprisingly, there are many GIS organizations throughout Europe and the world. In addition to the certification of GIS Professionals, which was discussed above, GIS organizations serve the following roles:

- Dissemination of Knowledge
- Conferences/Meetings
- Publications

- Development of Standards

The dissemination of knowledge is accomplished through professional organizations such as the Urban and Regional Information Systems Association (URISA) and the American Society of Photogrammetry and Remote Sensing (ASPRS). URISA focuses on the use of GIS for municipal management, whereas ASPRS focuses more on the use of GIS and Photogrammetry in the Earth Sciences. Both organizations have national and regional conferences, and publish periodicals and books.

The development of GIS standards is handled by national and international organizations. National organizations, such as the Ordnance Survey in Britain, establish national standards for mapping, GIS use, and data exchange.

Europe-wide, the European Umbrella Organization for Geographic Information (EUROGI) works to unite national GIS organizations to promote the use of Geographic Information throughout Europe. The group is non-profit, non-governmental and independent, and represents the GIS industry as a whole (http://www.eurogi.org/).

The European Community is also working toward the interoperability of European GIS organizations through its INSPIRE (Infrastructure for Spatial Information in the European Community http://www.ec-gis.org/inspire/) initiative, which aims to make national datasets easier to locate, access, and read, through the creation of Internet portals, conversion of data to standard formats, and the production of standardized metadata to describe the GIS data sets.

International organizations are also working towards standardization of GIS data. The Global Spatial Data Infrastructure Association (GSDI) is an independent group promoting the development of national spatial data infrastructures (http://www.gsdi.org)

The Open Geospatial Consortium (OGC) is developing a set of common standards for the transfer of GIS data between systems. Their most notable success has been the Geography Markup Language (GML), an XML-based language for transmitting geographical data. The OGC has many other initiatives relating to the discovery and transmission of XML-based files (http://www.opengeospatial.org/).

Finally, the International Organization for Standardization (ISO), is aiming towards standardizing geographic information and GIS activities through Technical Committee 211 (TC 211), and related ISO standards (6709, 8211, 15046) (http://www.iso.org/).

2.3.7 Software

There are two common forms of GIS software available today. GIS software has traditionally been commercially produced, but in recent years, a very strong Open Source movement has developed. Commercial GIS software remains dominant in the industry today, because it has a multi-decade lead in developing technology, and because the commercial model allows for sustained development of challenging computer algorithms over a period of years. Commercial GIS remains the most common choice for large GIS projects, because the software is well developed and generally reliable.

A number of Commercial GIS vendors dominate the market today (number are as of 2001). These include Environmental Systems Research Institute (ESRI), with 35% of the market, Intergraph with 13%, Leica (Mapinfo) with 12%, Autodesk (Autocad) with 7% and GE Network Solutions (Smallworld), also with 7% of the market (Howarth, 2004).

A large number of Open Source GIS projects are also available. These tend to be small, problem-specific projects, which rarely extend beyond the storage and display of data, although a few more comprehensive packages are available as well. Mitchell, 2005 provides a comprehensive list of Open Source GIS projects, and a more in-depth analysis of this sector can be found in Ramsay, 2006.

2.3.8 GIS Data

We have already discussed the disciplines in GIScience, which generate data for GIS, but we have not discussed the mechanisms by which these disciplines are able to transfer their data into Geographic Information Systems.

There are many different data exchange formats available for the exchange of GIS data. Part of the reason for the existence of so many standards has to do with the period of rapid technological advancement in the 1980's and 1990's. When GIScience was in its infancy, fundamentally new technologies were being developed on an annual basis. In order to transfer new types of spatial data, entirely new data transfer standards had to be developed.

Unfortunately, because a great deal of data had already been archived using the old formats, and because organizations were reluctant to convert old records to new data formats, the old standards were never formally abandoned, and instead began a long period of transition to obsolescence.

There are currently many proprietary data standards in use. Well-defined data exchange standards that are not tied to a particular GIS vendor facilitate the use of GIS by multiple agencies, both public and private.

Table 2 contains some examples of non-proprietary GIS standards that are available.

Format	Data Type	Description (Organization Responsible)
CCOGIF	Vector Data	Canadian Council on Geomatics Interchange Format (Canadian Council on Geomatics)
GeoTIFF	Raster data	Geographic Tagged Image Format File (GeoTIFF consortium)
GML	Vector Data	Geography Markup Language (XML based) (Open Geospatial Consortium)
NEN 1878	Vector Data	Netherlands Transfer Standard for Geographic Information (Netherlands Normalization Institute)
NTF	Vector Data	National Transfer Format (UK Ordnance Survey)
RINEX	Vector Data (GPS)	Receiver Independent Exchange Format (Astronomical Institute, University of Berne, Switzerland)
Format	**Data Type**	**Description (Organization Responsible)**
SDTS	Vector Data	Spatial Data Transfer Standard (American National Standards Institute)
TIGER	Vector Data	Topologically Integrated Geographic Encoding and Referencing System (U.S. Bureau of the Census)

Table 2. Non-proprietary GIS data exchange standards

There are also many proprietary data exchange standards that are in common use (Table 3). These are a reflection of the popularity of particular brands of software. There are many more proprietary data formats available through different GIS vendors.

Format	Data Type	Description (Company Responsible)

DXF/DWG	Vector Data	Data Exchange Format (DXF is ASCII, DWG is binary) (Autocad)
IMG	Raster Data	Imagine raster file (ERDAS - Inexpensive handheld GPS receivers require translation software to convert from the receiver's internal data format to a GIS-compatible data format. More expensive mapping computers combine the capabilities of GPS and GIS into a single handheld unit. Leica)
Coverage	Vector Data	Arc/Info file format (ESRI)
E00	Vector Data	Arc Export Format (compressed coverage) (ESRI)
SHP (+ SHX, DBF)	Vector Data	Shape Files (non-topological) (ESRI)
IGDS	Vector Data	Intergraph Graphics Design System -- file stored in Intergraph Standard File Format (Intergraph)
DGN	Vector Data	Design File (successor to IGDS) (Intergraph)

Table 3. Proprietary Data Exchange Standards in Common Use.

2.3.9 Applications

A GIS without an application is a solution without a problem. Most GIS projects are centered on one or more key pieces of functionality that GIS does well. In this section, we examine those things that make it worthwhile to purchase and implement a GIS.

Cartography: GIS are now able to produce very high quality maps; these maps match or exceed the cartographic quality of maps produced using traditional processes. Standard cartographic tools are available in many GIS, however, for the production of entirely new cartographic representations, or non-spatial data representations, such as cartograms, GIS data or maps need to be exported to an illustration tool, such as Photoshop.

Spatial Queries: One of the simplest, most widely used applications of GIS is simply to retrieve data from a spatial database. By clicking on a map, the GIS user is able to access the underlying data from which the map was generated.

Spatial Analysis: Spatial Analysis involves the manipulation of existing GIS layers to create new layers, or the combination of two or more layers to produce a desired result. This was discussed in Section 2.2.4, and will be examined in greater detail in Module 4.

Location/Allocation Problems: GIS can determine optimal locations for facilities (location), or which people should be assigned to which facility (allocation). For example, based on the distribution of elementary school students, we might use GIS to determine the best location for a new school. Once that school has been built, we will need to determine which students go to the new school, and which students go to neighbouring schools. Travel time to each school can be based on Euclidean distance (straight line), Manhattan distance (regular grid), or they can be based on distances along an irregular road network.

Geocoding: One common problem in many organizations is the determination of the location of particular addresses within a city. Geocoding allows the GIS to determine the approximate position of an address, based on address ranges for each road segment in a road theme. To accomplish this, the GIS interpolates the location of the address, based on the length of the road segment. The roads need not be straight, since the GIS uses the actual length of the road segment in its calculations.

For example, a building located at 542 Speth St. is located in the 500 block of Speth St. (this simply requires a matching of attributes). From our database, we know that the 500 block of Speth St. begins at building number 501 and ends at building number 600. Odd number addresses occur on the left side of the road (when we are traveling in the direction of increasing addresses), and even number addresses occur on the right. From this information, the GIS determines that the address can be found on the right-hand side of the street about 41% of the way between the beginning and the end of the block.

Routing: With a little more work, we can build our layer of streets into a topological network, in which we can model the movement of vehicles in one place to another. Intersections between roads can be modeled so that there are time delays set up for traveling straight through the intersection, for turning left, or for turning right. Temporary obstacles, such as road closures or vehicle accidents can also be modeled, so that routes that avoid these obstacles are chosen.

The most obvious application for GIS routing is for guiding emergency vehicles. Combined with the Geocoding and Allocation applications discussed previously, an emergency operator can enter the address of a house with a medical emergency, determine the nearest ambulance station, dispatch an ambulance to the location, and predict the amount of time that it will take the ambulance to arrive. The route from the house to the nearest hospital can then be modelled in a similar fashion.

Digital Terrain Models (DTMs): Digital terrain models are methods for representing topography in a GIS. These models may be used not only for display, but also for analysis, since physical phenomena, such as the downhill flow of water, solar illumination, and slope steepness can be modeled. Two different types of digital terrain models exist, the Digital Elevation Model (DEM), which is a raster of elevation values, and the Triangulated Irregular Network (TIN), which uses vector techniques to divide the surface into a series of irregular triangles, each representing an area of approximately uniform slope.

DTMs can be used to display the effects of physical changes on the landscape, such as landslides, or the logging of an area. DTMs can also be used for ecological studies, since the movement of animals upslope is more difficult than downslope movement, the amount of vegetation on a slope is determined by the amount of solar illumination on that slope, and elevation and slope aspect together determine the amount of snow pack. All these factors can be modeled using a DTM.

Path Functions: These are raster functions, which can be used to determine the cost of moving across the surface. Difficulties in moving up and down slope can be modeled by incorporating a DTM; difficulties in moving across particular types of terrain can be modeled by incorporating a grid showing travel cost, and uniform physical phenomena, such as winds can also be modeled to produce sophisticated models of overland travel.

One application of a Path Function is the prediction of forest fire travel over a surface. A digital terrain model represents the surface (fire is one of the only things in nature that travels uphill more easily than downhill), a separate grid showing inflammability by class of forest cover is used to model the cost, and the winds on the day of the fire control the direction in which the fire travels. The result is a model of forest fire movement over time from an initial source.

Geostatistical Analysis: Many different functions are available to determine the spatial distribution of points, lines, or polygons on a map. For example, Geostatistical Analysis can determine whether point locations or attribute values in a polygon map are clustered, random, or dispersed. For lines, the mean direction and length of the lines can be calculated, and for points the mean centre and standard distance can be determined and plotted. These tools are very handy for scientific analysis of spatial data.

Hydrological Analysis: Because a DTM is able to model the slope and aspect of every piece of terrain, it can be used as a tool for hydrological analysis. By modeling precipitation onto a DTM, the amount and direction of downhill flow can be modeled over the entire model. At some locations, all water flows downhill away from the point. These locations can be connected to create polygons showing watershed boundaries. At other locations, all water flow is inward, since most surrounding surfaces are uphill. These locations represent streams and rivers, and the amount of water coming from the uphill direction determines their flow. Thus, it is possible to predict with some degree of accuracy the amount of water flow at a particular point and time, given particular patterns of rainfall.

Summary

In this module, we have discussed the nature of GIS in a fair bit of detail, looking at the many different ways that people have defined GIS. Because GIS is inherently multidisciplinary, people from many different backgrounds have come up with their own definitions of what GIS is, based on their own perspectives. None of these perspectives is wrong, but none is completely right either. We have discussed the roles the people play in operating a GIS and the skills that they must bring forward to make a GIS project successful. The definitions of GIS that these people bring into a GIS project are important, because they influence how the people react to the challenges that occur in the GIS project. A good balance of perspectives may be just as important as a good set of skills in ensuring that a GIS project is successful.

GIS is the centre of Geographic Information Science. It is the place where most analysis is done, and it has a strong influence on what happens in other GIScience disciplines. Conversely, GIS would not be what it is today without other GIScience disciplines, which have created enormous amounts of spatial data that needs to be processed. It was this explosion of spatial data that turned GIS from a highly specialized niche area of information technology, into the dynamic, exciting field that it is today.

Finally, we examined the components of GIS, including hardware, software, people, data, procedures, and applications. As you may recall that these are the components of the Systems View of GIS, which is the most comprehensive perspective that we examined. By examining each of the components, you should now have a more thorough understanding of Geographic Information Systems.

Module Self-Study Questions

- Which of the five definitions of GIS are *top-down* and which are *bottom-up*? What difference does each approach make to the quality of the definition that results?
- Three of our GIS definitions focus on particular components described in the Systems View of GIS. Can you come up with alternative definitions that focus on people, procedures, or applications?
- Consider the synergies produced by all the disciplines within GIScience. How does GIS benefit from other disciplines such as Remote Sensing and Surveying? Do you think that GIS has changed the way that these other disciplines work?
- Describe the ways in which Global Positioning Systems are important for surveying, and the ways that they are important to GIS. Do you feel that GPS "belongs" to the Surveyors or to GIS Practitioners? Should it be a discipline on its own?
- You have been appointed as GIS Manager for a small city, and need to assemble a team to turn some vague plans into reality. Based on your own skills, what skills would you look for in a GIS Analyst? What specialties should your GIS Technicians have?
- What are the advantages and disadvantages of GIS Certification?
- Under what circumstances would you consider choosing a Commercial GIS package? When would you consider an Open Source product?
- Based on some of the GIS applications described in Section 0 can you think of a unique way to combine these capabilities to create a new GIS project?

Required Readings

- What is GIS? (http://www.gis.com/whatisgis/)
- United States Geological Survey. Geographic Information Systems (http://erg.usgs.gov/isb/pubs/gis_poster/)
- Howarth, Jeff (2004). SPACE-Spatial Perspectives on Analysis for Curriculum Enhancement (April 9, 2007) (http://www.csiss.org/SPACE/resources/gis.php?printable=true - Sector)

ESRI Virtual Campus Module

- Learning ArcGIS 9 Module 2: Creating Map Symbology
- Learning ArcGIS 9 Module 3: Referencing Data to Real Locations

Assignments

- Lab 1: Retrieval of Graphical and Attribute Data from a GIS
- Lab 2: Comparison of ArcGIS and Topographic Maps. How does a GIS Differ from Traditional Maps?

References

- Aalders, Henri J.G.L. (2000). Some Experiences with Managing Standards. (GIS Technology II:Standards and Tools for Data Exchange) (http://repository.tudelft.nl/consumption/idcplg?IdcService=GET_FILE&RevisionSelectionMethod=latestReleased&dDocName=202134) (Mar. 26, 2007)
- Arizona Electronic Atlas (http://atlas.library.arizona.edu/atlas/index.jsp?theme=NaturalResources) (Mar. 2, 2007).
- Aronoff, Stan (1989). *Geographic Information Systems: A Management Perspective.* Ottawa: WDL Publications .
- Burrough, Peter & Rachael McDonnell (1998), *Principles of Geographical Information Systems (2nd Ed.).* Oxford: Oxford University Press.
- Carkner, Larry & Egesborg, Paul (1992). "Use and Maintenance of Cadastral Data in a GIS for Canada Lands" in *Proceedings, GIS 92, the Canadian Conference on GIS, March 24-26th, 1992 Ottawa Ontario*. Ottawa: Canadian Institute of Surveying and Mapping, pp. 271-281.
- Carmack, Carmen & Jeff Tyson. "How Computer Monitors Work". Howstuffworks (Apr. 8, 2007) (http://computer.howstuffworks.com/monitor6.htm).
- Clark, Keith C. (1997), *Getting Started with Geographic Information Systems.* Upper Saddle River, New Jersey: Prentice-Hall.
- Covey, Randall J. (1999). Remote Sensing in Precision Agriculture: an Educational Primer (http://www.amesremote.com/title.htm) (Feb. 24, 2007)
- DeMers, Michael (2005). *Fundamentals of Geographic Information Systems, 3rd Ed.* Wiley.
- Donner, John (1992). "The Production of Topographic Maps Using a Totally Digital Cartographic Editing System." In *Proceedings, GIS 92, the Canadian Conference on GIS, March 24-26th, 1992 Ottawa Ontario*. Ottawa: Canadian Institute of Surveying and Mapping, pp. 71-80.
- European Umbrella Organization for Geographic Information (April 8, 2007) (http://www.eurogi.org/).
- Fuller, Alex (2005). "Edgecombe County, North Carolina, Maps and Inventories Water/Wastewater Infrastructure with GIS and GPS." *ArcNews* (Summer, 2005). Redlands, California: Environmental Systems Research Institute. (http://www.esri.com/news/arcnews/summer05articles/edgecombe-county.html) (April 7, 2007).
- Geographic Data Files. Wikipedia, The Free Encyclopedia. (April 6, 2007) Wikimedia Foundation, Inc. 10 Aug. 2004 http://en.wikipedia.org/wiki/Geographic_Data_Files
- GIS Certification Institute (April 8, 2007) (http://www.gisci.org/).
- GSDI Association (April 8, 2007) (http://www.gsdi.org).
- Howarth, Jeff (2004). SPACE - Spatial Perspectives on Analysis for Curriculum Enhancement (April 9, 2007) (http://www.csiss.org/SPACE/resources/gis.php?printable=true - Sector)
- INSPIRE Directive (April 8, 2007) (http://www.ec-gis.org/inspire).

- ISO TC 211 Newsletter 8 (April 8, 2007) (http://www.isotc211.org/Outreach/Newsletter/Newsletter_08_2005/TC_211_Newsletter_08.doc).
- Mitchell, Tyler (2005) An Introduction to Open Source Geospatial Tools (http://www.oreillynet.com/pub/a/network/2005/06/10/osgeospatial.html) (April 8, 2007).
- Ocean Instruments "Kongsberg-Simrad EM300 Multibeam Echo Sounder" http://www.whoi.edu/instruments/viewInstrument.do?id=15267 (April 9, 2007)
- Ramsay, Paul (2006). The State of Open Source GIS. Victoria, B.C.: Refractions Research (http://www.refractions.net/white_papers/oss_briefing/2006-06-OSS-Briefing.pdf) (April 8, 2007).
- Shannon, C.E., 1948. "The Mathematical Theory of Communication." *Bell System Technical Journal,* 27: 379-423, 623-656.
- Smith T.R., S. Menon, J.L. Starr, and J.E. Estes, (1987) Requirements and principles for the implementation and construction of large-scale geographic information systems. International J. of Geographical Information Systems, 1: 13-31.
- What is GIS? (http://www.gis.com/whatisgis/) (Feb. 28, 2007).

Terms Used

- Accuracy
- Active Remote Sensor
- Cartesian Coordinate System
- Check-In and Check-Out of data
- Commercial GIS Packages
- Coordinate Geometry (COGO)
- Data Collection Computer
- Database View of GIS
- Digitizer
- Echosounder
- Functional View of GIS
- Geodatabase
- Geographic Information Systems (GIS)
- Geoprocessing View of GIS
- GIS Analyst
- GIS Manager
- GIS Technician
- Global Positioning System (GPS)
- Hyperspectral Scanner
- Inkjet Plotter
- Logical Data Storage
- Map View of GIS
- Multispectral Scanner
- Object Model
- Open Source GIS Packages
- Orthophotographs
- Passive Remote Sensor
- Photogrammetry
- Physical Data Storage
- Plotter
- Polar Coordinate System
- Precision
- Raster Data Model
- Remote Sensing
- Scanner
- Satellite Navigation Systems
- Stereoplotter
- Systems View of GIS
- Three-Dimensional Printers
- Tiled Data
- Total Station
- Ultraspectral Scanner
- Vector Data Model
- Workstation Computer

3 Geographic Data

3.1 Introduction to Geographic Data

A fundamental piece of any Geographic Information System is geographic data. In fact the entire purpose of having a GIS is to store and analyze geographic data. As a GIS professional, it is very important to understand that not all spatial data are of equal value. Thanks to the Internet there are an enormous amount of spatial data available at no cost. However not all of this spatial data are equal in quality.

How do we utilize the enormous possibilities this enormous collection of data offers us? The first step is to identify the question we are trying to answer. Once we have identified a question we can then begin identifying data sources that we think will help us answer the question we have in mind. Data sources come in two forms, primary data and secondary data. Primary and secondary data sources are discussed in section 3.2.

Once we have identified our data sources the next step is to determine what type of geographic data are necessary. There are two types of geographic data, raster-based data, and vector-based data. The advantages and disadvantages of these two data types will be examined in section 3.3.

After we have identified appropriate sources of data and decided which types are required to answer our question, we need to assess the quality of the data sets we have decided to use. There are several very important indicators we need to consider when assessing data quality. These indicators include measures of accuracy and precision. We will also talk about the validity of our data. The availability of quality metadata can greatly increase data quality, and finally we need to think about error propagation and how it affects our data sets. These measures of data quality will be reviewed in section 3.4.

And finally we need to start thinking about how we are going to store our spatial data. Spatial data storage occur in a wide variety of formats and file types and we will introduce these in section 3.5.

3.2 Sources of Geographic Data

Geographic data come from a wide variety of sources. In order to conduct a meaningful geographic analysis it is important to select data sources that are appropriate to the research question at hand. A simple distinction can be made between primary data sources and secondary data sources.

3.2.1 Primary Data

Primary data are acquired directly from the source. These are data that are collected “in the field” by you or your organization. Primary data collection usually results in the creation of new data sets. Primary data come from a variety of different sources. Three examples of primary data are:

1. Surveying - considered a primary data source because the analyst is taking direct measurements of the things that they are interested in. Land surveys are one of the fundamental methods of acquiring spatial data. Land Surveys are typically entered into the GIS using COGO, which stands for Coordinate Geometry (Figure).

2. GPS mapping - also a primary data source. The analyst is on the ground collecting coordinates with a GPS unit. Once the analyst has collected a large number of coordinate points he or she can input them into the GIS to create an entirely new data set

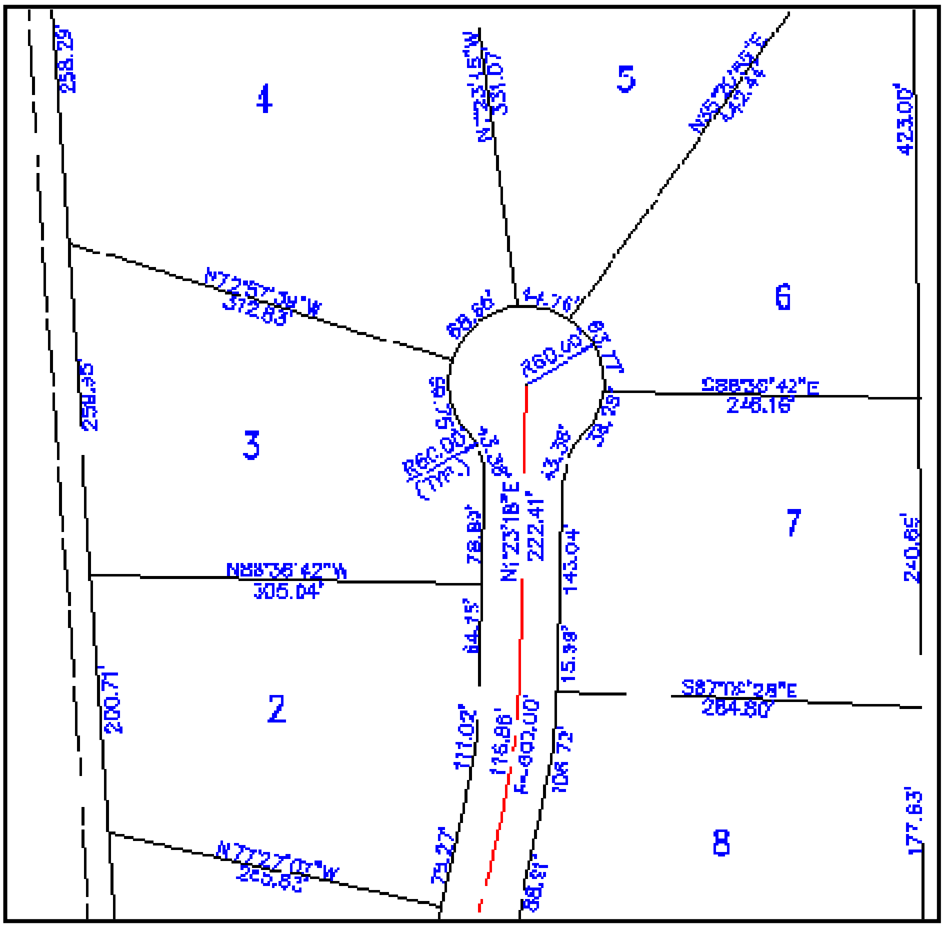


Figure1. Coordinate Geometry (COGO) is one way in which land surveys are entered into a GIS

3. Digital Cameras – the photographer in this case is also taking direct measurements. The photographer is measuring and storing light levels in order to reproduce data in the form of an image. The photographer has total control over how he shoots the image.

The first advantage of using primary data is having complete control over data collection. This ensures that the data are relevant and appropriate to the geographic question being asked. The analyst can design and implement data collection procedures, can answer questions and address concerns as they arise, or alter the design at any point in the data collection process if necessary.

The second advantage of using primary data is that the analyst has an intimate knowledge of the data. They know every aspect of the data set and can recognize and identify any data anomalies immediately. They can answer every question about the data set or are able to ask the person who collected the data.

The third major advantage of using primary data is that there are absolutely no questions regarding data accuracy. As the collector of the data, the analyst know exactly how current the data set is, can make accuracy assessments concerning measurements, and verify the accuracy of the attribute data. Data quality will be discussed in more detail in section 3.4.

However, there are also drawbacks to using primary data that must be carefully weighed against the benefits. Primary data collection requires careful thought and consideration when choosing a sample design. This requires the collector or analyst to have a basic understanding of statistical sampling techniques.

As well, the collection of primary data can often be time consuming and costly. Depending on the complexity of the project, data collection may overshadow the actual GIS analysis. Primary data often require post-processing and error checking, which only adds additional time and cost to the project.

The pros and cons must be carefully evaluated before a person makes a decision to use primary data as the foundation of their GIS analysis. There may be situations where primary data are the only option available to the analyst. In other cases, another individual or organization may already have collected appropriate data. This type of data is referred to as secondary data.

3.2.2 Secondary Data

Secondary data are a term used to describe data that have been collected by anyone other than the person doing the analysis. These data are collected from a variety of sources, which may include government agencies, private companies, or even files distributed freely over the Internet. Just like primary data, secondary data can take numerous forms. Some examples of secondary data sources are:

1. GIS Data – data that has been processed and already lives in a GIS database is considered secondary data. GIS data have a definite advantage for geographical problem solving since it is typically a simple matter to load existing data into a geographical database.

2. Remotely Sensed Data – data collected by remote sensing devices such as satellites or aerial photography cameras is also considered secondary data. Professional companies that specialize in remote sensing typically carry out data collection using these methods. The researcher then typically purchases these data. The researcher

may have some control over the area they are interested in but the actual data collection is left up the private company offering the service

3. Tabular data – tabular data such as census data or property assessment data can be very useful in performing geographic analysis. There are large numbers of tabular data sets readily available for a variety of purposes. Tabular data are considered secondary data when another individual or another organization carries out the data collection.

Again just like primary data there are many advantages to using secondary data in a GIS analysis. The biggest advantage to using secondary data is time. The data have already been collected and processed saving the researchers hours of time worrying over sampling design, and data collection. In addition to being time efficient secondary data are also generally much less expensive to acquire. The researcher doesn't need to invest a large sum of money in developing the data set. As well, other organizations may have particular expertise in collecting certain types of data.

However there are also many things one must consider before deciding to use secondary data. The researcher doesn't have the same familiarity with the data set as he would have if he had been involved in the collection process. This means that the researcher must rely heavily on the accompanying documentation. The researcher has no idea of the accuracy of the data unless it is explicitly defined. As well, it may be difficult to locate the data collector in order to determine this information. Acquiring data from organizations with a strong reputation for producing quality data will greatly improve the usefulness of the data

Another drawback to secondary data is that the data collection and sampling has almost always been designed for another purpose. Therefore, the researcher has to carefully consider whether the data they have is appropriate for the purpose they want. There is no control over what attributes are collected or stored. Data collected for one project may not be appropriate for another.

Another important consideration when evaluating secondary data is scale. The collection scale of these data must be carefully considered since it has a very large impact on the effectiveness of the analysis. Data collected for large-scale studies will not have the spatial coverage required to analyze a small-scale region. On the other hand, data collected for a small-scale study will not have the data density or spatial accuracy required for a local study.

Secondary data may also come in a variety of formats. In the best of cases secondary data will already come in a geodatabase or any one of the common file formats for storing geographical data. Data may also come in formats that will require some manipulation before they can be used. Some examples of these are JPEG files, TIFF files, PDF files, and Excel tables. In the worst-case scenario you may receive secondary data as paper maps, paper surveys, or even notes written on napkins! These forms of secondary data require a great deal of processing before they can be used, but sometimes they may be the only source.

As we can see above, there are many questions to consider when choosing appropriate data sources. If time and cost are not an issue and customized data are required, then collecting primary data may make the most sense, however if you are in a small organization with a tight budget, you may be forced to rely on secondary data sources that are close enough to your goals. With careful consideration of data sources you can greatly increase the effectiveness of your GIS projects

3.3 Types of Geographic Data

Now that we understand a little bit more about data sources and how they can affect spatial analysis, we can start to examine the data itself in more detail. However, before we start to look at how geographic data are represented in the GIS we might consider what kinds of things we might want to store in a GIS. We may be interested in the location of houses on a street, or we might be interested in storing the boundaries of a park reserve, or a city. We can also use a GIS to store things such as soil pH variations over an area, or temperature changes.

Generally, features stored in a GIS fall into two categories, discrete and continuous. Discrete and continuous data can generally be differentiated by whether or not the data are uniform within its boundaries. Discrete data stores features that have well defined, solid, unambiguous boundaries and are uniform within them. A house is a house everywhere within its walls. A city is a city everywhere within its boundaries. Therefore a house and a city in this context are discrete features.

On the other hand, continuous data varies continuously over an area. A continuous feature is not uniform within the data extent. Soil pH will vary from one place to another. Air temperatures are not consistent throughout an air mass. Both of these examples would be considered continuous data because there is an infinite amount of variation within them. Now that we understand the primary difference between these two data types we can explore the two basic ways these types of data are stored in a GIS.

3.3.1 Vector Data

The Vector data model is the most popular ways to store geographic data. This model is ideally suited for representing discrete objects. In fact, the Vector data model can only accurately represent discrete objects. The Vector data model uses points and edges to represent three basic types of spatial features: points, lines, and polygons. All of these types are capable of storing attribute data about the particular feature they represent. However these three types store their spatial data in different ways.

Point Data

Point data are data that can be represented as a single location on a map. To continue our example from above, point data can be used to represent house locations on a street. Each house is given a single X coordinate and Y coordinate, which allows the GIS to place it on a map (Figure). The GIS represents the house as a single point or a dot using the coordinate pair provided. This is by far the simplest data type and is very good for storing data when all we are concerned with is the location of a feature and not its length or width. Point data are zero dimensional and have no width, length or height.

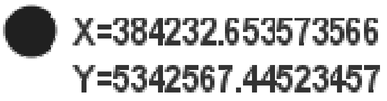


Figure 2. Point data are represented by an XY coordinate pair.

Line Data

Not all the features we'd like to store and represent on a map are going to be zero dimensional locations in space. Many of the features that we are interested in are going to have a length associated with them as well as a location. Features that have a location, a length, but no width are represented in the vector model by lines. Examples of some features well represented by lines are contours, administrative boundaries, roads, rivers, and sewers. It is important to mention that while some line features such as rivers and roads, may have an area, we use lines to represent them at scales where their width cannot be accurately reflected.

Storing line features using the vector data model is a somewhat more complicated procedure than storing point data. Line features are stored using a collection of points called nodes and vertices which each have their own unique coordinate pair. Nodes are the endpoints of a line while vertices are intermediate points located between the two end points (Figure3). Each node is connected to one line segment, while each vertex is connected to two line segments. These line segments connecting the nodes and vertices are referred to as edges. Each edge stores the ID of the two nodes or vertices that lie at its endpoints. The GIS uses all this information to correctly represent the location of the line feature. It is this collection of nodes, vertices, and edges that make up line features in the GIS vector data model.

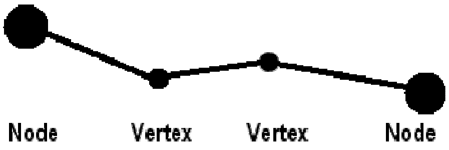


Figure 3. Line data are made up of nodes, vertices, and edges.

The vector model also allows us to join lines together to form more complicated line features (Figure 44). Collections of line features can represent more complicated real world features such as utility networks and street maps. Lines can only connect to other lines at nodes, or endpoints.

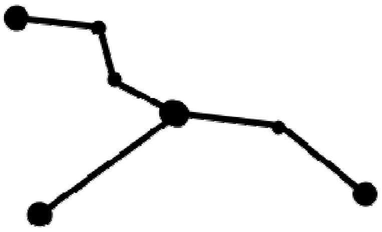


Figure 4. Lines can be connected at nodes to form networks.

Polygon Data

Many more features we may wish to represent in a GIS are going to have a width and an area associated with them (Figure 55). Examples of these are property boundaries, lakes, and political boundaries. The GIS stores polygon data much the same way it stores line data. The major difference between polygons and lines however, is that a polygon must be composed of at least one line, and must enclose an area.

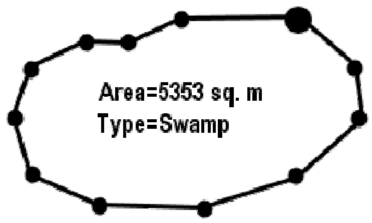


Figure 5. Polygons are formed by a line or set of lines enclosing an area.

Nodes and Vertices[3]

As we've demonstrated above the three main feature classes of the vector model are all based on collections of points with XY coordinate pairs. It is important though to spend a final moment clarifying the terminology. A node or vertex is always a point, but a point is not always a node or vertex. As shown above points may be represented as stand alone features. Points are only nodes or vertices when connected to an edge, or line segment. Nodes form the beginning and end points of lines, while vertices form the intermediate points.

We also have points that are called pseudo nodes. Pseudo nodes are nodes that only connect two line segments and can be replaced by a vertex. Not all two-line nodes are pseudo nodes however. Nodes that form the beginning and ending point of the line surrounding a polygon are necessary and therefore, are not pseudo nodes.

3.3.2 Raster Data

The second common data model used to store spatial data is the raster model. Just about every single person who has ever been exposed to a computer has seen a raster before, even though they may not have known it at the time. A raster model uses a grid of square cells to store spatial data. The most common rasters show up as images in web pages, and computer graphics. GIF, TIFF, and JPEG files, are all common formats for Raster-based images.

Raster Elements

There are many elements that are common to all rasters (Figure 66). Firstly rasters are composed of a grid of squares. Each square is referred to as a pixel, which is short for picture element. All squares are uniform in size and shape. A pixel may be any size however each pixel in the grid must be the same size.

Like vector point data, each pixel in the raster has an X coordinate and a Y coordinate. In a point vector feature the X and Y coordinates represent that point's absolute position. In the raster model however the X and Y coordinates represent the relative location of that pixel to a point of origin on the raster grid. The point of origin of the raster grid is usually the bottom left corner. The point of origin stores the grids absolute location and all pixels then become relative to that point.

[3] Mazgai ir Viršūnės

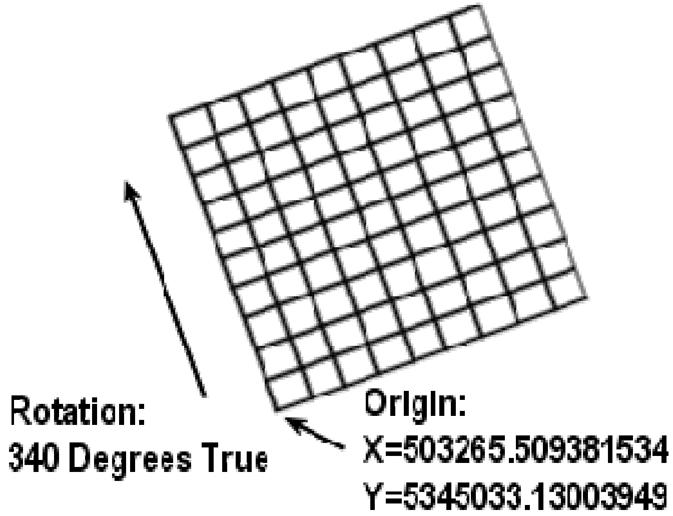


Figure 6. Basic elements of the raster model (10X by 10Y)

Another consideration of raster grids is rotation. Rotation refers to the raster's orientation to grid north. A raster does not always have to be perfectly aligned to grid north, although typically many of them are. Rasters can actually be rotated in just about any direction in order to store the data they are designed to represent. The raster's orientation is stored along with its point of origin. This allows the location of any cell in the grid to be quickly calculated based on the grid's orientation, and its relative location to the point of origin.

Each cell in the raster model is able to store a single value. Rasters are capable of storing any type of data, however the storage of nominal (kelp, blue algae, clay soils, etc.) and ratio data (2.345, 3.65) are the most common uses of a raster.

Resolution

Another important consideration when discussing rasters is resolution. The size of each pixel in the grid becomes important. The size of the pixels in the grid is referred to as spatial resolution. For example, a pixel that represents an area on the ground that is 10m x 10m has a 10-metre resolution As you would imagine, a raster with a resolution of 1m captures more detailed data than a raster with a resolution of 10m. A Landsat image, for example, has a 30m resolution, while digital aerial photography may have a resolution of 15cm or less.

Rasters with more detailed data and smaller pixels have a "higher" resolution than data sets with less detail. The Landsat images for example have a much lower resolution than the digital aerial photography in the above example. In general, higher resolution data sets are preferable to lower resolution data sets although there are some considerations to make when determining the appropriate cell size.

The first consideration when determining an appropriate resolution is the size of the features you are interested in measuring. If the features you are trying to store are very large, having a high resolution leads to an unnecessary amount of redundant data. For example if we are looking at surface temperatures of a 10km x 5km lake it makes little sense to use a 10cm resolution raster; given a choice, we would probably choose a raster with a resolution of 50m or more. However, if we were interested in measuring the surface temperatures of a 10m x 10m pond then a raster of 10 cm cells might be appropriate.

The second consideration when determining appropriate resolution for a raster dataset is the uniformity of the features being measured. For very uniform features we can increase the cell size and for very non-uniform features we should decrease the cell size. When looking at something like soil types, that are fairly uniform over large surfaces, it is possible to use a larger cell size because a smaller cell size would just store redundant data. When looking at something with more variation over a smaller area such as algae blooms in a pond we would have to use a higher resolution to capture the greater degree of variation. As a rule of thumb, we want to use a pixel size that is no larger than half the size of the smallest feature we want to capture. This ensures that no features are missed because they were too small to be represented.

The final consideration when selecting a pixel size is linked to computer processing and storage. Higher resolution images require a larger number of pixels to cover the same area. Rasters with large amounts of pixels require much more memory to store and process. Very large rasters may have so many pixels they become cumbersome to work with that they require very large amounts of computer memory.

Raster Storage

As described above one of the biggest problems facing the use of high-resolution raster data sets is the amount of memory required to store the data. To help solve this problem various types of raster encoding have been designed to help reduce the file sizes of raster data sets.

One of the simplest types of compression is called run length encoding, or RLE. RLE works by grouping pixels with the same values. RLE goes through the raster horizontal row by row looking for strings of pixels with the same value. When it finds a string of pixels with the same value it stores the value of the pixel and the number of pixels behind it. So in Figure 7, the second row might be represented as wwwwbbbwbwwwwwwwwwww, or as w4b3w1b1w11 using RLE. It becomes very apparent that RLE compression works very well in situations where we have long strings of pixels with the same value, and not so well where there is a large degree of variation from pixel to pixel (Figure 77).

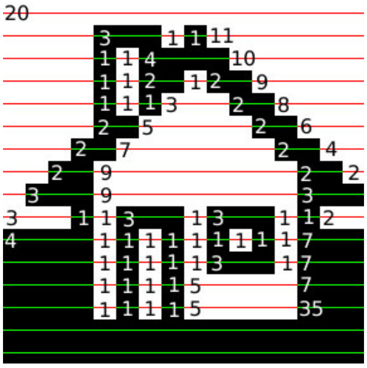
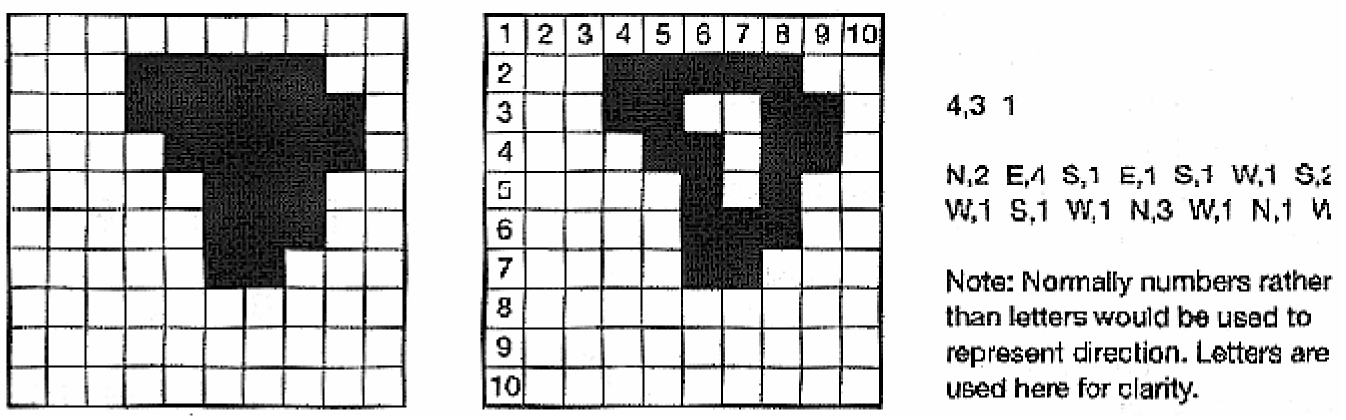


Figure 7. Run Length Encoding

A second form of raster data compression is called chain encoding. Chain encoding is similar to RLE encoding. However in chain encoding the outer boundaries of contiguous areas are stored. The location and value of the first pixel are stored and after that a set of directions are stored. The directions are stored as numbers, with each number representing the direction to the next pixel of the same value. Each successive number is stored until the pixels form a complete loop around the area (Figure 88).



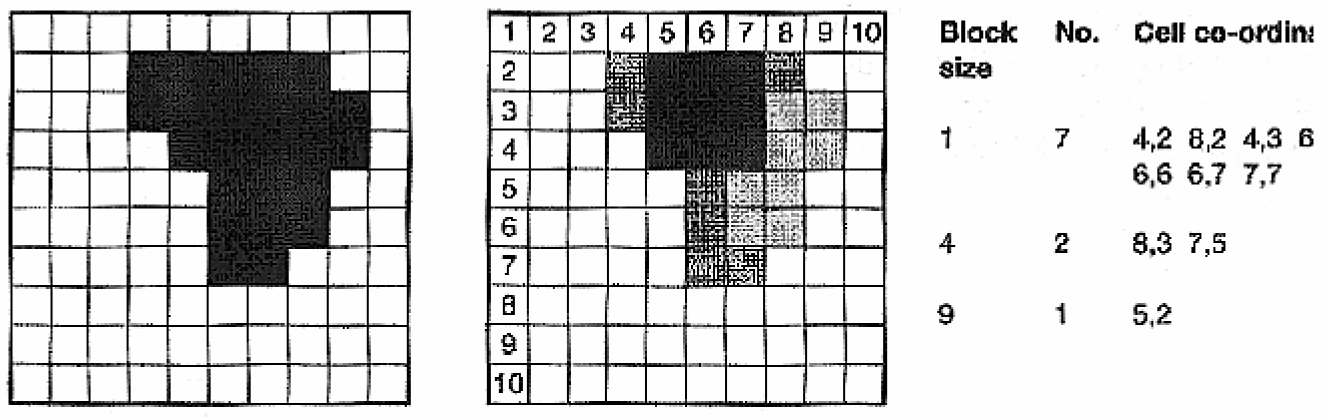
4,3 1

N,2 E,4 S,1 E,1 S,1 W,1 S,2
W,1 S,1 W,1 N,3 W,1 N,1 W

Note: Normally numbers rather than letters would be used to represent direction. Letters are used here for clarity.

Figure 8. Chain Encoding

A third approach to data compression is called block encoding. Block encoding differs from the two previous forms of encoding in that it does not store chains or rows of pixels. Block encoding breaks the raster image into square blocks with the same pixel value. The largest blocks are placed first and only the location of the origin, the size of the block, and the value of the pixels are stored. This process is repeated with successively smaller and smaller blocks until the entire raster image has been encoded (Figure 99).



Block size	No.	Cell co-ordin
1	7	4,2 8,2 4,3 6 6,6 6,7 7,7
4	2	8,3 7,5
9	1	5,2

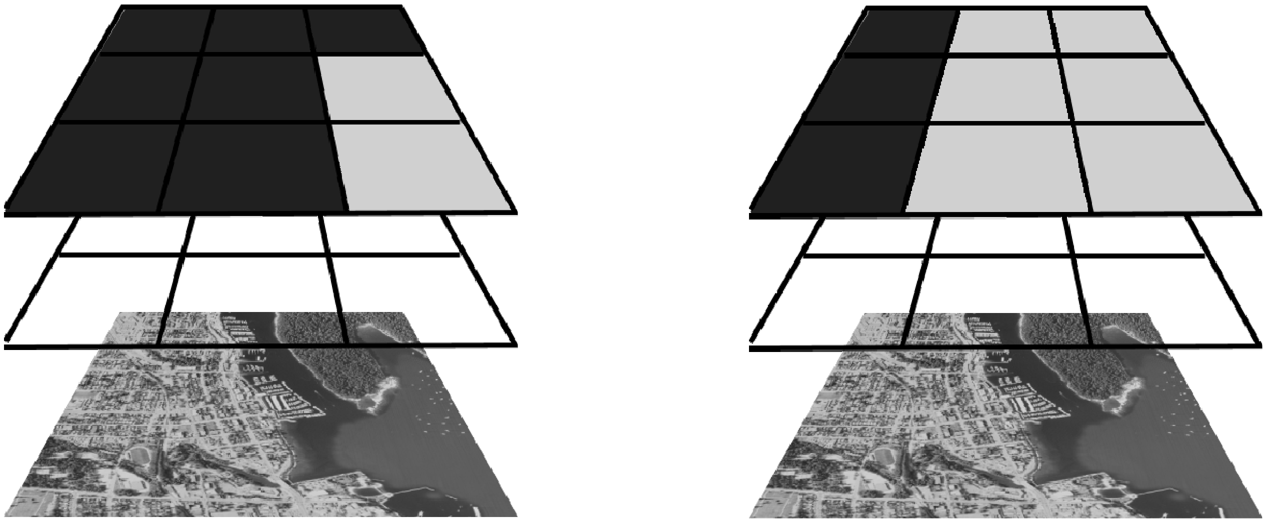
Figure 9. Block Encoding

Adding Data to a Raster

Not every feature we want to measure in the real world is going to fit nicely into a square grid of cells. In order to populate our raster grid we need to create rules to determine what falls into each cell. For many applications this choice fits very easily into the sample design. Each grid may contain a measurement of something located at the centre of that cell, or an average of several measurements made throughout that grid cell. However when measuring spatial features that do not always fit nicely into a square grid design we need to have some rules for determining what gets stored in which cell.

The rule of dominance (Figure 101010) states that whatever underlies the majority of the cell is what gets stored in the raster. For example if a farmer's field covered 75% of the grid cell then the farmer's field is the feature that is stored in that grid cell. This rule works very well for the majority of raster-based data. However, it may not always be the case that the feature that underlies the largest portion of the cell is the feature that is of most importance.

The rule of importance states that the most important feature is stored for each pixel. The highest importance feature will be stored in that cell based solely on its presence and not because it represents the largest portion of that cell. In this example the thin river that flows between the farmers fields is stored because it is the most important feature, even though it only covers 25% of the grid cell.



Rule of Dominance **Rule of Importance**

Figure 1010. Comparison of the rules of "Dominance" and "Importance". In the first Image the land that makes up the largest portion of the raster cells is stored. In the second image the water which is more "important" is stored in the raster cells.

Pros and Cons of the Vector Data Model

There are many significant advantages to using the vector data model to store geographic data. The first advantage of using a vector-based approach is that it tends to produce a much more attractive cartographic product. The vector model represents discrete features much more closely to their real life shapes. We will see why this is when we examine raster data in the next section.

Vector-based data are also much more compact to store. Vector-based data are stored as a collection of points and lines. Therefore we only store the features we are interested in storing and don't have to store any of the information found between the features. Because we are working with discrete data, we can safely assume that all data within a polygon is the same, so only one set of attributes needs to be stored for each polygon.

Another important advantage of the vector-based model is that it is much more precise than the raster-based model. Raster-based models are reliant on grid resolution. Vector-based features can be placed precisely where the feature is located. These advantages will become clearer when we discuss rasters and precision.

We can see that there are many advantages to using a vector-based data model. However, a vector-based approach may not be appropriate for every type of data. One of the disadvantages of the vector model is that we need to construct topology rules to prevent vector features from overlapping. Topology refers to the way features are arranged in relationship to one another.

In addition the vector-based model does not lend itself to spatial analysis. Many vector operations such as overlays are computer intensive to solve. Despite these issues, the vector data model remains one of the most popular ways to store spatial data.

Pros and Cons of the Raster Model

Much like the vector model, we can see that there are some uses for which the raster-based model is very well suited. In order to best utilize raster data in your analysis it is important to understand when it is appropriate and not appropriate to use raster-based data.

Raster-based data lends itself very well to spatial analysis. It is simple to overlay two grids with the same angle of orientation, grid size, resolution, and origin point. Not only is spatial analysis easier on raster-based data, many types of analysis are only possible on raster-based data. Cost surface modelling is an example of analysis that is only possible on a raster-based dataset.

As well the concepts behind the raster model are very easily understood. Storing data values in a grid of cells is very intuitive to anyone who is familiar with computer graphics.

So, if rasters are easier to understand and allow more as far as spatial analysis is concerned then why don't we use them exclusively in GIS? Well there are limitations to the raster-based data model. The first of these is size. As mentioned above raster files can store thousands of pixels. Rasters that cover a large area, and have a high resolution take up a lot of computer memory, despite the encoding techniques mentioned above. Large raster datasets can take quite a long time to analyze and load.

In addition to being very large, raster files are always an approximation. No matter how small a pixel size is used, the raster is always based on a square grid and does not exactly capture the shape of features in the real world. This is an important consideration whenever precision and accuracy are important in spatial data.

3.3.3 File Formats

Now that we've discussed the differences between raster-based data, and vector-based data we will talk about some of the most common file formats used for storing and maintaining spatial data. First we will examine raster-based file formats. Some of the most common file formats for storing raster-based files are formats that people encounter every day in the digital world. Raster files can be stored as JPEG or GIF files. These files all share the same characteristics that they are grid based and each cell contains a value. In these examples the value corresponds to a color that is then shown on the computer monitor. These three file formats, and the large number of other image formats out there are relatively primitive for storing spatial data since their point of origin is always stored as 0,0.

A better format for storing spatial raster data are TIFF files, and more specifically GeoTIFF files. GeoTIFF files are useful because they can be read just like a regular TIFF file. GeoTIFF files also contain geographical reference information. This allows the GIS to correctly locate and represent the image in its appropriate location relative to other types of geographic data. It is important when working with GeoTIFF files not to corrupt the image header that contains the spatial reference information. However, TIFF files are still mostly limited to aerial photography and other remote sensing applications. What happens if we want to store spatial data other than imagery?

For storing spatial data other than imagery we have to start looking at more specialized file formats. ESRI GRID files are raster files that are designed for use in a GIS system. GRID files are capable of storing attribute data that is nominal (Swamp, Desert), as opposed to the earlier image formats that are only capable of storing numeric values.

Vector file formats also come in a variety of formats. Most GIS systems out there are capable of utilizing a wide variety of vector-based data. One of the most common file formats for vector-based data are DWGs or AutoCAD files. AutoCAD files are not technically GIS data. AutoCAD is simply a vector-based drawing and drafting program. While AutoCAD operates using a Cartesian coordinate system, AutoCAD doesn't take into consideration any of the issues regarding map projection. Therefore while it is possible to use AutoCAD to draw spatial features in the correct locations, you will not achieve the same usefulness of the data that you can achieve when storing spatial data in a GIS system. As well, AutoCAD files are primarily used in drafting and typically show spatial features without carrying any attribute data. Perhaps the biggest issue with AutoCAD is simply that AutoCAD files do not support topology. One of the most important functions of a GIS is being able to determine the neighbours of selected polygons; because it does not support topology, you cannot do this in AutoCAD. AutoCAD files are however, a very good starting block when creating vector-based data.

In addition to the GRID raster format, ESRI has a format called a coverage for storing vector data. Coverages consist of a series of linked files in a subdirectory to represent different types of discrete features. For example, a .PAT file contains the attribute linkages for polygons, and a .PAL file contains a list of the lines that make up the polygons. Although ESRI has come up with two newer file formats since coverages were commonly used, there still exists a large store of data out there in coverage files and GIS analysts regularly come across them when performing spatial analysis.

Another ESRI vector file format that is more commonly used in today's GIS world is the shape file. Shape files were first used in the mid nineties, and have a very simple file structure, which does not store any spatial relationships (topology). However, shape files remain one of the most popular file formats in the world. They are popular because they are very small and easy to transfer, they use a DBase (.dbf) table that is very easy to edit in most spreadsheet or database applications, and the software to read them is inexpensive or free.

One of the newest vector formats is the geodatabase. A geodatabase is an object-relational data format, which will be examined in more detail in Section 4. Geodatabases store feature classes, which are specially designed vector formats offering many specialized functions that were not available in coverages and shape files.

3.4 Geographic Data Quality

After examining the most common spatial data formats and their sources we can now start to examine ways in which we can assess the quality of our data. There are many aspects to consider when talking about data quality. Data quality assessment is a very important and often overlooked part of a GIS spatial analysis.

It's important to understand how the use of non-quality data can lead to error propagation in a GIS analysis. There are many different ways we can assess the quality of spatial data. The first way to assess data quality is to examine the precision and accuracy of the data. The second way is to confirm the validity of the data. The third way to measure the quality of spatial data is based on the quality of its metadata.

3.4.1 Precision and Accuracy

The first consideration in measuring the quality of spatial data is to examine its precision and accuracy. In order to evaluate the precision and accuracy of a data set we must first understand the subtle but very important difference between precision and accuracy (Figure 111111).

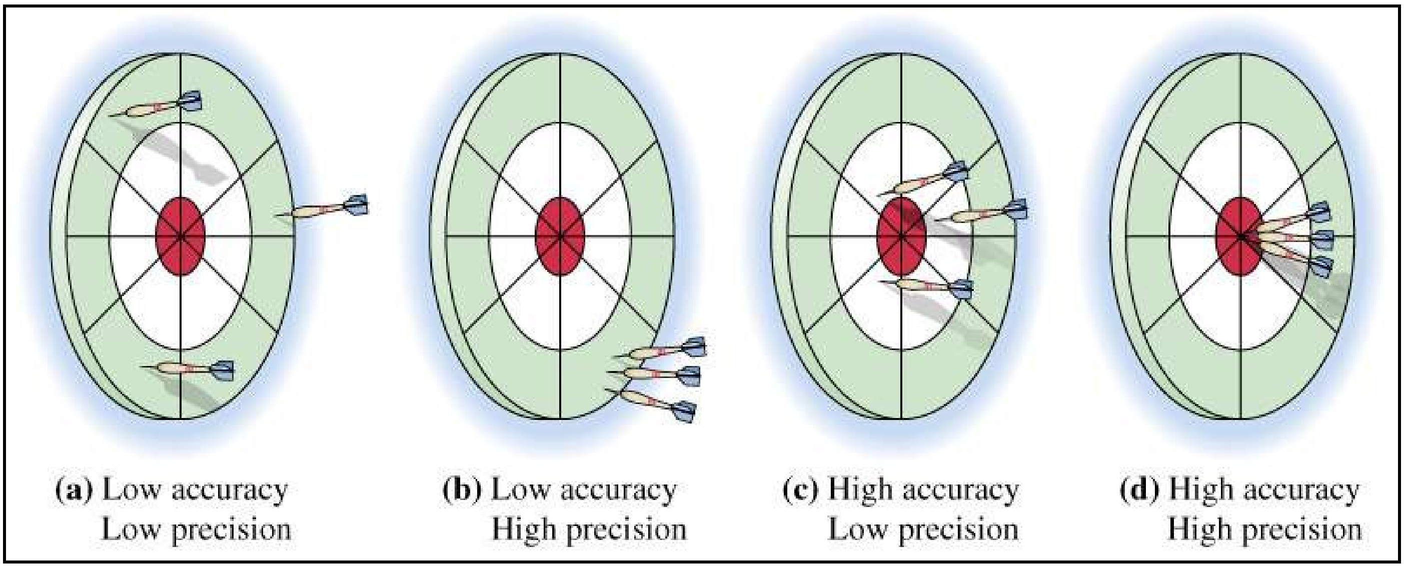


Figure 1111. Comparison of Accuracy and Precision. Source: Hill, John W. & Ralph H. Petrucci (2002). *General Chemistry, An Integrated Approach, 3rd Edition.* Upper Saddle River, New Jersey, Prentice-Hall.

Precision and accuracy are terms that are sometimes used to describe the same concept. However, scientifically these terms have very specific and completely different meanings.

Precision is the word used to relay the smallest possible division in the scale used to measure each observation. Precision therefore determines the smallest feature we can measure and therefore map or store in our spatial data. In terms of spatial data highly precise data will have very small divisions in its measurement scale and thus will be able to store spatial measurements to a very fine degree. Data that are not very precise would have very large divisions in its measurement scale and therefore would not be able to store spatial measurements with as fine a degree of measurement.

Perhaps the easiest way to understand precision is to think of measuring something using two different rulers. The first ruler measures to the nearest millimetre, whereas the second measures to the nearest half-millimetre. As a general rule, we only consider a ruler to be precise to about one half of the smallest division shown. Thus, our first ruler is precise to half a millimetre, and the second is precise to one-quarter of a millimetre. Therefore, a data set with a precision of 1 cm can store spatial features at a hundred more positions than a data set with an accuracy of 1 m.

Accuracy on the other hand is a different subject altogether. It is possible that a highly precise data set can be totally inaccurate and a highly accurate data set may not always be precise. Accuracy refers to how well the measured value corresponds to the real world value. For example if a feature sits at 4X and 4Y in the real world and we measure it at 4X and 4Y in our data set then our data set is considered accurate. However if we place the same feature at 2X and 2Y in our data set then our data set is not accurate. Accuracy is usually measured in metres in a GIS database. A data set may say it has an accuracy of less than 10 m, which means that all features in the dataset are within 10 m of their location in the real world. Depending on the scale you are working at 10 m may be high accuracy or low accuracy.

3.4.2 Validity

Data validity refers to the completeness and appropriateness of the attribute data. It is one of most important things to consider when talking about data quality and one of the most difficult to catch and correct. Valid data are basically data that has appropriate attributes for each of the attribute fields. If the attribute data are a date field then it only contains dates. If the attribute data are numeric then the attributes are all numeric. Data that are not valid would have text in numeric fields, dates in text fields and numbers.

Other examples of data validity errors are dates that use 14 months, or proper names that are misspelled. Data validity errors are quite common and are typically the result of erroneous entries in the attribute table during data entry.

One of the ways we can ensure data validity in our GIS data is by specifying field types and defining ranges in our attribute tables. Different software packages have different ways of accomplishing this but essentially by defining field types and ranges we limit the data to acceptable values.

Some common field types in a database are:

- Text – Capable of storing any kind of string data. Numbers entered into a text field are stored as text and not as values.
- Integer – Capable of storing any kind of whole number data. Sometimes integer data fields are further broken into small and large integers to further the range of acceptable values
- Float – A field type designed to store ratio numbers. Double precision a more accurate way of storing numbers than float (also known as single precision)
- Date – Stores date information. Will not allow any dates that are beyond the range of 12 months.

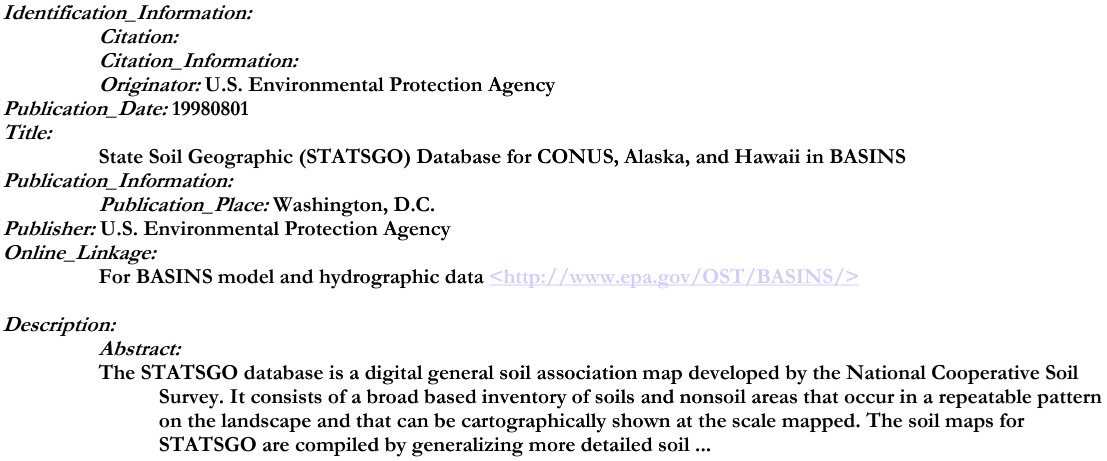
Domains operate somewhat differently than field types. Domains act to further restrict what an operator may add to an attribute field by storing a list of acceptable values. For a text field this list would include all of the values you would expect to find in that field. If a value shows up that is not in the list, the domain will not let the operator enter it. Domains on number fields act in a similar way by specifying a range of acceptable values. If an operator tries to enter a value that falls outside that range the domain will not let the operator enter it into the database.

By using geographic data with field types and domains we can be much more confident that our data are valid. As well, by using field types and domains when we create data we can ensure that our new data will also be valid when we use it for analysis.

3.4.3 Metadata

Another critical step in assessing the quality of our spatial data is the quality of the metadata. Simply put, metadata is data about data. Metadata stores such things as when the data were collected, by whom, for what purpose, and contain a general accuracy assessment among other things (Figure 12). Often the metadata will also include a legal disclaimer describing the acceptable uses of the data. A contact name is usually included in metadata stating where the GIS Analyst can go to look for more information.

Metadata is most useful when using secondary data. In this case the analyst wasn't present when the data were collected and would have no idea about how or why the data were collected. Metadata becomes critically important. It is also important to have metadata when using data from the web. The Internet allows for the exchange of cheap and often free spatial data. Without metadata there is no way to determine whether the free data are accurate, valid, or even recent.



Identification_Information:
Citation:
Citation_Information:
Originator: U.S. Environmental Protection Agency
Publication_Date: 19980801
Title:
State Soil Geographic (STATSGO) Database for CONUS, Alaska, and Hawaii in BASINS
Publication_Information:
Publication_Place: Washington, D.C.
Publisher: U.S. Environmental Protection Agency
Online_Linkage:
For BASINS model and hydrographic data <http://www.epa.gov/OST/BASINS/>

Description:
Abstract:
The STATSGO database is a digital general soil association map developed by the National Cooperative Soil Survey. It consists of a broad based inventory of soils and nonsoil areas that occur in a repeatable pattern on the landscape and that can be cartographically shown at the scale mapped. The soil maps for STATSGO are compiled by generalizing more detailed soil ...

Figure 12 Sample Metadata

There are many ways to store metadata. One of the most common ways to store metadata is as a separate file that accompanies the digital data. Often these are simply text files containing information about the digital data. In some cases a data dictionary is also included with the digital data. A data dictionary describes the data in more detail and includes information about attribute fields and any coded attribute values that may exist. Without a data dictionary, coded attribute data are nearly impossible to understand.

More recently, metadata files are being stored in an XML format. This allows the user to create attractive metadata files that have a standard format. One of the groups leading the move to standardize metadata is ISO and they have created metadata templates that are incorporated in many of today's GIS packages.

As a GIS professional, it is very important to create and maintain metadata for any datasets you may create or edit in your career. By maintaining up to date and complete metadata we can greatly increase the usefulness of any geographic dataset.

3.4.4 Error Propagation

Error is one of the most feared terms in the GIS world, but unfortunately it is one of the things we must consider when working with any type of spatial data. There are basically two types of error to consider when talking about GIS data.

1. Systematic errors - Errors that are created through some flaw in the analysis process

2. Random errors - Errors that cannot be predicted or expected.

Systematic errors are the easiest errors to identify and correct. Systematic errors typically affect the entire data set and are quite easy to spot. Correcting systematic errors is a matter of identifying the step in the analysis that is producing the error. It could simply be a problem of multiplying a data value by the wrong number, or an incorrect calibration on a measuring device.

Random errors are somewhat more problematic since they are harder to identify and harder to fix. Random errors can be created through a number of processes. Operator error and data entry errors are all types of random errors. Most data sets usually have some degree of random error associated with them. The measurement of error in any given data set is referred to as uncertainty.

A data set with a high uncertainty therefore is likely going to have a much larger amount of error present than a data set with a lower degree of uncertainty. You would imagine then by always selecting the data set with the lowest degree of uncertainty we will always end up with the most accurate final analysis. However another important concept to understand is error propagation.

When we combine two data sets with any degree of uncertainty in them we increase the overall uncertainty of the final product. This should be intuitive. If we are unsure of some of the data in one data set, and unsure of the data in another data set, when we combine them we are combining errors with errors.

There are many ways to measure the propagation of error. Most of which are covered in most basic statistics courses. For our purposes it is only important to understand that performing analysis on data with uncertainties increases the overall uncertainty of the output. Therefore it is critically important to make every effort to start a GIS analysis with data that are as error free as possible.

3.5 Spatial Databases

We've now discussed sources of data, basic types of data models, and factors to consider when assessing the quality of geographic data. The final topic of this module deals with the organization of geographic data.

As GIS analysis becomes more and more complicated, the need for efficient and intelligent data storage is becoming more and more important. We've already briefly mentioned the sheer volume of geographic data that is now available. Being able to store and find pieces of data as they become useful is critically important. Another consideration is that some of our data sets may be spatially related. Changing one feature in one data set may require changes to features in another data set. And finally, by creating specially designed storage systems for geographic data we can make our analysis of data much more efficient.

We will discuss three of the most common forms that geographic data takes. One of the most common is the shape file. Shape files have many advantages and have briefly been mentioned before. Shape files are an older data format, and there are many challenges to using them for spatial analysis. A newer approach to spatial data organization and storage is the geodatabase. The geodatabase has many advantages over shape files as will be illustrated. And finally we will discuss object-oriented models of data storage.

3.5.1 Shape files

As was mentioned earlier in this module shape files remain one of the most popular and numerous forms of spatial data. Shape files have a vector-based file structure and were created by ESRI in the early to mid nineties. The popularity of shape files stems from the fact they have a very simple file structure. It is this simple file structure that allows shape files to remain very small in size, while storing an enormous amount of features. By comparison shape files remain the smallest file type in terms of physical memory required to store them. The viewers for shape files are inexpensive or free and are distributed throughout the Internet. As well, shape files use a separate dbf table to store attribute data. Shape files are compact and usually only take seconds to move, copy, or transfer.

However, shape files also have several drawbacks. In organizational terms shape files exist as standalone files on the hard disk. Each shape file can be a point file, a line file or a polygon file. Each shape file is completely independent of other shape files. This means that shape files must be organized in windows folders, or on CDs, or some other manual organizational method to keep them in order. It is impossible to define relationships between two different shape files. Therefore, if we have a line shape file of water mains and a point shape file of pressure valves that sit on top of that water main then we would have to move both the line and the points separately should that water main ever be moved. Topological rules are completely absent from shape files which means that polygons may overlap each other and lines can cross each other any that way they want leading to potentially inaccurate data.

Shape files also have a tendency to be very unstable, and can be corrupted quite easily. It is often wise to have a backup copy of any shape files on which you are performing analysis. Shape files do not automatically store geometry information such as perimeter length and area, and these fields, if they exist, are not updated as the feature changes. Shape files are also limited to DBase database conventions regarding field names, including a 13-character limit on column names and a 255-character limit on text fields.

Because of these limitations, there was demand to build a better way to store quality spatial data while maintaining the small sizes of the shape files. It was because of this demand that the geodatabase was designed.

3.5.2 Geodatabases

Geodatabases are based on standard relational database applications such as Oracle and SQL Server. Geodatabases store data as a series of feature classes. Feature classes are similar to shape files in that each feature class may be a point, line or polygon file. Each individual point, line or polygon is referred to as a feature. One major difference between shape files and geodatabases is that feature classes do not exist as stand alone files on the hard disk. Feature classes need to live within a geodatabase. In addition to the three standard vector data types, feature classes may also represent annotation in a geodatabase. Tables, rasters, and Triangulated Irregular Networks (TINs) can also be stored as part of a geodatabase. We can see that the geodatabase offers a convenient central location to store a very diverse amount of data types.

Feature classes may also be organized into feature datasets within the geodatabase. Feature data sets represent features that share some commonalities. For instance the data types may exist within the same spatial extent, or the feature classes all represent parts of the same thing. Storing features as parts of data sets allows us to further organize our data.

Aside from the obvious organizational advantages, geodatabases have several other advantages. One of these advantages is that all the spatial, attribute, tabular, and topological data are stored in a single relational database. These databases therefore can be opened and edited with a number of different database software packages. Geodatabases also allow the creation of “intelligent” behaviours as part of their storage.

Some examples of “intelligent” storage include domain restrictions. As discussed above domain restrictions can greatly increase the validity of our datasets. By using domains we ensure that each of the attributes we enter must have a reasonable value for that particular field. When we enter the land use of a parcel of land for example it must have a value such as residential, commercial, or industrial. Another example of intelligent behaviour is the enforcement of topological restrictions. In a geodatabase it is possible to define and enforce topological rules. For example, it is possible to make rules stating that polygons or features from feature class A, industrial waste sites, cannot be located with 10 km of polygons or features. We therefore have greater data integrity since there is much more error checking.

Geodatabases may also enforce geometric restrictions. All lot boundaries must have corners that are right angles. And finally geodatabases can be used to create relationships between different feature classes. If an overhead wire is moved in the database then the related transmission towers that support it are moved as well. This “intelligent” storage gives us much richer data.

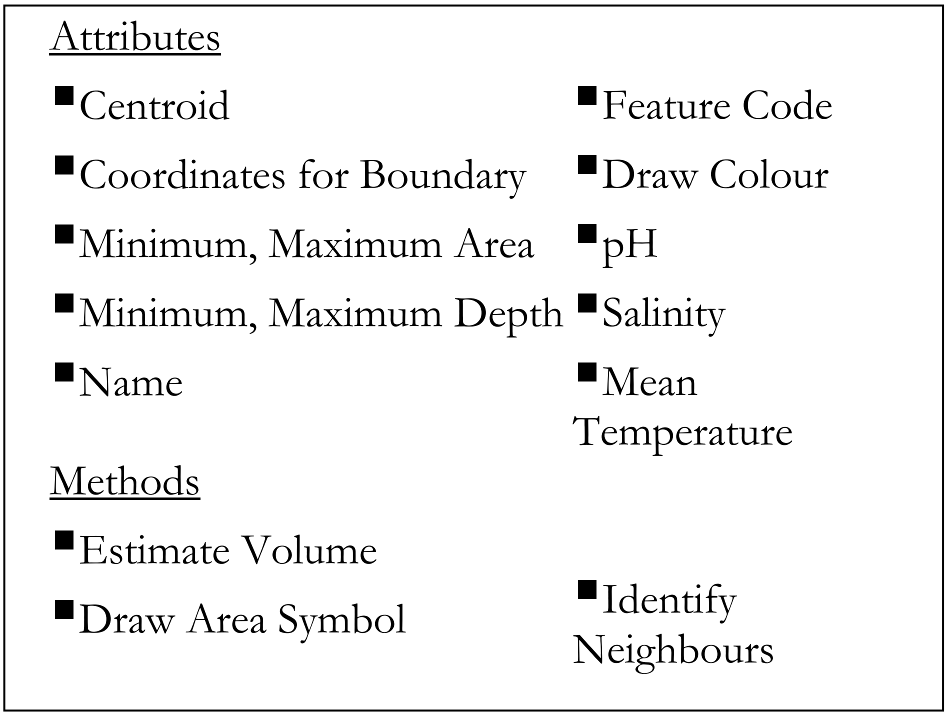
Another feature of geodatabases is that they can allow versioning, meaning many users can edit the data simultaneously without corrupting it. The versioning feature is simply unavailable in the simple shape file.

3.5.3 Object-oriented Databases

Object-oriented databases have their roots in computer programming. The premise of object-oriented systems is that users should interact with objects and operations as opposed to features and attributes. In the object-oriented system, objects know what they are and what their relationships are to nearby objects. Object-oriented systems are still not widely used because they are quite expensive to develop and there is enormous investment already in the current relational database model used by the most popular database engines such as SQL Server and Oracle. As well object-oriented systems require a great deal more processing power than the standard GIS databases

Object-oriented systems are quite different from the raster and vector models used by the previous two storage systems. Object-oriented models are less concerned with the details of storing computer data and are more concerned about the relationships between the objects being stored.

While the object-oriented concept can seem quite abstract to begin with, there are several important concepts to understand. The first is abstraction. In an object-oriented model only the essential details are stored. Anything non-essential or unimportant is omitted. Instead of storing every single detail about an object, the object-oriented model sticks to the ones that are most important. The second concept is encapsulation. Each object has a set of methods (Fig. 13) by which the object can be created, deleted or modified. Each individual object, feature or table, in the object-oriented database has its own unique methods, although similar features have similar methods.



Attributes

- Centroid
- Coordinates for Boundary
- Minimum, Maximum Area
- Minimum, Maximum Depth
- Name
- Feature Code
- Draw Colour
- pH
- Salinity
- Mean Temperature

Methods

- Estimate Volume
- Draw Area Symbol
- Identify Neighbours

Figure 13. Sample object-oriented water body. Source: Introduction to Object Oriented GIS Technology (http://home.klebos.net/philip.sargent/oo-gis/ppframe.htm)

The third concept is classes and subclasses. Each object contains its own attributes and the attributes of other higher-class objects. Each object is unique, and each object is an instance of a class. Related to this is the next concept, inheritance. Each object inherits the properties of its class. For example a school might exist as an instance of the class "building." The school therefore inherits all the properties of a building such as address, age, and size. The final concept is polymorphism. Methods may be defined differently for different objects. Therefore each object will have its own unique set of methods.

The object-oriented database is not used as widely as the other two methods of data storage. There is, however, a great deal of promise applying object-oriented concepts to GIS, and in the near future objected oriented systems may become much more prevalent.

Summary

We have taken a very thorough look at geographic data and its importance in GIS. First we examined data sources. We learned the advantages and disadvantages to using both primary and secondary data sources. Knowing the merits and drawbacks of these will help you to effectively find data sources that are appropriate to the needs of your particular projects.

We looked at the two fundamental ways in which a GIS stores spatial data. We discussed vector data, and introduced the three fundamental elements of vector data. Vector data are currently the most popular form of spatial data available. We discussed some of the pros and cons of using this method of data storage. We also looked at raster data and discussed many of the technical issues surrounding the use of raster data. These two types of data are fundamental to GIS and it is very important to know and be able to distinguish the difference between the two of them. We also briefly looked at some of the common file formats that are used to store and distribute these two types of data.

Next we looked at data quality. The very important and very subtle difference between accuracy and precision was explained. It will be important throughout your GIS career to be able to distinguish between the two since they have different meanings. We investigated data validity and how using attribute fields and domains can help ensure that that attributes we are storing are reasonable for that particular field. Metadata was discussed and its role in data quality was examined. Metadata is perhaps one of the most overlooked pieces of information that accompanies any spatial data set. It is important to always create metadata when creating or altering any spatial data set. We briefly touched on error propagation and how combining uncertainties increases the uncertainty of the final result.

Finally we examined spatial databases. We discussed some of the details surrounding shape files. We examined why they are so popular and numerous, as well as some of the drawbacks to storing, organizing, and analyzing them. We examined the geodatabase. We briefly listed some of the significant advantages it presents for GIS analysis. The geodatabase allows far more functionality out of our spatial data than ever before. Object-oriented models were also discussed. Although they have yet to really take off in the GIS realm they do show a great degree of promise and it is good at the very least to be aware of them.

As was stated at the top the fundamental piece to any GIS analysis is good quality data. With the tools and information in this module you will be able to make informed and intelligent choices when selecting and using spatial data.

Module Self-Study Questions

- Discuss a number of different types of data, and whether the vector, raster, or geodatabase model would be the best storage method.

Required Readings

- ESRI: GIS Data - (http://www.gis.com/implementing_gis/data/index.html)
- Kenneth E. Foote and Donald J. Huebner, The Geographer's Craft Project, Department of Geography, The University of Colorado at Boulder.
- Data Sources - (http://www.colorado.edu/geography/gcraft/notes/sources/sources_f.html)
- Database Concepts - (http://www.colorado.edu/geography/gcraft/notes/datacon/datacon_f.html)
- Error, Accuracy, and Precision - (http://www.colorado.edu/geography/gcraft/notes/error/error_f.html)
- ESRI: Geodatabases –
- (http://downloads.esri.com/support/whitepapers/ao_/arcgis_geodb_multiuser.pdf)
- Wikipedia: Object Oriented Databases – (http://en.wikipedia.org/wiki/Object_database)

ESRI Virtual Campus Module

- Learning ArcGIS 9 Module 4: Organizing Geographic Data
- Learning ArcGIS 9 Module 5: Creating and Editing Data

Assignments

- Lab 1: Updating Maps from Orthophotos
- Lab 2: Conflation of Unregistered Maps Practical Applications:

References

- Hill, John W. & Ralph H. Petrucci (2002). General Chemistry, An Integrated Approach, 3rd Edition. Upper Saddle River, New Jersey, Prentice-Hall.
- Introduction to Object Oriented GIS Technology (http://home.klebos.net/philip.sargent/oo-gis/ppframe.htm)

Terms Used

- Accuracy
- Block Encoding
- Chain Encoding
- Classes
- Coordinate Geometry (COGO)
- Coverages
- Encapsulation
- Error Propagation
- Feature Classes
- Geodatabases
- Inheritance
- Lines
- Metadata
- Methods
- Nodes
- Object-Oriented Data Model
- Points
- Polygons
- Polymorphism
- Precision
- Primary Data
- Raster Data Model
- Resolution
- Rule of Dominance
- Rule of Importance
- Run-Length Encoding (RLE)
- Secondary Data
- Shape Files
- Subclasses
- Vector Data Model
- Vertices

4 Methods of Data Analysis in GIS

This module outlines fundamentals of geospatial data analysis in GIS. The module reinforces vector and raster data models in GIS and introduces basic methods to analyze corresponding data. The fundamental operations, discussed in the module, are based on measures of proximity and distance, and include approaches such as buffering, overlay and map manipulations. The module elaborates on similarity of analysis of vector and raster data and, at the same time, emphasizes the differences in model analysis, based on differences in data representation. Each method of analysis is based on and utilizes specifics of data representation and has its advantages and disadvantages. The module also discusses possibilities of raster-vector data conversion that opens an avenue to employ the best methods of data analysis in GIS. The module outlines the significance of geospatial analysis for effective use of geospatial data and GIS for a variety of applications.

4.1 Vector Data Analysis

4.1.1 Data Analysis with GIS

The term “Data Analysis” is used here to describe the collection of methods, techniques and approaches to extract meaningful information from sets of data, represented in geospatial form in modern GIS packages. In other words, the role of analysis in GIS is to turn data into information and create new data by manipulating collected data. Modern GIS packages are well equipped with analysis toolboxes that employ many techniques, methods and tools to analyze geospatial data. As GIS operates with two major data modes, vector and raster, it is not surprising to have two corresponding branches of data analyses. Vector is pre-dominant data type in most GIS and this topic will discuss major techniques applied to vector data.

Spatial Analysis has several levels of sophistication: manipulation, queries, statistics and modeling. Spatial data manipulation is one of the classic GIS capabilities. This includes spatial queries and measurements, buffering and map layer overlay. Spatial data analysis falls into two categories: descriptive and exploratory analysis, implemented through visualization, data manipulation and mapping. The next level is hypothesis testing based on spatial statistical analysis which tests whether the data are “to be expected” or are they “unexpected” relative to some statistical model, usually of a random process. Spatial modeling is the most sophisticated level of spatial analysis. These methods construct models of different processes to predict spatial outcomes and possible patterns.

Geospatial data analysis is based on several fundamental spatial concepts. *Distance* is the magnitude of spatial separation; there are a number of distances used in GIS analysis, and

Euclidean distance (straight line) is the most frequently employed. Adjacency is a nominal (or binary) equivalent of distance, expressed as levels of the neighbourhood. Interaction shows the strength of the relationship between entities and is usually calculated as an inverse function of distance. Neighborhood is defined as an association between one entity and those around it and may be based upon flows or functional connections or similarity of formal attributes.

4.1.2 Vector data properties

Vector analysis is based on vector data properties: geometry and structure. Vector data models use mathematical primitives (points and their x- and y-coordinates) to construct fundamental geometric spatial features such as points, lines and polygons. As a most complex geometric feature, polygons evolve from point and line geometric primitives which compose its boundary using three line segments as a minimum. The length of these lines defines the perimeter and the area of the polygon. It is important to mention here that as a geospatial feature, polygons have attributes which allow their identification and manipulation. The location of a polygon in any given space is defined by its centroid.

As mentioned earlier, basic vector analysis is primarily based on proximity operations and tools that are used to implement the following fundamental spatial concepts:

- Buffering
- Overlay
- Distance measurement
- Pattern analysis
- Map manipulation

While the concept of distance measurement is rather intuitive, we will concentrate on buffering, overlay and map manipulation methods and leave pattern analysis for further modules dedicated to advanced geospatial analysis.

4.1.3 Buffering

Buffering creates new polygons by expanding or shrinking existing polygons or by creating polygons from points and lines. Buffers are based on the concept of *distance* from the neighbouring features.

Buffers are generated for spatial analysis to address proximity, connectivity and adjacency of features in a geospatial place.

A buffer is a spatial zone around a point, line or polygon feature (Fig. 1).

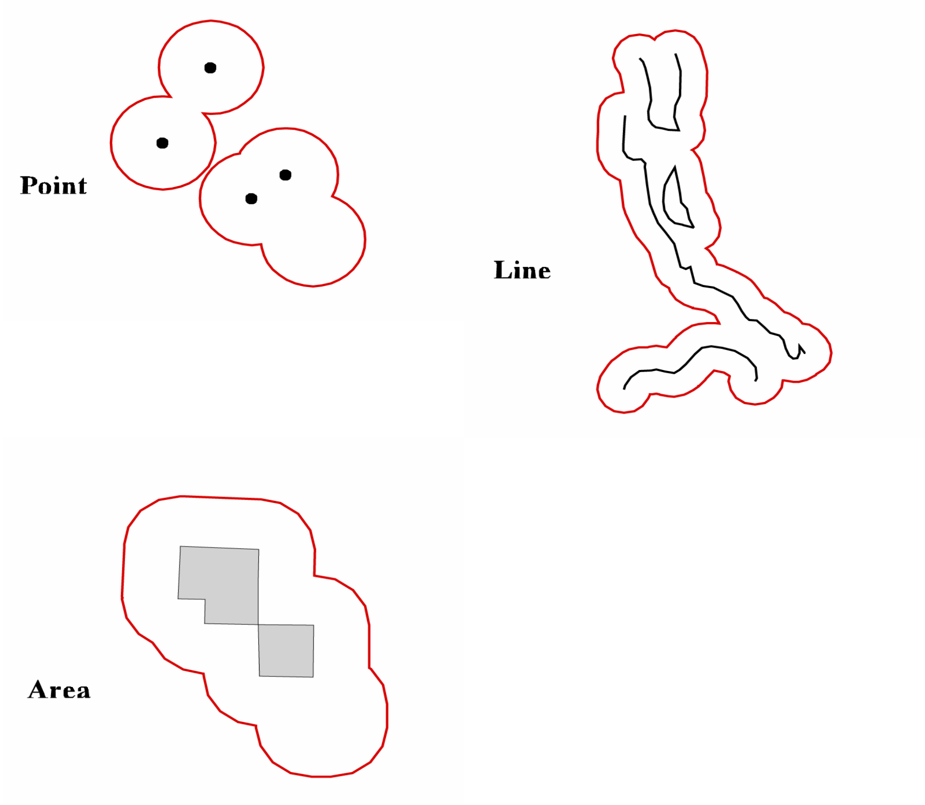
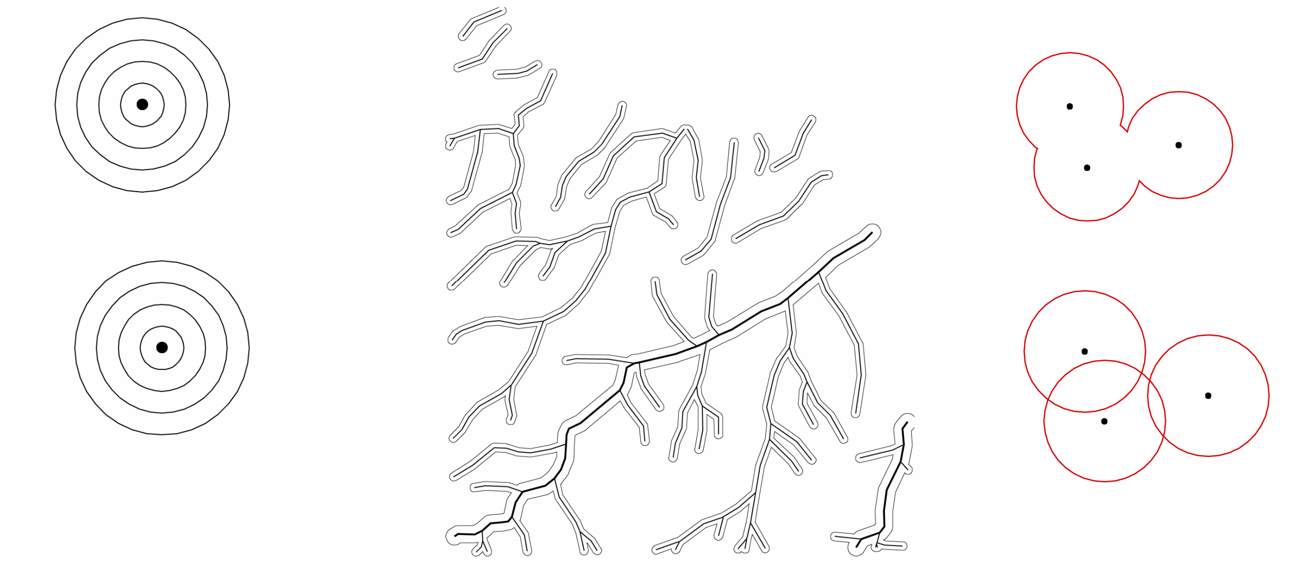


Figure1. Point, line and polygon (area) buffers

There are many variations of buffers. The shape and size of buffers can be defined by variable distance (distance based on a feature's attribute), buffers can be defined by multiple zones and can have dissolved or merged boundaries (Fig.2).



Multiple zones

Different buffer size defined by distance

Dissolved (top) and not dissolved buffers (bottom)

Figure 2. Variations of buffering

How does a buffer process work? Buffer processes use mathematical algorithms to identify the space around a selected landscape feature. First, features are selected for buffering through a variety of selection processes. Then a buffer distance is specified – it can be entered directly, specified by an attribute, or even borrowed from another table. Based on that information, the software draws a line in all directions around the features until a solid polygon has been formed and, finally, a new database containing the buffer results is created.

The point and line buffers are the simplest form of buffering with not much choice compared to buffering for polygons. In this case, users may select whether a buffer is created that represents (1) only the area outside of the polygon that is being buffered; (2) the area outside of the polygon plus the entire area of the polygon; or (3) the buffer area that is created both inside and outside of the polygon boundary. As buffers create separation zones around features, the interest can be within or outside the buffer zone.

Where and how are buffers useful? There are some examples:

- Buffering a proposed new road path to determine if wetlands are within 50 meters of a proposed road
- Buffering the point of discharge to determine if it is within 100 m of a shellfish bed
- Buffering trail systems or roads to delineate areas of visual sensitivity within which logging operations may be limited
- Buffering research areas to prevent (or hope to prevent) the planning and implementation of logging operations within them
- Buffering stream systems to delineate the distance herbicide operations must stay away from water systems. Local buildings (particularly houses), roads, agricultural fields, and orchards may also require buffering

- Buffering is a simple operation but has the potential to introduce certain problems. For example, incorrect buffer units could be created, buffering sub-selected features instead of the entire layer (theme), and mixing contiguous and non-contiguous results.

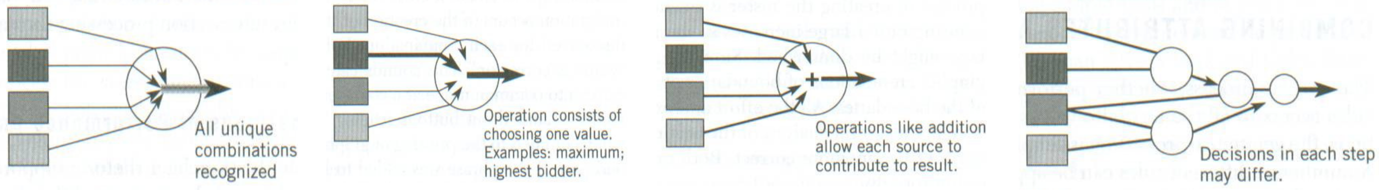
4.1.4 Vector overlay

Overlay functions, when associated with geometrical (or “physical”) overlays of data layers, are implemented by certain mathematical operations – both arithmetic and logical. Arithmetical operations commonly used, but not limited by, are addition, subtraction, division and multiplication. Logical operations are aimed on finding where specified conditions occur and use logical operands such as AND, OR, >, <, etc.

As discussed later in this module, methods for overlaying vector data differ from those of raster data related methods. However, a few brief statements could assist as a starting point: Vector methods are good for sparse data sets while raster grid calculations are faster and easier.

While overlaying, a good practice would be to follow the four basic rules (Figure 3):

- Enumeration Rule: each attribute preserved in output and all unique combinations recognized
- Dominance Rule: one value wins, means that the only one value should be chosen
- Contributory Rule: each attribute value contributes to result (example: operation of addition)
- Interaction Rule: pair of values contribute to result, i.e. decision in each step may differ


All unique combinations recognized
Operation consists of choosing one value. Examples: maximum; highest bidder.
+
Operations like addition allow each source to contribute to result.
Decisions in each step may differ.

Enumeration Rule Dominance Rule Contributory Rule Interaction Rule

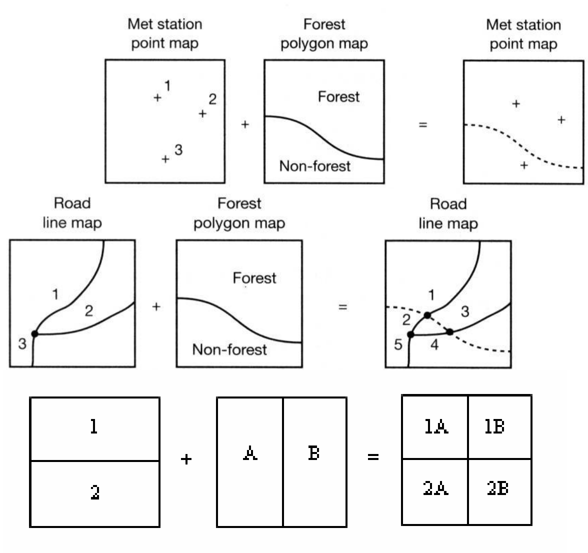
Figure 3. Overlay Rules

Points, lines and polygons form three main combinations of vector overlays: point-in-polygon, line-in-polygon and polygon-on-polygon. A point feature layer, overlaid with a polygon feature layer results in point-in-polygon combination. An example of such a combination is when a meteorological station map (a point map) overlays a land cover map (polygon map).

Lines split at the overlay polygon boundary into line segments with assigned attributes from the overlay polygon. An example of such a combination is when a road map (a line map) is laid over a forestry map (polygon map). In line-to-polygon overlays the output differs from the input in two aspects: the line is broken into two segments, and the line segments have attribute data from the overlay polygon map.

Polygon-in-polygon overlays is the most complicated type of overlay and is characterized by the following: polygons split at the overlay polygon boundary; polygon segments assign attributes from the overlay polygon; and the output combines the geometry and attribute data from the two maps into a single polygon map. Polygon overlays are used when comparing two or more data layers and described by a simple formulae: “input + overlay = output”, where

- input is a point, line or polygon
- overlay is always a polygon layer
- output is the same as input



point-in-polygon

line-in-polygon

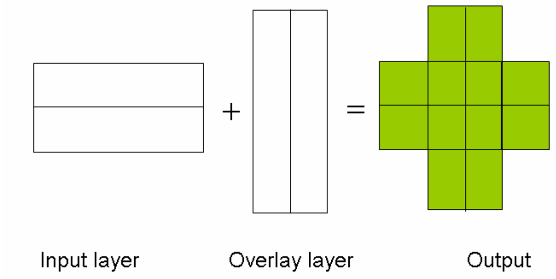
polygon-on-polygon

Figure 4. Point, Line and Polygon Vector overlays

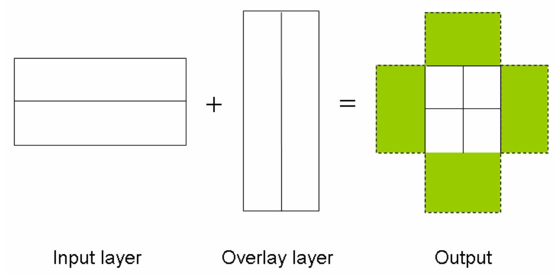
4.1.5 Map manipulations

The majority of all overlay manipulations are based on Boolean operations such as AND, OR, NOR and XOR (see http://en.wikipedia.org/wiki/Boolean_operators for details). These operations result in the following four common overlay methods: Union, Erase, Intersect, and Identity.

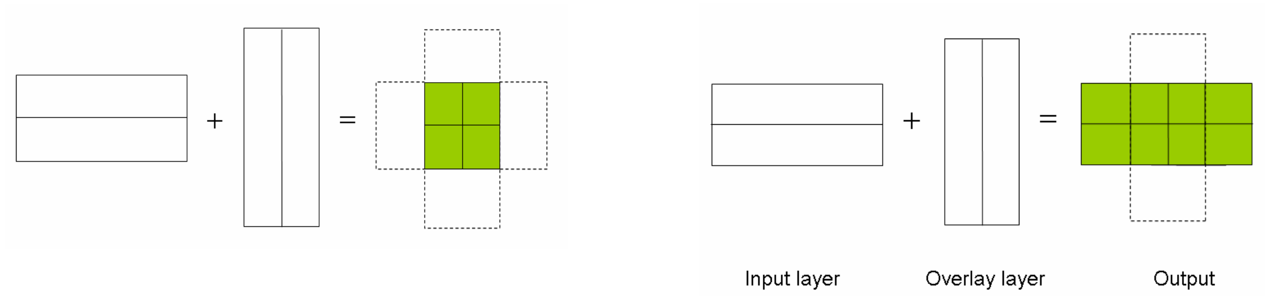
Union preserves and combines all features from both input and overlay layers and requires that both input and output layers be polygon layers. Erase discards areas of the input layer that fall inside the overlay layer, while intersect combines features that fall within the same area from both input and overlay layer. Finally, Identity produces an output that has the same extent as input layer and preserves only features that fall within the input layer. These four overlay methods are illustrated in Figure 5.



Union



Erase



Intersect

Identity

Figure 5. Overlay methods

The choice of an overlay method becomes relevant only if the inputs have different area extents.

These tools and methods are designed to be used in GIS for many manipulations with maps. Map overlay involves combining features and attributes. Map manipulation methods use multiple layers in different ways:

- Dissolve by attribute
- Clip - cutting layer using another layer
- Merge – joining maps at boundary

Map manipulation outputs show the geometric intersections of input and overlay, which need to be correctly geo-referenced.

Clip methods cut features based on extent of other features and fit one layer to the edge of another. *Merge* puts adjacent layers into a single layer and creates larger areas from smaller ones. This method merges pieces together creates a single map from two adjacent maps, but does not remove the shared boundary between the maps. *Dissolve* simplifies a feature by attribute and removes unnecessary boundaries.

One of the typical errors from overlaying polygon layers involve slivers – very small polygons created along shared boundaries during the overlay of inputs (Figure 6.). This problem in map manipulation is usually the result of digitizing errors and, sometimes, non-precise geo-referencing or data export.

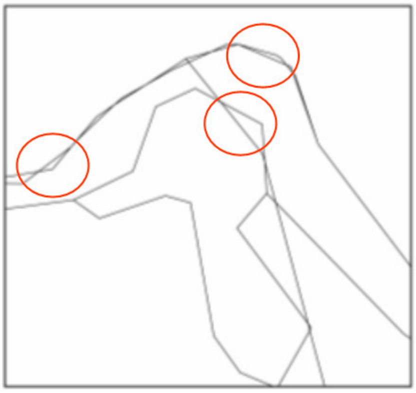


Figure 6. Slivers

These slivers are formed between the coastlines from two maps used in an overlay operation.

If the coastlines register perfectly between the two maps, then slivers will not be present.

Most GIS packages are well equipped with vector analysis tools. Below is a list of basic tools, provided by ArcGIS.

- *Extract* toolset, including
    - *Clip* which limits one layer to the exact outer boundary of another layer (e.g. limit a Texas road theme to Dallas county only)
- *Overlay* toolset, including
    - *Intersect*, which combines two polygon layers--with output limited to common area
    - *Union*, which combines two polygon layers--with output covering full extent of both layers
- *Proximity* toolset, including
    - *Buffer*, for creating buffer polygons at a specified distance around points, lines or polygons
    - *Point Distance*, for calculating distances between points within a specified radius
- *Statistics* toolset, including
    - *Frequency*, which gives you counts of attribute value combinations
    - *Summary Statistics*, which gives you summary descriptive statistics for columns in a table, including sum, mean, min, max, etc.

These tools are available in *ArcToolbox*, particularly in the *Analysis Tools*. Other tools useful for analysis of vector data are located in other toolsets as well. For example, *Data Management Tools* under *Generalization* contain *Dissolve* which removes boundaries between polygons.

4.2 Raster Data Analysis: Fundamentals

The fundamentals of raster data analysis are discussed in this section. Topics include the characteristics of raster data, operations and functions used for raster data analysis, filters and filtering raster data, the concept of map algebra and basic techniques used for such manipulations.

4.2.1 Raster data: a review

Raster data is another (along with vector) type of data used in GIS. As with vector data, methods and techniques of raster data analysis are fully defined by characteristics and internal structure, storage, and representation of raster data in GIS. Generally speaking, raster is a grid setup that has the origin (usually upper left-hand corner) and where the location of each pixel is defined by this origin and an offset. A data grid is usually a rectangular shape and its size is defined by a number of rows and columns; the grid extent is calculated by multiplying the size of the grid (number of columns by number of rows) by the size of a pixel (expressed in a metric system). Although a wide variety of raster shapes are possible (e.g. triangles, hexagons) generally a series of rectangles, or more often, squares, called grid cells, are used.

The grid could be referenced, that is, oriented relative to a known coordinate system, which could be a local or World coordinate system. Each element of a grid should have some value that represents encoded phenomena in a quantitative form. For example, elevations in DEM, temperature in thermal fields, or brightness or color in images.

Raster models are used to represent continuous data (such as elevation surfaces and fields) in ordinal or rational form, classes and groups of thematic data (such as forest species), and, finally, digital photographs and images.

Like vectors, raster can also be used to represent fundamental graphic primitives such as points, lines and polygons and confining cells for representing corresponding boundaries (Figure 7).

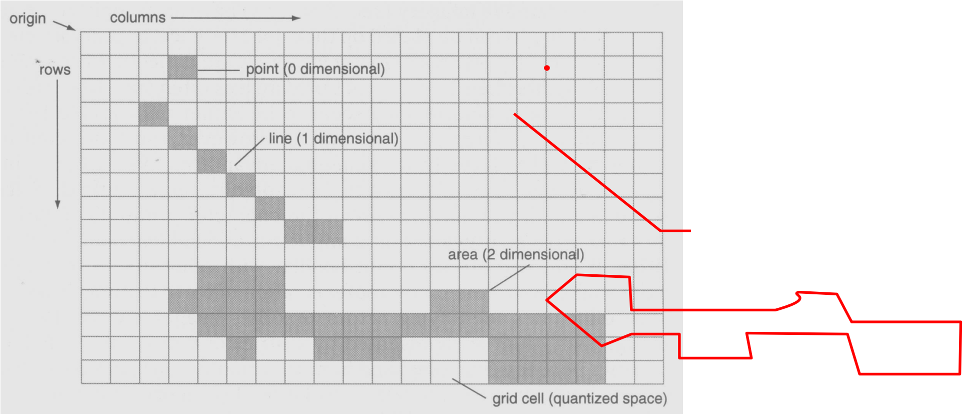


Figure 7. Representing graphic primitives in raster and vector data models

Displaying a large raster data sets at full resolution is time consuming and many GIS packages use the approach called *Pyramids*, otherwise known as a “Reduced Resolution Dataset”. Here multiple, generalized versions of rasters are pre-computed and stored to ease displaying rasters at differing scales. The original raster is only seen when one zooms in until the native raster resolution when one pixel of an image is displayed by one pixel on a computer screen. When one zooms out, one views a generalized version of the original raster. It is important to mention that pyramids are used only for display but the original raster is used for all raster calculations.

4.2.2 Raster operations

Advantages of any modern GIS system is clearly seen through its possibilities not only for displaying spatial information, but analyzing and manipulating geospatial data and information. GIS data manipulation uses map algebra and image algebra.

Generally speaking, algebra is a mathematical structure consisting of operands and operations. Applied to geospatial data, this definition could be loosely extended and interpreted as map algebra and image algebra, reflecting the specifics of vector and raster data and models:

- Map Algebra
  - Operand: rasters
  - Operations: local, focal, zonal and global
- Image Algebra
  - Operand: images
  - Operations: crop, zoom, rotate

The concept of raster operations is of fundamental value to understand map algebra methods and effectively use corresponding techniques.

There are four types of raster operations (Figure 8):

- Local: only those pixels that overlap a particular pixel are used to calculate that pixel's value (must have multiple input rasters)
- Focal : all pixels in a predetermined neighbourhood are used to calculate a pixel's value
- Zonal : use zones defined in one layer to make calculations on another (variable shaped and sized neighbourhoods)
- Global : all cells in a raster are used as inputs to calculate the value of a single pixel

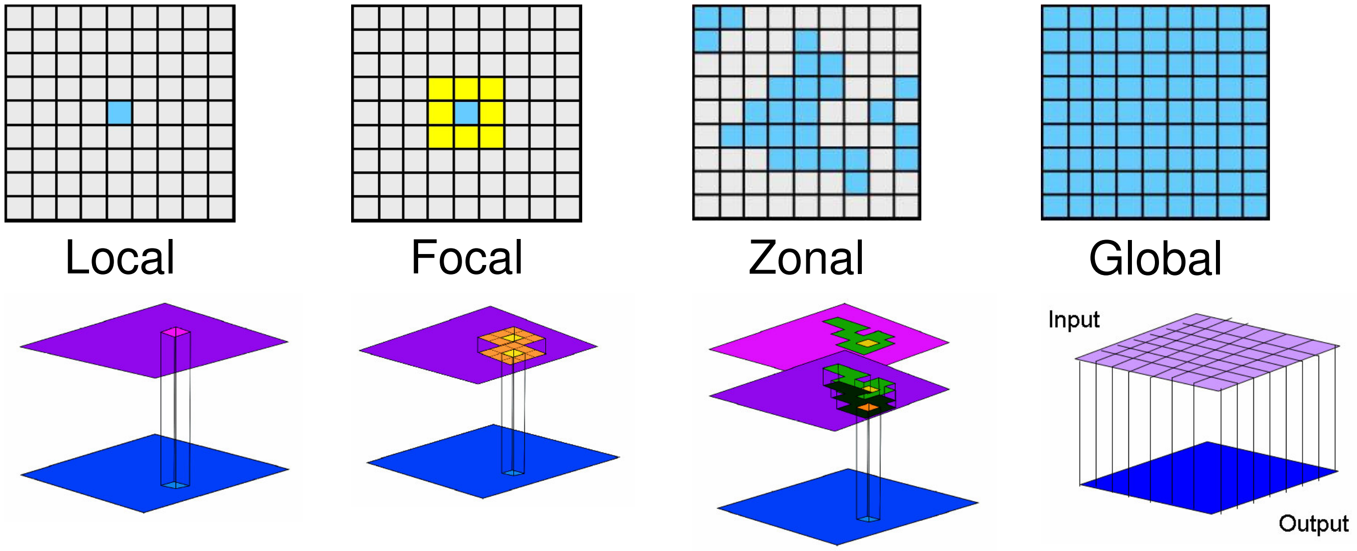


Figure 8. Four types of raster operations

Local operations:

- Perform calculation on single cell at a time
- Surrounding cells do not affect the calculation
- Can be applied to one raster layer or several

Focal operations:

- Perform calculation on a single cell and its neighbouring cells
- Also known as local neighbourhood functions

Zonal operations:

- Perform a calculation on a zone, which is a set of cells with a common value
- Cells in a zone can be discontinuous

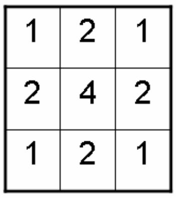
Global operations:

- Perform calculations on the raster as a whole
-

Local, zonal, focal and global operations define basic manipulations used in map algebra.

4.2.3 Filters and filtering

Many zonal, focal and global operations are implemented using filters. A filter is a small matrix with cell values designed for a specific operation (Figure 9).



1	2	1
2	4	2
1	2	1

Figure 9. Simple 3x3 filter

Spatial filtering is designed to highlight or suppress specific features in an image based on their spatial frequency. Spatial frequency is related to the concept of image texture. It refers to the frequency of the variations in tone that appear in an image. "Rough" textured areas of an image, where the changes in tone are abrupt over a small area, have high spatial frequencies, while "smooth" areas with little variation in tone over several pixels, have low spatial frequencies.

In practical implementations, filters are convoluted with the source raster by means of moving windows (kernels). A common filtering procedure involves moving a 'window' of a few pixels in dimension (e.g. 3x3, 5x5, etc.) over each pixel in the image, applying a mathematical calculation using the pixel values under that window, and replacing the central pixel with the new value. The window is moved along in both the row and column dimensions one pixel at a time and the calculation is repeated until the entire image has been filtered and a "new" image has been generated. By varying the calculation performed and the weightings of the individual pixels in the filter window, filters can be designed to enhance or suppress different types of features. The moving filter process is illustrated in Figure 10.

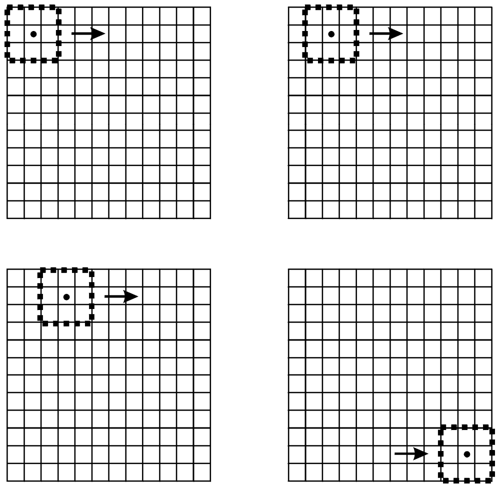


Figure 10. Filtering procedure: a kernel moving over the source raster grid

Typical kernel shape is square or rectangular, but circle and annuli are also used.

Filtering is widely used in many raster data analyses. Generic applications might include edge detection, blurring (smoothing), and noise removal. Noise may be erroneous data values, or spikes one wishes to remove. For example, spikes in a DEM may be removed through 3x3 median filtering.

Thematic applications of filtering: surface slope and aspect calculations using DEM, calculations of weighting functions for advanced multi-criteria raster analysis and many others.

4.2.4 Map algebra

The concept of map algebra could be illustrated by this simple example: "*How to identify all areas without vegetation with slope greater than 15% as high risk* "(see Figure 11).

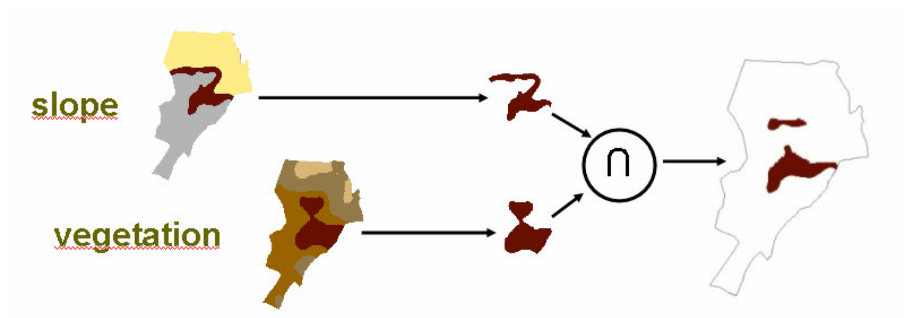
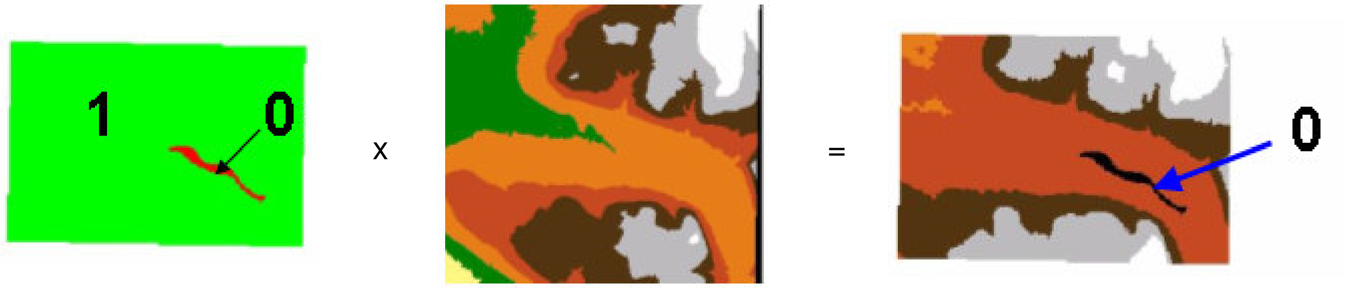


Figure 11. Map algebra concept

What is map algebra? It is an algebraic framework for performing operations on data stored in a geographical information system (GIS). Map algebra allows the user to model different problems and to obtain new information from the existing data set.

Map algebra is implemented as a cell-by-cell combination of raster layers using basic mathematical operations such as addition, subtraction, division, exponent, max, min, etc. Map algebra incorporates strong analytical functions, which allows us to perform virtually any mathematical calculation.

It is worth mentioning that some calculations applied to raster data will make sense, others won't. For example, you can create a grid where water features are 0 and land values are 1 (Figure 12). Then, you can multiply this grid with an elevation map. The output will include 0's where water existed (x * 0 = 0), and the original elevation value where land existed (x * 1 = x). Or, you can add the elevations and the grid with 0's and 1's together - but it would be meaningless.



Grid 1 x Grid 2 = Grid 3

Figure 12. Grid multiplication example

Since raster grids are arranged as array, the use of them makes map algebra very computationally efficient. Nonetheless, in either vector or raster systems, it is sometimes advisable to reduce complex data to the simplest format for site analysis.

Map algebra is an approach that defines many applied techniques and operations with raster data models. These procedure permit data resampling, reclassification, zonal operations, slicing, lookup and combinations of these. Most GIS packages have a variety of generalization functions such as *Boundary Cleaning* to smooth boundaries between regions, *Majority Filter* to replace pixels based on a majority of its neighbours' values, *Region Group* to group cells into a region based on a moving window, and *Shrink, Thin, Expand* , as well as other functions.

4.3 Raster Data Analysis: Techniques

Topic 2 provided an introduction to fundamental raster operations. The aim of this topic is to enhance our knowledge about raster data by looking into analysis techniques in detail. We will learn techniques used to implement basic raster operations (local, focal, zonal, global). We start with single layer techniques, and then will extend them to multiple raster layers.

4.3.1 Local operations

Local operations perform calculations on a single cell at a time with the condition that the surrounding cells do not affect the calculation. Local operations can be applied to one raster layer or several layers. Local operations assume any arithmetic operation with each cell in a single input layer. Example: Multiply each cell by 25.4 to convert rainfall values from inches to millimetres (Figure 13).

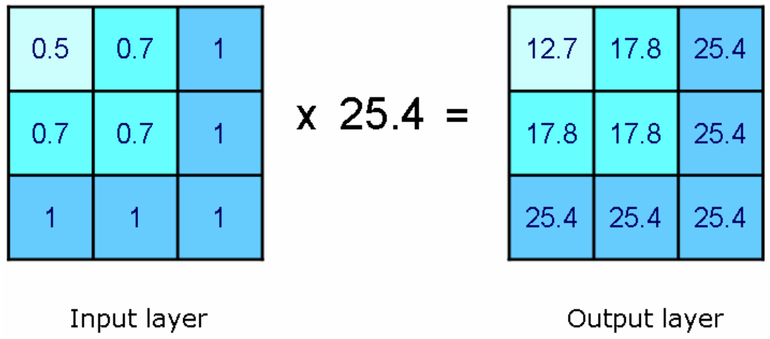
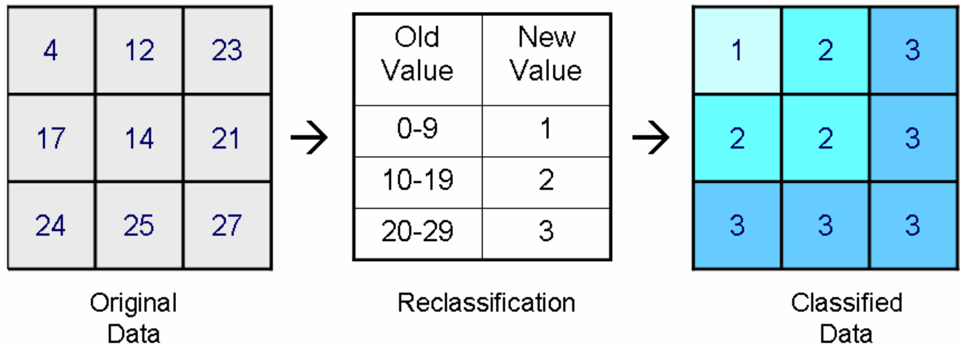


Figure 13. Local operations. Single layer

Another technique based on local operations is called reclassification. Reclassification is a generalization technique used to re-assign values in an input raster to create a new input raster.

This technique classifies cell values to be changed according to predefined rules (Figure 14).



Old Value	New Value
0-9	1
10-19	2
20-29	3

Figure 14. Principles of reclassification

Reclassification is widely used to change NoData values to something else, this operation is very useful for data sets that have gaps in cell values.

There are a number of methods for reclassifing data, such as binary masking, classification reduction, classification ranking, changing measurement scales to name a few (details of reclassification methods are far beyond the scope of the current module and could be discussed in topics related to advanced raster data analysis).

Local operations can also be applied to multiple layers. A common case is adding and subtracting layers where each cell from input layer A is added (subtracted) from another layer B (see illustration in Figure 15).

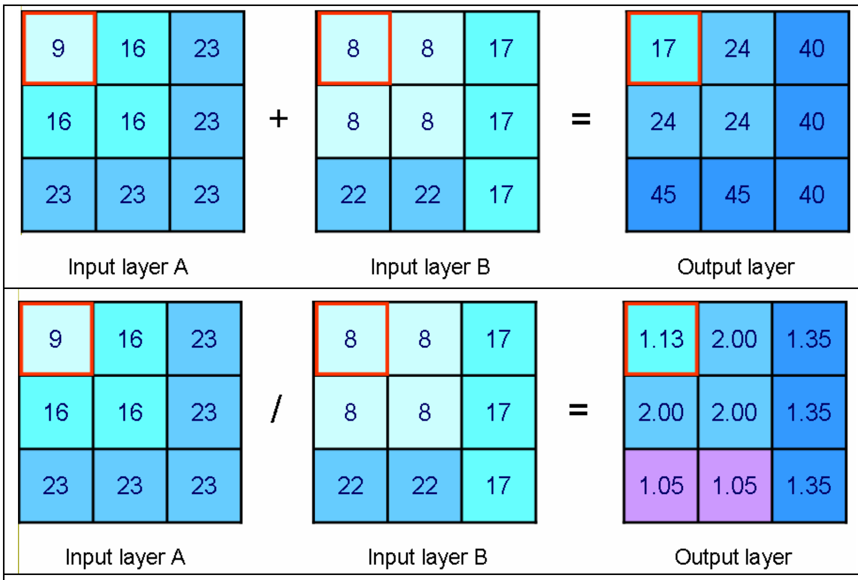


Figure 15. Local operations applied to multiple layers

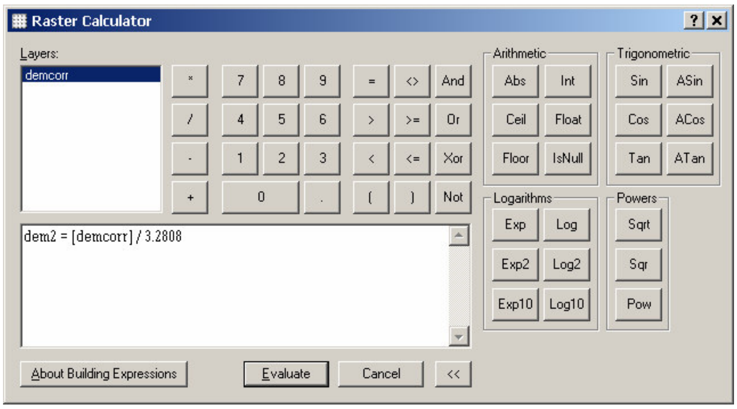
Another popular local operation is cell statistics. This operation is usually applied to multiple layers. Finding the maximum value for each cell through all layers provides a good example: each cell in the output map is based on the values of each cell of multiple input maps. For example, if you want to find the highest rainfall at each location over a 5 year period, you have to find the maximum for each cell throughout all layers and generate an new layer.

Statistics, calculated on a cell-by-cell basis between several layers, employs a number of common functions:

Maximum:	Highest value
Minimum:	Lowest value
Majority:	Value that occurs most often
Minority:	Value that occurs least often
Sum:	Total
Mean:	Average value
Median:	Halfway point (half values are above, half below)
Std. dev:	How close the values are to the mean
Range:	Difference between highest and lowest values.
Variety:	Number of unique values

Apart from arithmetical functions, it is not uncommon to implement Boolean (AND, OR, XOR, NOR) and logical operands (> , < , = etc)

ArcGIS has implemented so-called *Raster Calculator* – a tool that allows rasters to be manipulated with algebraic notation to perform local raster operations. Raster Calculator uses virtually all arithmetic operations, including Boolean operators (Figure 16).



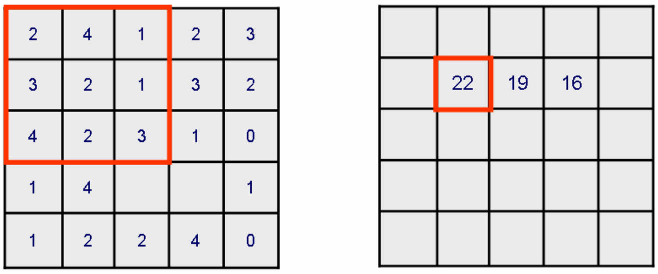
dem2 = [demcorr] / 3.2808

Figure 16. ArcGIS Raster Calculator

4.3.2 Focal operations

Focal operations perform calculations on a single cell and its neighboring cells. Focal operations are often called local neighborhood functions, where the neighborhood configuration determines the output for the cell in question. Neighborhoods can be of any size; typical shapes are rectangle, circle, annulus (doughnut) and wedge.

Principles of focal operations could be illustrated by operation of summation: look at the target cell and the surrounding cells and calculate the total value for the cell in question.



2	4	1	2	3
3	2	1	3	2
4	2	3	1	0
1	4			1
1	2	2	4	0

	22	19	16	

Figure 17. Focal operations: Sum

For many focal operations filter techniques are used. Figure 18 outlines typical spatial filters for neighborhood statistics.

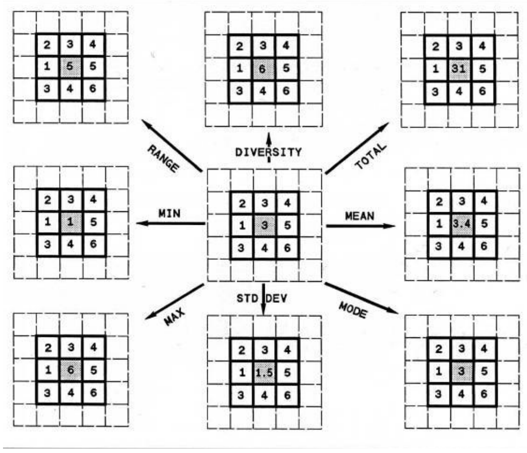
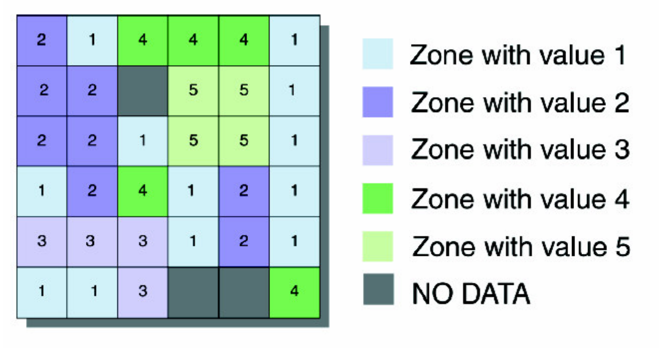


Figure 18. Neighborhood Statistics for spatial filtering

One of the practical aspects in raster data analysis relates to gaps in cell values. Such cells can be assigned a *NoData*. NoData means that no information or not enough information was available to assign the cell a value (Figure 19).



2	1	4	4	4	1
2	2		5	5	1
2	2	1	5	5	1
1	2	4	1	2	1
3	3	3	1	2	1
1	1	3			4

Figure 19. NoData in raster data

Cells with NoData can be processed in one of two ways: (1) assign NoData to output cell regardless of the combination of input cell values, that is, as long as one input cell layer is NoData, then the output will be NoData, or (2) ignoring the NoData cell and completing the calculation without it (e.g. calculating the maximum value in a neighbourhood, and ignoring the NoData cell). In practice NoData cells are ignored. That is, the calculation (total sum) is done anyway based on cells with values.

4.3.3 Zonal operations

Zonal operations perform a calculation on a zone, which is a set of cells with a common value. Zones may be continuous or non-continuous. A continuous zone includes cells that are spatially connected, whereas a non-continuous zone includes separate regions of cells.

A zonal operation may work with a single or two raster layers. Given a single input raster, zonal statistical operations measure the geometry of each zone (area, perimeter, thickness, centroid, etc.). Given two raster layers in a zonal operation (one input raster and one zonal raster), a zonal operation produces an output raster layer which processes the cell values in the input raster in zones, outlined in the zonal raster layer. A zone layer defines the zones (shape, values, and locations) and an input value raster contains the input values used in calculating the output for each zone (Figure 20).

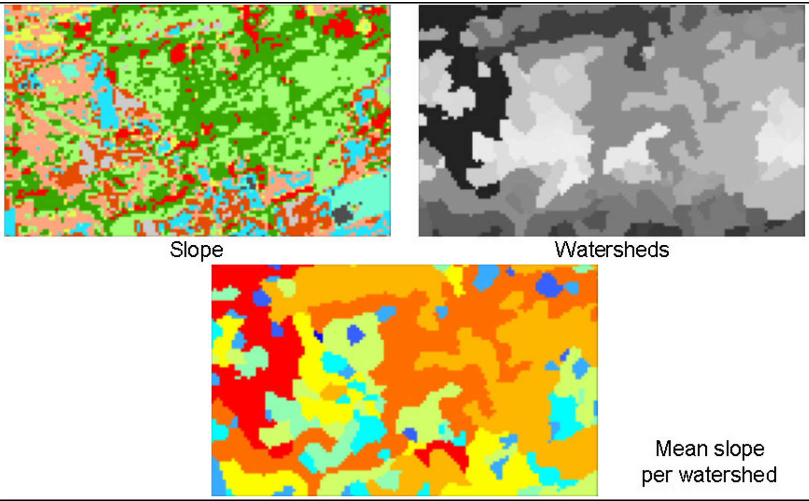
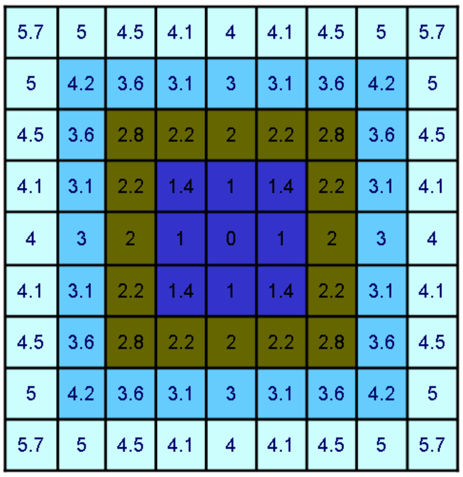


Figure 20. Illustration of zonal statistics

4.3.4 Global operations

Global operations perform a calculation on a raster as a whole. The output value at each location is potentially a function of all the cells in the input data sets. Global operations are sometimes called "operations on extended neighborhoods". The most common global operations are *Distance, Density* and *Interpolation*, along with a number of zonal operations that comprise a group of *Surface Analysis* operations.

Distance operators calculate distances across grids or find the most suitable path across the surface. In many applications weights (or costs) can be assigned to affect travel. Typical weight factors are high slopes and dense cover, but the weights could be any functions and are entirely defined by the user. Figure 21 illustrates a simple distance calculation that measures the distance from each cell to the closest source. These kind of distances are also known as Euclidean distances (or straight line distance in ESRI's terminology).



5.7	5	4.5	4.1	4	4.1	4.5	5	5.7
5	4.2	3.6	3.1	3	3.1	3.6	4.2	5
4.5	3.6	2.8	2.2	2	2.2	2.8	3.6	4.5
4.1	3.1	2.2	1.4	1	1.4	2.2	3.1	4.1
4	3	2	1	0	1	2	3	4
4.1	3.1	2.2	1.4	1	1.4	2.2	3.1	4.1
4.5	3.6	2.8	2.2	2	2.2	2.8	3.6	4.5
5	4.2	3.6	3.1	3	3.1	3.6	4.2	5
5.7	5	4.5	4.1	4	4.1	4.5	5	5.7

Figure 21. Simple distance calculation

Density. The density function is used to create grids showing the density of features within a given radius. Density tools might include kernel density, line density or point density.

Interpolation. Interpolation takes values from points and distributes them across a grid, estimating the values at points in between the source measurements. Interpolation is well studied in mathematics, and among the wide variety of available functions there are some more commonly used in geospatial analysis:

- Inverse Distance Weighting
- Kriging
- Nearest Neighbour
- Splines

Analysis of geospatial data in GIS will be discussed in detail in Module 7 which will be focusing on generic and applied (task-specific) techniques for advanced analisys and modeling geospatial data. Example of this analysis is surface analysis that includes a set of local, zonal, neighbourhood and global functions to calculate properties of a surface. Surface analysis operations are commonly applied to elevation data, but can be used on any type of continuous grid. Examples of surface analysis are slope map, aspect map, hillshade, viewshade, slicing, flow direction, hydrological functions and many others.

4.4 Raster vs Vector Analyses and Data Conversion

The main goal of topics 1 and 2 was to provide an introduction to two main types of data used in geospatial applications: Vector and Raster. These topics have emphasized that types, methods and techniques of geospatial data analyses entirely depend on corresponding data models. The goal of this topic is to re-emphasize commonality and differences in these data model types, to outline basic techniques and approaches in corresponding data analyses and highlight their advantages and disadvantages. To get the full benefit from both analyses for both data types, a common solution is introduced: convert the source data into appropriate data model (vector-to-raster or raster-to-vector) and implement the best suitable method for analysis needed. To evaluate such possibilities, a brief introduction to vector-raster data conversion is provided. Finally, as many GIS projects involve both vector and raster data, the issues of raster and vector data integration are discussed in the final part of this section.

4.4.1 Raster analysis versus Vector analysis

Vector Data Analysis and Raster Data Analysis represent two basic types of GIS analyses. There are commonalities and differences between raster and vector analyses, caused by differences in data structures and representation. The two most common techniques implemented both in vector and raster analyses are buffering and overlay.

4.4.2 Raster data models versus Vector data models

There are a number of advantages of raster data and models. These include:

- Simple data model
- Multiple spatial analysis functions often simpler and faster
- Efficient for data with high spatial variability
- Efficient for low spatial variability when compressed
- Easy to integrate with satellite and remotely-sensed data
- Topological relationships are not explicitly encoded; some analysis is more difficult

On the other side, vector data and models have their own advantages, such as:

- Can store data efficiently with high precision
- Requires about 10% of space to store same data in raster format
- Certain types of topological analysis are more efficient or only possible with vector
- Gives much greater precision and accuracy
- Greater flexibility in storing and manipulating attribute data

To see the differences between these two data models, let's compare implementation issues of two kinds of basic analysis operations, applied to both these data models.

1. Compare an arithmetic operation in raster and vector
   - Raster
     - Simply add two cells together
   - Vector
     - Must subdivide or intersect polygons first to build new polygon coverage
     - Then values in each new polygon may be added together

2. Buffering (finding all areas within a certain distance of a feature)
    - Raster
        - o Change values of cells within specified range from the target feature
    - Vector
        - o Create circular polygons at regular intervals along the feature arcs
        - o Intersect all the circles
        - o Dissolve arcs inside the circles

This comparison could be used as a good illustration on how the internal structure of the raster data model allows the best implementation of one or another techniques for geospatial analysis compared to vector data. From the other side, vector data are quite suitable for another group of operations which, in turn, require very complicated implementation schemes being applied to raster data.

In general, one could use the following pro-and-contra list as a guidance to choose the best appropriate analysis technique to analyse raster and vector data:

Operations best suited to raster analysis

- Overlays and arithmetic, Boolean, and map algebra operations
- Buffering
- Proximity, cost-distance
- Viewshed analysis (what parts of a surface can be seen)
- Any operations requiring continuous surfaces
- Projects involving data with high spatial variability
- Projects in which original data is raster (e.g. satellites)

Operations best suited to vector analysis

- Connectivity, network modeling
- Point-in-polygon and line-in-polygon overlays
- Overlays when many attributes are involved
- Evaluating contiguity
- Projects requiring high precision of stored data
- Projects in which attributes are primarily character data

The choice of the method or technique for data analysis in many cases is pre-defined by availability of a GIS package to the user. Some GIS platforms are primarily vector or raster oriented packages. For example:

- Raster-oriented GIS: GRASS, IDRISI, MOSS
- Vector-oriented GIS: Intergraph, MapInfo
- Integrated GIS platforms: ArcGIS

Some GIS packages allow you to convert between data types with ease, some do not, and the user faces the challenge of data conversion and compatibility of vector and raster formats, supported by different GIS packages. It is often convenient to use one model primarily, but convert to the other for certain operations, and then convert the results into the native (original) data format to ensure compatibility and further analysis. Such conversion-analysis-conversion operations are rarely repeated at different stages of data analysis during the project.

In summary:

1. Raster analysis is generally faster than vector analysis

2. Vector analysis provides more accurate and spatially correct results
3. To implement tools available in vector analysis for raster data and vice versa one should convert these data from one model to another
4. Methods and techniques for conversion form vector to raster data are based on fundamental features of corresponding data models

4.4.3 Vector-Raster conversions

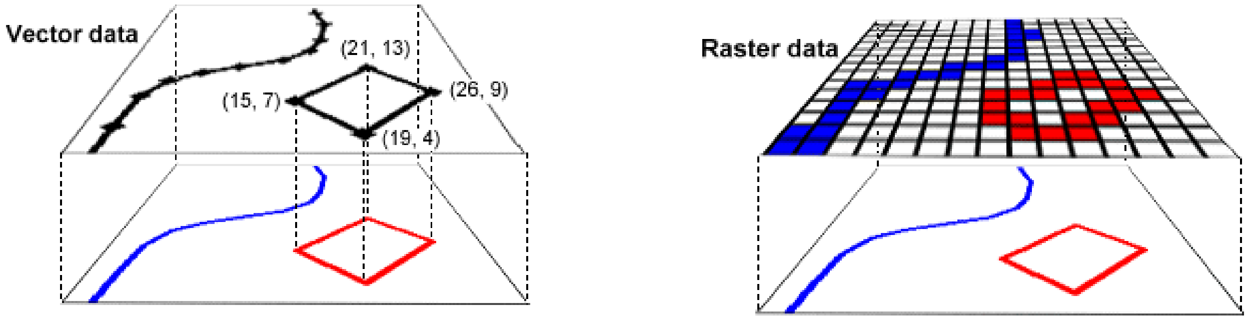
Data conversion methods, as well as data analysis methods, are entirely based on features of the source data model to be converted and characteristics of the target data model. Differences between vector and raster data structures are briefly outlined below and illustrated in Figure 22.

Raster data model:

- Resolution determined by pixel size
- Efficiently represents dense data
  - e.g. elevation

Vector data model:

- Resolution determined by precision of coordinates
- Efficiently represents sparse data
  - e.g. house locations



Vector data model: Points, lines, and polygons (areas)

Raster data model: Grid of equal sized cells

Figure 22. Vector and raster data model structures

As mentioned above, many GIS packages have tools for automatically converting raster and vector and vice versa. These processes are called *vectorization* (if data converted from Raster to Vector (R2V)) and *rasterization* (if data converted from Vector to Raster (V2R)). Normally, some information and data are lost in the conversion process; consequently, converted data are less accurate than original data.

There are some subtle issues associated with conversion-related terminology. In a strict sense, one should make a distinction between data conversion and model conversion. Model conversion preserves not only formal features of the data format, but also the structure of the information, stored in this particular format. For example, topology links do not exist in raster, and when converted into vectors, these data should be further edited to reconstruct networks and relationships which are usually found in vector models. Another example – errors in data models, such as slivers and closed loops, can be generated as a result of inaccurate raster data conversion.

Raster-to-Vector conversion approaches:

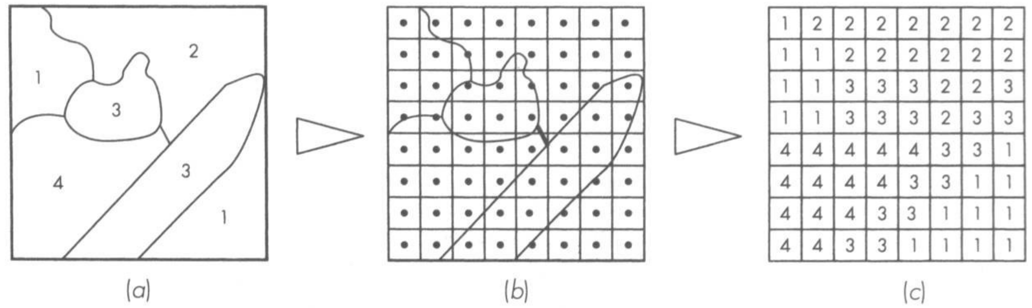
1. *On-screen digitizing*. On-screen digitizing usually implements conversion of raster data directly to vector model. It is a manual process to compile a complete vector data model by manual digitalization of raster data; depicted on a computer screen, raster data usually should be geo-referenced, so the final vector data possesses all correspond to geospatial attributes that are relevant to geospatial vector models in GIS.

2. *Vectorization.* Automated vectorization is usually completed in two steps. In the first step an automated algorithm is applied to convert raster data into vector data. The second step involves advanced data editing that allows converted vector data to be transformed into a full vector model.

Vector-to-Raster approaches are generally simpler and easier to implement and carried out as a single step process called rasterization.

4.4.4 Vector to Raster conversion (V2R)

Conversion of vector data to raster data is usually a straight-forwarded process (Figure 54). Source vector data are represented as coded polygons with (x,y)-coordinates, geo-referenced into a certain coordinate system (Figure 23, a). The process involves generating a grid with the appropriate cell size and then overlaying it on top of the source polygons - dots represent the center of each grid cell (Figure 23, b). Finally, each cell is assigned the attribute code of the polygon to which it belongs (Figure 23, c).



1	2	2	2	2	2	2	2
1	1	2	2	2	2	2	2
1	1	3	3	3	2	2	3
1	1	3	3	3	2	3	3
4	4	4	4	4	3	3	1
4	4	4	4	3	3	1	1
4	4	4	3	3	1	1	1
4	4	3	3	1	1	1	1

Figure 23. Conversion of vector data to raster data

Basic graphic vector primitives (points, lines, polygons) are converted into raster data according to certain rules:

- Points are converted to single cells
- Lines are converted to groups of cells oriented in a linear arrangement
- Polygons are converted to zones

The following are usual steps are available in most packages to implement the rasterization of vector data:

1. Setting up a raster grid with a specified cell size to cover the area extend of vector data
2. Assigning all cell values as zeros

3. Change the values of those raster cells that correspond to points, lines or polygon boundaries
4. Fill the interiors of polygons by re-assigning corresponding cells values

Rasterization can be reduced to the process of resolving which feature attribute a grid cell should be labelled – in modern GIS packages this process is usually automated. In most cases grid layers, as a whole, can only be converted directly to polygon vector layers, but only selected features are converted, or all features if no selection of feature objects is made and active.

As already mentioned, any data conversion involves a loss of quality and precision of output data. Grid cell size, position, and orientation of the grid, selected by the user, greatly impacts the speed and quality of the conversion process from vector or point formats to raster format. Rasterization generally involves a loss of precision, and this precision loss is retained if data are re-converted to vector (see Figure 24).

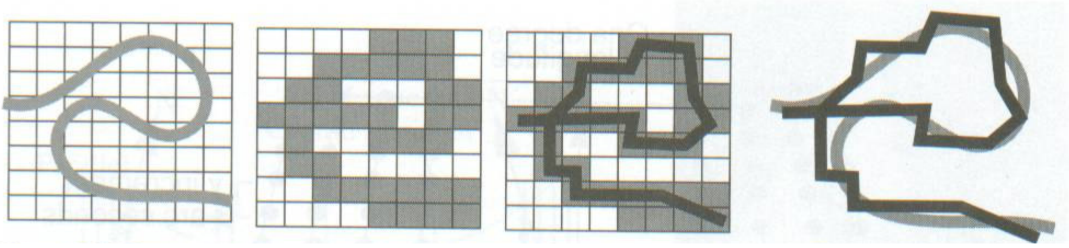


Figure 24. Example of errors caused by conversion between raster and vector data models: the original river after raster-to-vector conversion appears to connect the loop back

Vector to raster conversion can cause a number of errors in the resulting data, such as: topological errors, loss of small-sized polygons, effects of grid orientation (blurring, moiré, etc), variations in grid origin and datum and other. Further problems with converting vector data to a raster structure include creation of stair-stepped boundaries, small shifts in the position of objects, deletion of small features, etc.

4.4.5 Raster to Vector conversion (R2V)

Vectorization, as well as rasterization, is a standard data conversion mechanism in many integrated GIS packages. Vectorization involves three basic steps: line thinning, line extraction and topological reconstruction. Conversion from raster to vector varies in difficulty with the type of data.

Points and polygons are relatively easy to deal with, especially for classified remote sensing data. On the other hand, lines are relatively difficult. The typical workflow is outlined in Figure 25.

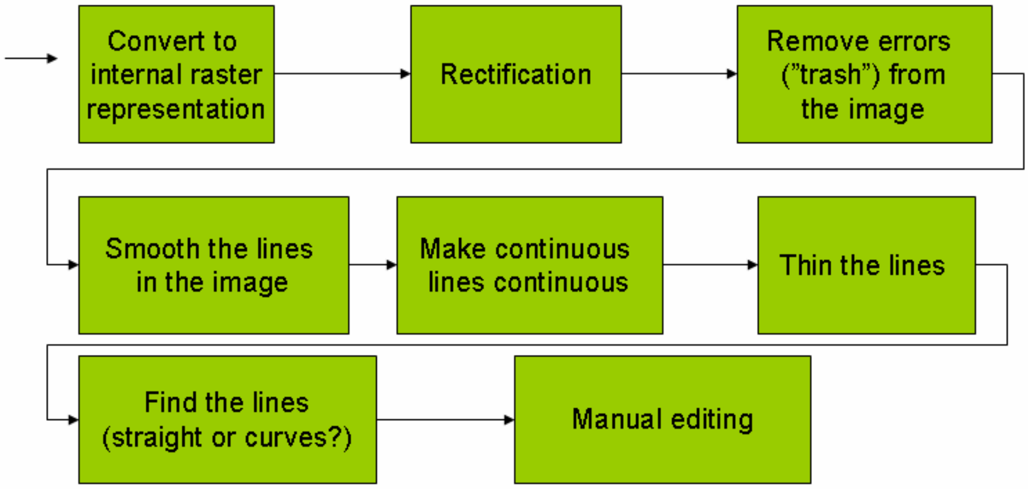

Convert to internal raster representation
Rectification
Remove errors ("trash") from the image
Smooth the lines in the image
Make continuous lines continuous
Thin the lines
Find the lines (straight or curves?)
Manual editing

Figure 25. Line vectorization workflow

Lines as a geometric primitive in vector data models have lengths but no widths. Raster lines in a scanned map, however, usually occupy several pixels in width. Thus, raster lines should be thinned, preferably up to a one-cell width, for vectorization. An algorithm decides whether a certain pixel belongs to a line by analyzing cell values in a grid. In practice, cell values could significantly vary, even for the same graphical object. As a result, it is common to see that lines break during the thinning process**.**

The goal of line extraction is to determine where individual lines begin and end. Topological reconstruction connects extracted lines as well as shows where digitizing errors exist. A typical shortcoming of vectorization is that it creates step-like features along diagonal lines, A subsequent smoothing operation is required to eliminate these step-like artifacts.

Vectorization is difficult to program and implement. Nevertheless, there are a number of stand-alone software packages and applications as well as plug-ins to GIS platforms that implement vectorization at different levels of sophistication. ARCScan is an extension that helps to create vector maps from scanned paper maps in the ArcINFO / ArcView / ArcGIS family of products. This module has a set of tools for Raster – Vector conversion, such as Raster to polygon conversion, Contour Generation, and Surface Interpolation from point data - to name a few.

From another side, vectorization is a time-consuming and error-prone process which has it's advantages and disadvantages:

- Advantages
  - Could be very fast and cost effective
  - Relatively inexpensive
  - Provides a very accurate representation of the analog map
- Disadvantages
  - The analog map needs to be in a pristine condition with minimum extra features and annotation
  - False recognition of different features and text
  - Editing can be very labour intensive

Some further problems with converting raster maps to a vector structure include:

- Potentially massive data volumes

- Difficulties in line generalization
- Topological confusion

4.4.6 Raster and vector data integration

Working simultaneously with raster and vector data is common in many GIS projects. Raster data, represented in different forms of surface grids such as Digital Elevation Models, interpolations of spotted observations of different phenomena, as well as visual geospatial data– aerial photographs and satellite imagery - are widely used in conjunction with native vector data (e.g., road networks, watersheds, land use zones, etc.). In fact, many vector data are extracted from original raster data. For example, contour lines taken from DEM. A typical example of vector-raster data integration (and transition) is the reconstruction of spatial fields (raster data in the grid form) from measurements of irregular network of sensors (point data in vectors), and then extraction isopleth lines that represent this field again in the vector form.

Summary

- Spatial analysis operates on fundamental spatial concepts
  - Proximity
  - Buffering
  - Overlay
  - Distance
  - Map manipulations
- Vector analysis
  - Mainly based on buffers, overlays and map manipulation
  - Utilize fundamental set of Boolean and arithmetic operations
  - Incorporate features and attribute information from multiple databases into a single database (layer)
- Raster data analysis
  - Actively employs filtering techniques
  - Uses local, focal, zonal and global operations
  - Computation efficient
- Common raster and vector analyses techniques: buffers and overlays
- Raster data could be converted into vector data and vice versa to apply the most appropriate method for analysis
- Vector-to-Raster conversion is much easier to implement then Raster-to-Vector

Module self study questions:

- What fundamental spatial concepts form the basis of all spatial data analyses?
- Describe the main vector and raster characteristics and describe how they define the nature of corresponding data analyses?
- Explain the principles of the four types of basic operators used for raster data analysis.
- What similar techniques are used for analisys both vector and raster data?
- Elaborate on the principles of raster-vector data conversion. Outline typical rasterization and vectorization workflows.

Required Readings:

- Chang, K-Ts., 2006 Introduction to Geographic Information Systems, 3rd ed. McGraw-Hill, New York
- Price, M., 2006 Mastering ArcGIS, 2nd ed. McGraw-Hill, New York

ESRI Virtual Campus Course:

- Learning ArcGIS 9 Module 6: Getting Started with GIS Analysis
- Learning ArcGIS 9 Module 7: Working with Geoprocessing and Modeling Tools

Assignment:

- Vector analysis tools - buffer and overlay analysis
- Raster analysis tools - working with rasters
- Vector to Raster and Raster to Vector conversion

References

- Chang, K-Ts., 2006 Introduction to Geographic Information Systems, 3rd ed. McGraw-Hill, New York
- Price, M., 2006 Mastering ArcGIS, 2nd ed. McGraw-Hill, New York
- Bettinger, P., Wing, M., 2004 Geographic Information Systems: Applications in Forestry and Natural Resource Management, McGraw-Hill, New York
- Demers, M. N., 2000 Fundamentals of Geographic Information Systems, 2nd Ed.
- Bernhardsen, T., 1999 Geographic Information Systems: An Introduction, 2nd Ed.
- Clarke, K., 2001 Getting Started with Geographic Information Systems, 3rd Ed.

Terms used

- Geospatial analysis
- Vector and raster data analysis
- Proximity
- Buffering
- Overlay
- Map manipulations
- Overlay operations: union, erase, intersect, identity
- Map manipulation methods: clip, merge, dissolve
- Slivers
- Raster data: continuous, thematic, image
- Raster data pyramid
- Raster data: numerical and graphic presentation
- Raster operations: local, focal, zonal, global
- Filtering
- Map algebra
- Image algebra
- Cell statistics
- Zonal statistics
- Reclassification
- NoData
- Raster-to-vector conversion (R2V)
- Vector-to-Raster conversion (V2R)
- Rasterization
- Vectorization
- On-screen digitizing
- Conversion errors

5 Introduction to Remote Sensing

The aim of this module is to give a brief introduction to remote sensing, a discipline that studies the acquisition, processing and use of imagery in general and particularly in geospatial science and GIS applications.

The module begins with review of fundamental concepts of electromagnetic radiation to illustrate physical principles of image formation. To utilize the best possibilities of remote sensing data, it is important to understand different approaches to register electromagnetic energy. To do so, we introduce basic principles of acquisition of electromagnetic energy by two main types of sensors: analogue films and digital sensors.

Image geometry is another important issue discussed here. This type of imagery includes frame systems, represented by classic aerial photographic cameras, and line-by-line systems, represented by modern scanners. Imaging sensors can be installed onboard aircraft or satellites. An introduction to specific features of images, acquired by these systems, is also discussed in this module. An concept of great importance in remote sensing, is image resolution – spatial, spectral and temporal image resolutions will be discussed in this module. The different factors that characterize and impact the quality of images, acquired by variety of remote sensing systems, are outlined here.

It is essential to know the fundamentals of color formation, composition and decomposition in order to understand the concept of multi-spectral and hyper-spectral image acquisition and analysis. Understanding these fundamentals is essential to understanding the basis of remote sensing.

The last section of this module is dedicated to images *per se* – specifics of their structure, representation, interpretation and analyses. Starting with classical visual photo interpretation, we introduce the idea and principles of computer-aided image classification and analysis.

5.1 Fundamentals of Remote Sensing: Electromagnetic Radiation

5.1.1 What is remote sensing?

The term “remote sensing” is widely used as a method of acquiring information about the Earth's surface without being in contact with it. In remote sensing the reflected or emitted electromagnetic energy is recorded by a sensor in a form of imagery, and then processed and analyzed in order to extract meaningful information about the objects and phenomena depicted.

Remote sensing is a multi-stage process that includes several components and interactions between them. Remote sensing needs a source of energy. When the Sun’s energy travels to the Earth, it interacts with the atmosphere and objects on the Earth’s surface. The reflected part of this energy is captured by a sensor, coded into an electric signal, and then transmitted to a station on the ground. To become valuable, this data needs to be pre-processed, corrected and enhanced. Further processing involves intensive interpretation and analysis that allows data to be converted into more meaningful information.

5.1.2 Electromagnetic Radiation

The first requirement for remote sensing is to have a source of the energy that will illuminate the target, unless the sensed energy is being emitted by the target. This energy is called electromagnetic radiation, which is a form of energy that exists in a continuous spectrum, which, in turn, is also referred to as “electromagnetic”. A typical example of such electromagnetic radiation is light from the sun.

The fundamental unit in the electromagnetic (EM) force field is called a photon. All electromagnetic radiation (photon) has fundamental properties and behaves in predictable ways according to the basics of wave theory. Electromagnetic radiation consists of an electrical field (E) which varies in magnitude in a direction perpendicular to the direction in which the radiation is traveling and a magnetic field (M) that is oriented at right angles to the electrical field (Figure 1). Both these fields travel at the speed of light (c).

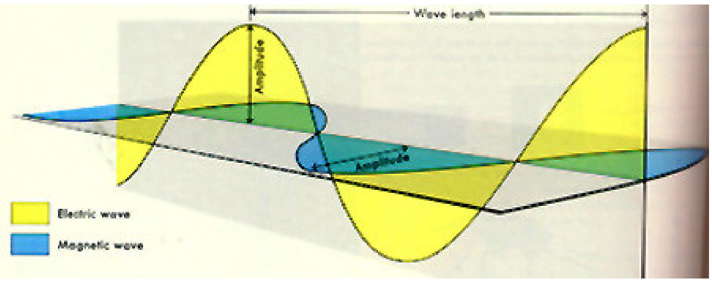


Figure 1. Electromagnetic field: electric and magnetic waves

Two characteristics of electromagnetic radiation are particularly important in remote sensing:

- Wavelength and frequency (Wave model)
- Energy or momentum (Particle model)

The wavelength is the length of one wave cycle. This can be measured as the distance between successive wave crests (Figure 2). A wavelength is usually represented by the Greek letter lambda (ג). Wavelengths are measured in meters (m) or some factor of meters such as nanometers (nm, 10-9 meters), micrometers (μm, 10-6 meters) or centimeters (cm, 10-2 meters).

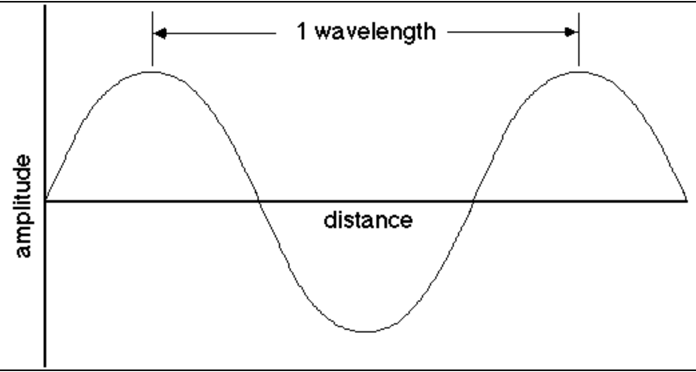


Figure 2. Wavelength

Frequency refers to the number of cycles of a wave passing a fixed point per unit of time (usually per second). Frequency is normally measured in hertz (Hz), equivalent to one cycle per second, and various multiples of hertz:

- Kilohertz (kHz, = 10^3 Hz)
- Megahertz (MHz, = 10^6 Hz)
- Gigahertz (GHz, = 10^9 Hz)

Wavelength and frequency are related by the following formula:

$c = \nu\lambda$,
where:
λ - wavelength
ν - frequency
c - velocity of electromagnetic waves in a vacuum, $c = 3 \times 10^8$ m/s

Figure 3 provides a graphical illustration of this relationship.

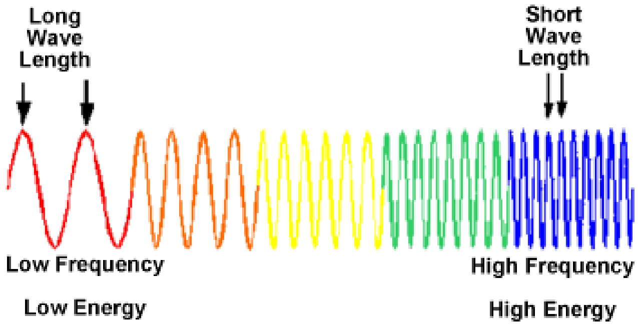


Figure 3. Relationship between wavelength and frequency of electromagnetic radiation

The electromagnetic spectrum ranges from the shorter wavelengths (including gamma and x-rays) to the longer wavelengths (including microwaves and broadcast radio waves) (See Figure 4).

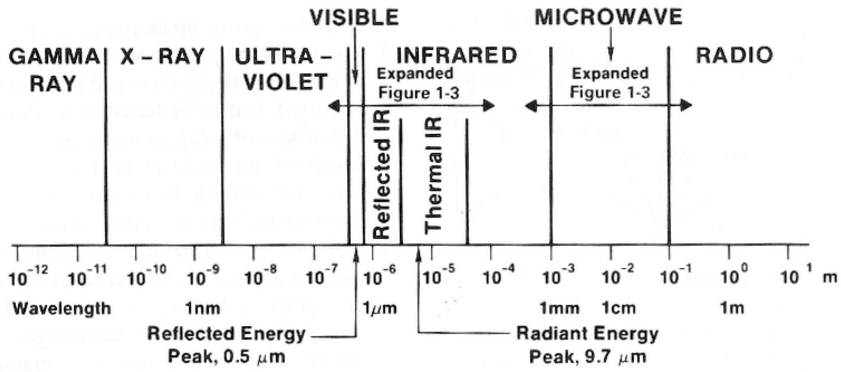


Figure 4 Electromagnetic spectrum and zones

There are several regions of the electromagnetic spectrum which are particularly useful for remote sensing: ultraviolet, visible, infra-red and microwave.

For most purposes, the *ultraviolet* (UV) portion of the spectrum has the shortest wavelengths which are practical for remote sensing. Some Earth surface materials, primarily rocks and minerals, fluoresce or emit visible light when illuminated by UV radiation.

The light which our eyes can detect is part of the *visible* spectrum. The visible wavelengths cover a range from approximately 0.4 to 0.7 μm. The longest visible wavelength is red and the shortest is violet. It is important to recognize how small the visible portion is relative to the rest of the spectrum (see Figure 4 for visible zones).

The following chart demonstrates the common colors and corresponding wavelengths in the visible zone.

Color	Wavelength	
Violet	0.400 - 0.446 μm	Wavelength ↓ 🟪
Blue	0.446 - 0.500 μm	🟦
Green	0.500 - 0.578 μm	🟩
Yellow	0.578 - 0.592 μm	🟨
Orange	0.592 - 0.620 μm	🟧
Red	0.620 - 0.700 μm	🟥

Blue, green and red are the *primary colors* or wavelengths of the visible spectrum. No single primary color can be created from the other two – all other colors can be formed by combining blue, green and red in various proportions. The visible portion of radiation can be shown in its *component colors* when sunlight is passed through a prism. The prism bends the light in differing amounts according to wavelength.

There is a lot of radiation around us which is "invisible" to our eyes, but can be detected by other remote sensing instruments and used to our advantage.

The infrared (IR) region covers the wavelength range from approximately 0.7 μm to 100 μm. Infrared region is more than 100 times as wide as the visible portion. The infrared region can be divided into two categories based on their radiation properties – reflected IR radiation and emitted (or thermal) IR radiation.

Radiation in the reflected IR region (also called *near infrared*) is used for remote sensing purposes in ways very similar to radiation in the visible portion. The reflected IR covers wavelengths from approximately 0.7 µm to 3.0 µm. The thermal IR region is quite different than the visible and reflected IR portions, as this energy is emitted from the Earth's surface in the form of heat. The thermal IR covers wavelengths from approximately 3.0 µm to 100 µm. The thermal IR region is subdivided into mid wavelength IR (MWIR, 3–8 µm), long wavelength IR (LWIR 8–15 µm) and far infrared (FIR 15 µm and longer).

The portion of the spectrum of more recent interest to remote sensing is the microwave region from about 1 mm to 1m. This covers the longest wavelengths used for remote sensing. The shorter wavelengths have properties similar to the thermal infrared region while the longer wavelengths approach the wavelengths used for radio broadcasts.

5.1.3 Interaction of EM-radiation with the environment

Any beam of photons from some source passing through medium 1 (usually air, water or glass) that impinge upon an object or target (medium 2) will experience one or more reactions with the mediums. While interacting with mediums (see Figure 5), EM radiation could be:

- Scattered
  - deflected in all directions
  - occurs on surfaces that are “rough”
  - wavelength of incident energy
- Reflected
  - returned from surface of material with angle of reflection equal and opposite to angle of incidence
  - happens at “smooth” surfaces
- Emitted by the substance
  - usually at longer wavelengths – black body radiation
- Transmitted/refracted
  - passing through a substance
  - change of direction and velocity at boundaries
  - index of refraction n = velocity in vacuum / velocity in substance
- Absorbed
  - giving up energy, usually to heating
  - change of intensity

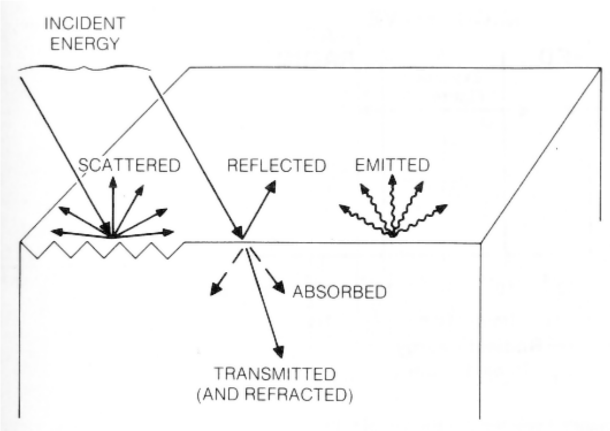


Figure 5. Interaction of EM energy and mediums

In remote sensing one studies how EM energy interacts with the atmosphere (as medium 1) and with objects on the Earth's surface (as medium 2). Both these interactions change the properties of electromagnetic radiation.

5.1.4 Interaction between EM radiation and atmosphere

EM radiation from the energy source travels some distance through the Earth's atmosphere. Particles and gases in the atmosphere can affect the incoming light and radiation. These effects are caused by the mechanisms of scattering and absorption. Ozone, carbon dioxide and water vapour absorb electromagnetic energy in very specific regions of the spectrum. Areas of the spectrum which are not severely influenced by atmospheric absorption, useful to remote sensors, are called *atmospheric windows* (see Figure 6).

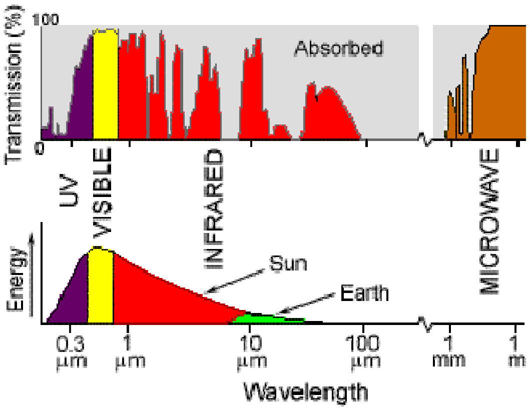


Figure 6. Atmospheric windows

By comparing the characteristics of the two most common energy/radiation sources (the Sun and the Earth) with the atmospheric windows available to us, we can define those wavelengths that we can use most effectively for remote sensing. The visible portion of the spectrum, to which our eyes are most sensitive, corresponds to both an atmospheric window and the peak energy level of the Sun. Note also that heat energy emitted by the Earth corresponds to a window around 10 mm in the thermal IR portion of the spectrum, while the large window at wavelengths beyond 1 mm is associated with the microwave region.

5.1.5 Interaction between EM radiation and the Earth's surface

Radiation that is not absorbed or scattered in the atmosphere can reach and interact with targets on the Earth's surface. There are three forms of interaction that can take place when energy strikes, or is incident (I) upon, the surface (see Figure 7):

- absorption (A)
- transmission (T)
- reflection (R)

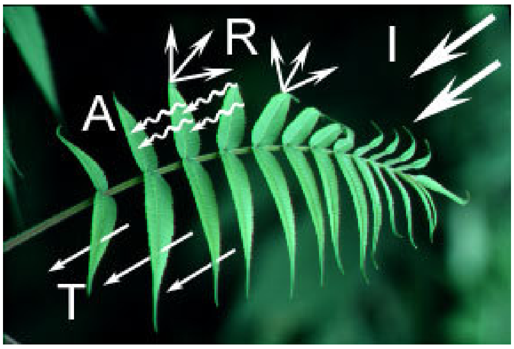


Figure 7. Interaction between EM radiation and Earth's surface

The total incident energy will interact with the surface in one or more ways. The proportions of each will depend on the wavelength of the energy and the material and condition of the feature. Reflection (R) occurs when radiation "bounces" off the target and is redirected. In remote sensing, we are most interested in measuring the radiation reflected from targets.

As an example, we will discuss the interaction of EM radiation and leaves as a target at the Earth's surface (see Figure 8).

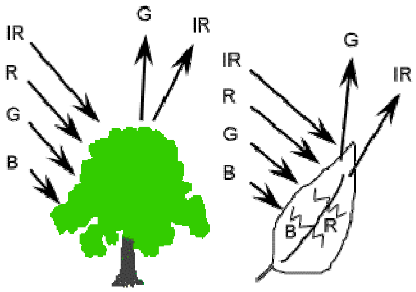


Figure 8. Interaction of EM radiations with leaves

How does energy at the visible and infrared wavelengths interact with foliage? A chemical compound in leaves called chlorophyll strongly absorbs radiation in the red and blue wavelengths but reflects green wavelengths. Leaves appear "greenest" to us in the summer, when chlorophyll content is at its maximum. In autumn, there is less chlorophyll in the leaves, so there is less absorption and more reflection of the red wavelengths, making the leaves appear red or yellow (yellow is a combination of red and green wavelengths).

The internal structure of healthy leaves act as excellent diffuse reflectors of near-infrared wavelengths. If our eyes were sensitive to near-infrared, trees would appear extremely bright to us at these wavelengths. In fact, measuring and monitoring the near-IR reflectance is one way that scientists can determine how healthy (or unhealthy) vegetation may be.

It is important to emphasize here that different objects reflect EM radiation differently in different parts of the spectrum. Figure 9 illustrates reflection of EM radiation by leaves through a portion of the spectrum (notice that the leaf reflects a lot of infrared radiation).

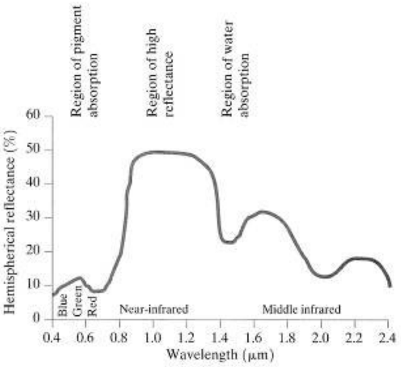


Figure 9. Reflection of EM radiation by leaves

5.1.6 Spectral signatures

As noted above, EM energy behaves very differently to the mechanisms of absorption, transmission and reflection. EM-target interaction depends on the physical nature of the object and the wavelengths of radiation involved.

By measuring the energy that is reflected (or emitted) by targets on the Earth's surface over a variety of different wavelengths, we can build up a spectral response for that object. These response patterns are called spectral signatures. By comparing the spectral signatures of different features we may be able to distinguish between them. In cases where we are unable to do so, we may only compare them at one wavelength.

For example, water and vegetation may reflect somewhat similarly in the visible wavelength band but are almost always separable in the infrared sector (Figure 10). Spectral response can be quite variable, even for the same target type, and can also vary with time (e.g. "greenness" of leaves) and location.

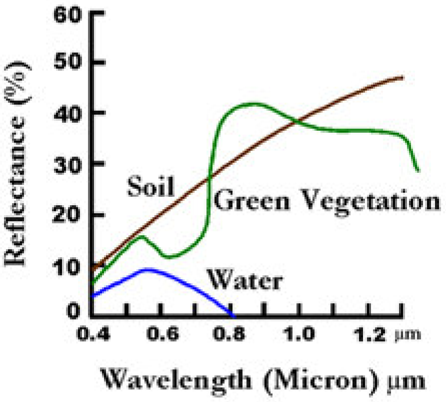


Figure 10. Spectral signatures for water, soil and vegetation

We can see from these examples that, depending on the complex make-up of the target that is being looked at and the wavelengths of radiation involved, we can observe very different responses to the mechanisms of absorption, transmission and reflection. Knowing where to "look" spectrally and understanding the factors which influence the spectral response of the features of interest are critical to correctly interpreting the interaction of electromagnetic radiation with the surface. Each surface has a unique spectral signature or “fingerprint”

5.1.7 Passive and active sensors

In the majority of remote sensing systems the Sun acts a natural source of EM energy. According to the general rules of EM radiation with the mediums, the Sun's energy can be reflected (in visible wavelengths), absorbed and then re-emitted (in thermal infrared wavelengths). How does remote sensing deploy EM radiation? Remote sensing systems measure energy that is naturally available or supplied by the sensor itself. The first systems are called *passive*, and the second one – *active*.

Passive sensors can only be used to detect energy when the naturally occurring energy is available. For all reflected energy, this can only take place during the time when the Sun is illuminating the Earth. There is no reflected energy available from the Sun at night. Energy that is naturally emitted (such as thermal infrared) can be detected day or night, as long as the amount of energy is large enough to be recorded.

Active sensors, on the other hand, provide their own energy source for illumination. The sensor emits radiation which is directed toward the target to be investigated. The radiation reflected from that target is detected and measured by the sensor. Advantages for active sensors include the ability to obtain measurements anytime, regardless of the time of day or season.

Active sensors can be used for examining wavelengths that are not sufficiently provided by the Sun, such as microwaves. Active systems require the generation of a fairly large amount of energy to adequately illuminate targets. Examples of active sensors are LiDAR (laser systems) and SAR (synthetic aperture radar systems).

Figure 11 summarizes the three models (types of sensors) for remote sensing discussed above.

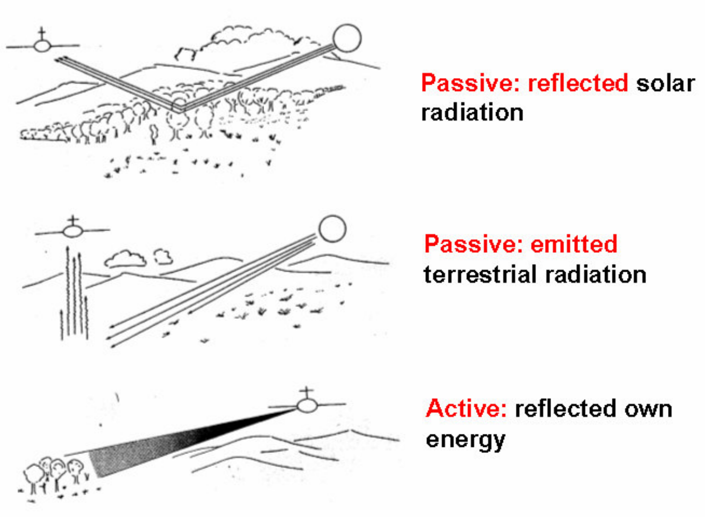

Passive: reflected solar radiation
Passive: emitted terrestrial radiation
Active: reflected own energy

Figure 11. Three models (types of sensors) for remote sensing

5.2 Sensors in remote sensing

This section is dedicated to sensors, used in remote sensing. How do remote sensing systems differ from each other? They can be installed on different platforms (ground, aerial, space). They use different principles and techniques for image acquisition and different media to register acquired radiation. Sensors operate in different spectral zones and utilize external or internal sources of energy. Finally, acquired images differ in spatial, spectral and temporal resolutions.

5.2.1 Sensors and platforms

Modern remote sensing systems can be installed on ground platforms (vehicles), aircraft or satellites. All these platforms are specifically designed and manufactured to carry the following common types of sensors:

- Photographic film cameras (frame)
- Digital cameras (frame, scanning)
- Multi-spectral digital cameras (frame, scanning)
- Radar (scanning)

Photographic film cameras are framing systems which acquire a "snapshot" of an area of the Earth’s surface. Camera systems use a lens or system of lenses (collectively referred to as the optics) to form an image at the focal plane – the plane at which a film is placed and the image is registered. There are two main types of aerial photographs: vertical and oblique. Vertical aerial photographs are most accurate for mapping, while oblique photographs are very convenient for image interpretation.

When obtaining vertical aerial photographs, the aircraft normally flies in a series of lines (Figure 12).

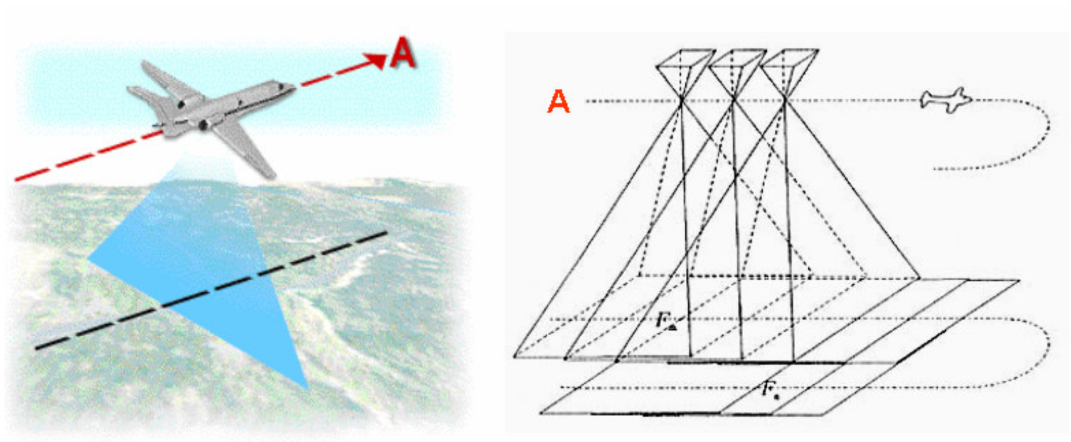


Figure 12. Flight lines in aerial photography

Aerial photos are usually taken in a way to make a certain overlap between photos. Two consequent aerial photographs form a stereopair allows us to perceive a stereo model within an overlapping area. The forward overlapping is usually 60% of the image area along the flight direction to ensure the best stereoscopic viewing; lateral (side) overlaps are usually 20 to 40% (Figures 13 and 14).

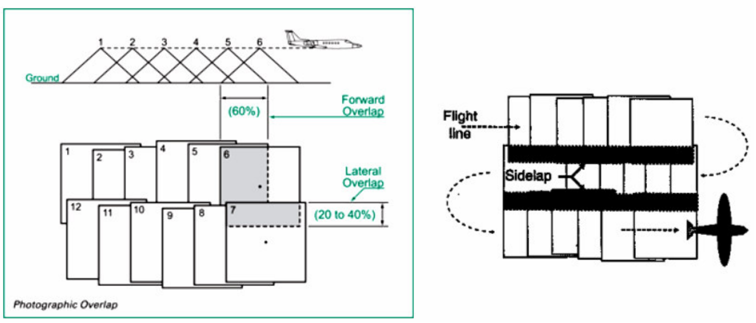


Figure 13. Aerial photography: overlaps

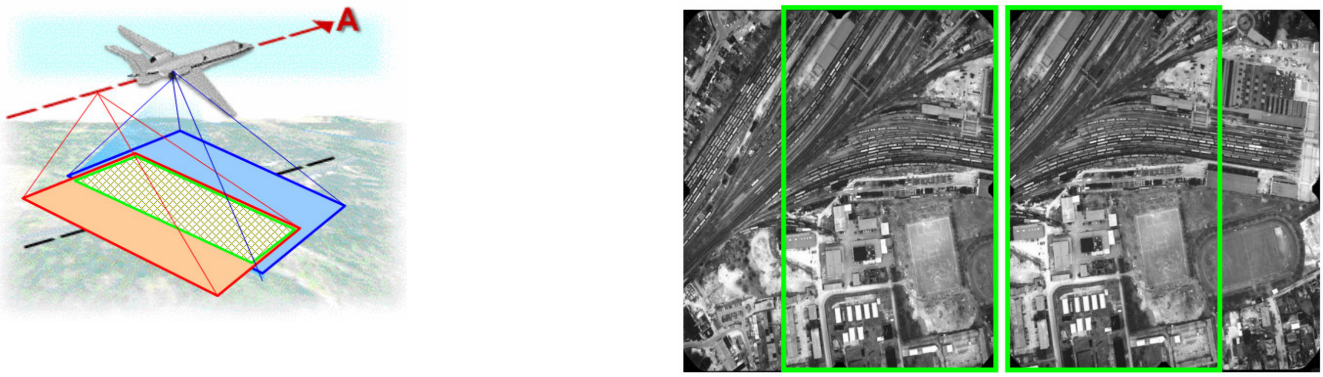


Figure 14. Stereopairs of aerial photographs

Single image allows us to measure (X,Y) coordinates of an object, while stereopairs of aerial photographs allows us to see 3D models and define (X,Y,Z) coordinates of any object within overlapping areas. Stereopairs are widely used to support image interpretation. Successive photo pairs display the overlap region from different perspectives and can be viewed through a device called a stereoscope where one can view a three-dimensional image of the area – generally referred to as stereo model.

5.2.2 Registering EM radiation as an image: photographic and digital cameras

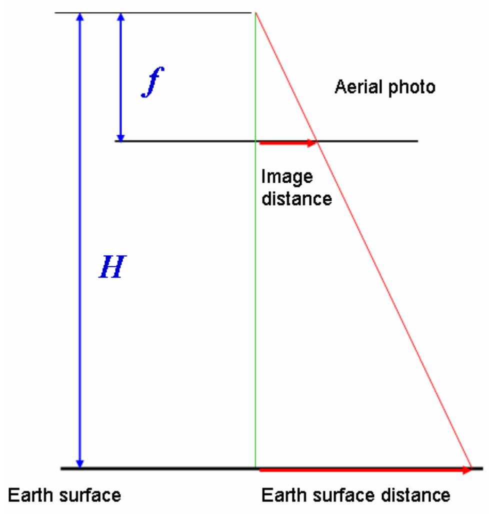
EM energy, reflected by an object and traveling through the atmosphere to a sensor, can be registered as an image. This section introduces basic principles of acquisition of electromagnetic energy by two main types of sensors: analogue films and digital sensors.

Analogue films are used in *aerial (* later, in *satellite) photographic cameras*. There are several types of films used in aerial photography:

- Panchromatic ("black and white")
- Color (natural colors, as seen by humans)
- Infrared (register thermal energy, non-visible to the human eye)

The typical aerial photograph is panchromatic (or more recently, color) square 9"x9" (23x23 cm) picture. Photos can be is taken by aerial cameras with different focal lengths (6" = 152mm is typical) and at different flight altitudes. A combination of focal length and altitude defines the scale of an aerial photograph.

The scale of a vertical photograph is approximately equal to the ratio of the flying height above the ground ***H*** and the focal length of the camera lens ***f*** (Figure 15).



$$Scale = \frac{imageDist}{surfaceDist} = \frac{f}{H}$$

Figure 15. Scale of aerial photograph

The scale of aerial photographs greatly impacts the appearance of surface details in the image. More details of the same phenomena are seen on large scale photos; on the smaller scale photos one sees more generalized images of phenomena over a larger geographic area (Figure 16). Different scales require the use of different identification features for image interpretation.

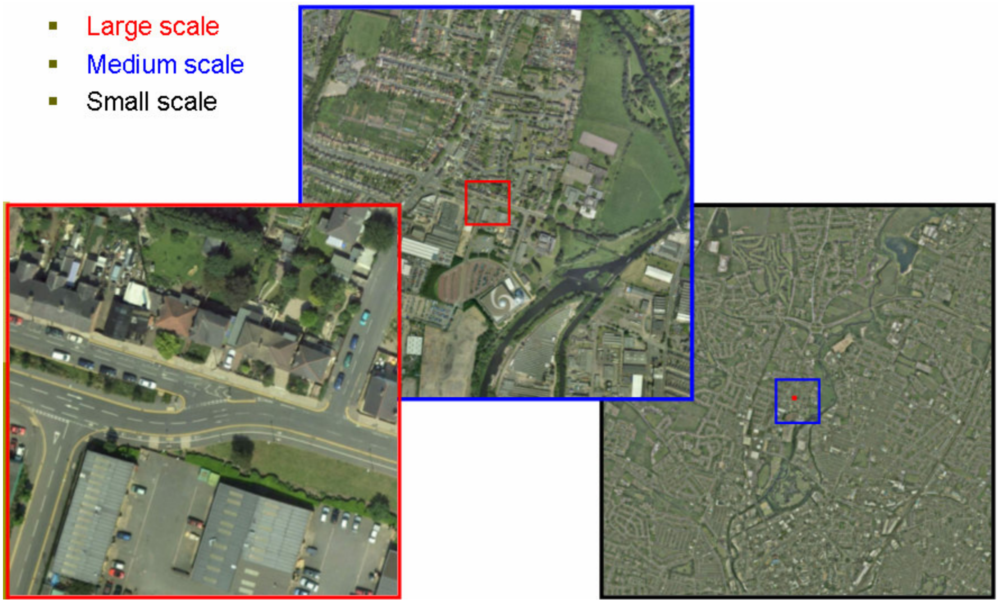


Figure 16. Scale of image and appearance of objects on the ground

Aerial photographs are created using a central (perspective) projection. Therefore, the geometry of the objects depicted in the image is burdened by certain forms of distortions. Major image distortions are caused by surface relief, tilt of an aerial camera and lens distortions. If the amount and direction of distortions are known, then the photo may be corrected (rectified).

Instead of using a film, *digital cameras* use a grided array of silicon coated CCDs (charge-coupled devices) that individually respond to electromagnetic radiation. The CCD elements equate to pixels on the ground.

Energy reaching the surface of the CCDs causes the generation of an electronic charge which is proportional in magnitude to the "brightness" of the ground area. A digital number for each spectral band is assigned to each pixel based on the magnitude of the electronic charge.

The digital format of the output image is amenable to digital analysis and archiving in a computer environment, as well as output as a hardcopy similar to regular photos. Digital cameras also provide quicker turn-around for acquisition and retrieval of data and allow greater control of the spectral resolution. The size of the pixel arrays (matrix) varies between systems. A typical custom camera has an array of 2560 x 1920 elements, equal to approximately 5 megapixels. This is changing rapidly as digital camera technology improves.

5.2.3 Registering EM radiation as an image: scanning systems

Many electronic (as opposed to photographic) remote sensors acquire data using scanning systems, which employ a sensor with a narrow field of view (i.e. IFOV) that sweeps over the terrain to build up and produce a two-dimensional image of the surface. Scanning systems can be used on both aircraft and satellite platforms and have essentially the same operating principles. A scanning system used to collect data over a variety of different wavelength ranges is called a multi-spectral scanner (MSS). This is the most commonly used scanning system today.

There are two main modes or methods of scanning employed to acquire multi-spectral image data: across-track scanning and along-track scanning. Across-track scanners scan the Earth in a series of lines. The lines are oriented perpendicular to the direction of motion of the sensor platform (i.e. across the swath). Figure 17 illustrates the principles of across-track and along-track scanning.

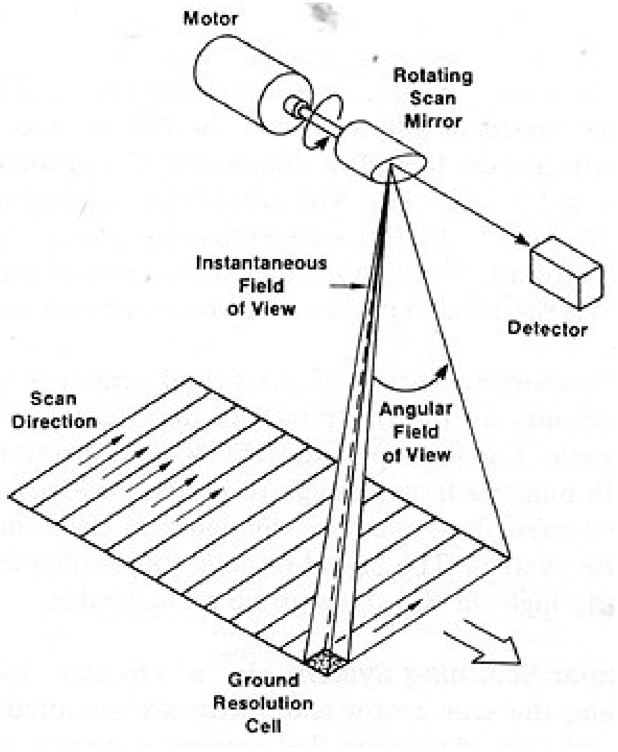


Figure 17. Across-track scanning

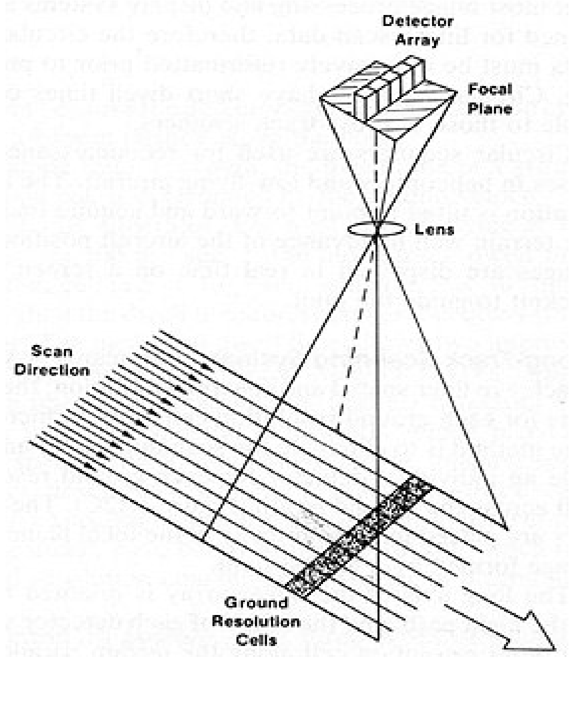


Figure 18. Along-track scanning

In across-track scanning each line is scanned from one side of the sensor to the other, using a rotating scan mirror. As the platform moves forward over the Earth, successive scans build up a two-dimensional image of the Earth's surface. The incoming reflected or emitted radiation is separated into several spectral components that are detected independently. The UV, visible, near-infrared, and thermal radiation are dispersed into their constituent wavelengths. A bank of internal detectors, each sensitive to a specific range of wavelengths, detects and measures the energy for each spectral band and then, as an electrical signal, they are converted to digital data and recorded for subsequent computer processing.

Along-track scanners (Figure 18) also use the forward motion of the platform to record successive scan lines and build up a two-dimensional image, perpendicular to the flight direction. However, instead of a scanning mirror, they use a linear array of detectors located at the focal plane of the image formed by the lens systems, which are "pushed" along in the flight track direction (i.e. along track).

These systems are also referred to as push-broom scanners, as the motion of the detector array is analogous to the bristles of a broom being pushed along a floor. Each individual detector measures the energy for a single ground resolution cell and thus the size and Instantaneous Field of View (IFOV) of the detectors determines the spatial resolution of the system. A separate linear array is required to measure each spectral band or channel. For each scan line, the energy detected by each detector of each linear array is sampled electronically and digitally recorded.

The following table provides a summary of photographic frame cameras versus scanning systems:

Photographic camera	**Scanner**
Image is a solid frame	Image is a set of lines, each line comprises separate pixels
The whole frame is acquired at once	The frame is acquired as accumulation of lines due to movement of the sensor
Strong geometry - central projection	Weak geometry - non-central projection
Stored as a film, but could be scanned into a digital form for further use	Stored as digital array (matrix)

5.2.4 Multi-spectral imagery in Remote Sensing

Many remote sensing applications are based on the use of multi-spectral imagery. Multi-spectral photography uses multi-lens systems with different film-filter combinations to acquire photos simultaneously in a number of different spectral ranges.

The advantage of these types of cameras is their ability to record reflected energy separately in discrete wavelength ranges, thus providing potentially better separation and identification of various features. However, simultaneous analysis of these multiple photographs can be problematic.

The whole concept of multi-spectral image analysis is based on fundamental features of color composition and decomposition. The palette of all natural colors is represented as a combination of primary colors: red, green and blue. All other colors are formed by combining color primaries in various proportions. Color images could be de-composed into three separate panchromatic images, where the density of photo tone corresponds to amount of energy, registered in red, green and blue visible zones. In the opposite, if one registers EM energy in red, green and blues zones as three separate images and then combines them, applying corresponding color filters, these three primary colors will produce a full color image (Figures 19 and 20).

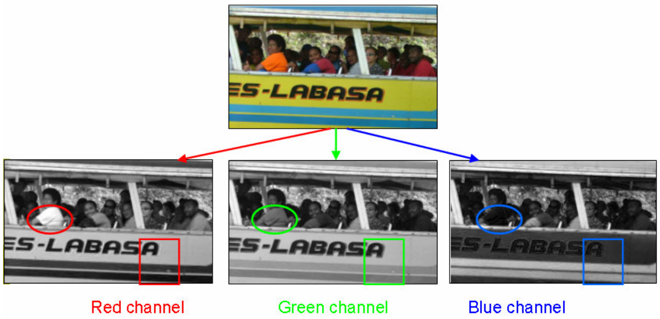


Figure 19. Color image de-composition

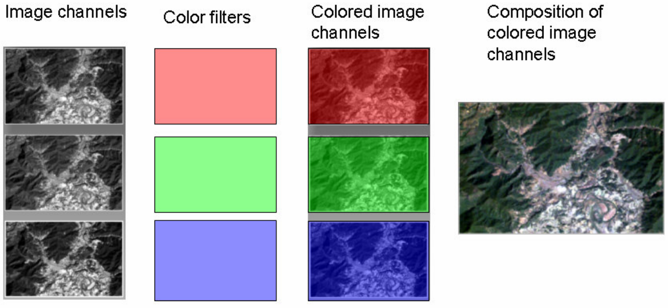


Figure 20. Color image composition

We see color because our eyes detect the entire visible range of wavelengths and our brains process the information into separate colors. Can you imagine what the world would look like if we could only see very narrow ranges of wavelengths or colors? That is how many sensors work. The information from a narrow wavelength range is gathered and stored in a *channel*, also sometimes referred to as a *band*. As discussed, we can combine and display channels of information digitally using the three primary colors (blue, green, and red). The data from each channel are represented as one of the primary colors and, depending on the relative brightness (i.e. the digital value) of each pixel in each channel, the primary colors combine in different proportions to represent different colors. Figure 21 illustrates this concept using seven channels of Landsat satellite system.

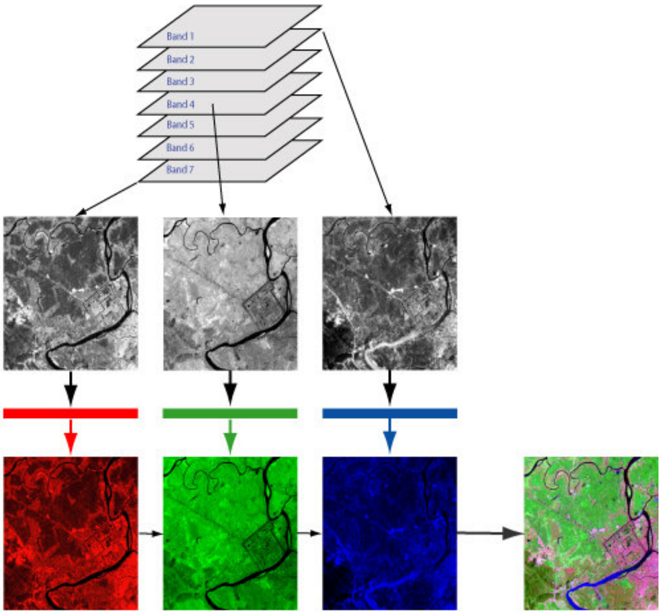


Figure 21. Combining channels of Landsat satellite system

Modern remote sensors can detect EM radiation in as few as one band to as many as hundreds of bands, but we can use only three primarily colors to make color combinations. The more channels in a sensor, the higher the number of color combinations. If a color, applied to a channel, corresponds to a spectral zone, such combinations produce a “true color image”.

Mixing three primary colors and channels produces a variety of pseudo-color or false-color composites.

5.2.5 Thermal, video , RADARs and LIDARs

Thermal sensors. Many multi-spectral (MSS) systems sense radiation in the thermal infrared as well as the visible and reflected infrared portions of the spectrum. However, remote sensing of energy emitted from the Earth's surface in the thermal infrared (3 µm to 15 µm) is different than the sensing of reflected energy.

Thermal sensors use photo detectors sensitive to the direct contact of photons on their surface, to detect emitted thermal radiation. The detectors are cooled to temperatures close to absolute zero in order to limit their own thermal emissions. Thermal sensors essentially measure the surface temperature and thermal properties of targets.

Video sensors. Although coarser in spatial resolution than traditional photography or digital imaging, video cameras provide a useful means of acquiring timely and inexpensive data. Applications with these requirements include natural disaster management (fires, flooding), crop and disease assessment, environmental hazard control, police surveillance and a host of other practical applications.

Cameras used for video recording measure radiation in the visible, near infrared, and sometimes mid-infrared portions of the EM spectrum. Acquired from an aerial platform, the image data are recorded onto cassette, and can be viewed immediately. Digital Video Technology is also developing rapidly and offers powerful mapping and monitoring potential.

RADAR stands for RAdio Detection And Ranging. RADAR systems are active sensors which provide their own source of electromagnetic energy. Radar sensors, whether airborne or space-borne, emit microwave radiation in a series of pulses from an antenna, looking obliquely at the surface perpendicular to the direction of motion.

When the energy reaches the target, some of the energy is reflected back towards the sensor. This backscattered microwave radiation is detected, measured and timed. By recording the range and magnitude of the backscatter from all targets as the system passes by, a two-dimensional image of the surface can be produced. Because RADAR provides its own energy source, images can be acquired day or night. Also, microwave energy is able to penetrate through clouds and most rain, making it an all-weather sensor.

LIDAR is an acronym for LIght Detection And Ranging, an active imaging technology very similar to RADAR. Pulses of laser light are emitted from the sensor and energy reflected from a target is detected. The time required for the energy to reach the target and return to the sensor determines the distance between the two. LIDAR is used effectively for measuring heights of features, such as buildings or forest canopy height relative to the ground surface, and water depth relative to the water surface (laser profile-meter).

5.2.6 Satellites and sensors for environmental studies

There are a number of satellite and airborne platforms and systems acquiring terabytes of image information about the Earth. The main groups of satellite sensors for environmental studies include:

- Land observation (Landsat, SPOT, IRS, JERS, Ikonos, QuickBird, etc.)
- Meteorological observation (NOAA AVHRR, METEOSAT, GOES, etc.)
- Marine observation (Nimbus, MOS, SeaWiFs, etc.)

5.3 Topic 3: Image Analysis in Remote Sensing

This section provides an introduction to images and image analysis in remote sensing. How do remotely sensed images differ from each other? Sections 1 and 2 review the basics of image acquisition and registration, spectral channels and image scales. Here we concentrate on another important characteristic of remotely sensed data – image resolution. We will introduce concepts of spatial, spectral and temporal resolution. In order to become valuable information, image data should be pre-processed, enhanced and treated using certain transformations. Finally, we briefly outline the fundamental principles of visual and computer-assisted image analysis in remote sensing.

5.3.1 How do remotely sensed images differ from each other?

Images, acquired by different remote sensing systems differ by

- Image scale (large, medium, small)
- Image geometry acquisition (frame, scanning)
- Radiation registration (analogue film, digital sensors)
- Wavelengths of operation (visible zone, infrared, microwave)
- Number of channels (single channel [panchromatic], three channel [color], multi-channel [multi-spectral])
- Resolution (spatial, spectral, radiometric, temporal)

5.3.2 Image resolution

Within the framework of remote sensing, there are several kinds of data resolutions:

- Spatial resolution: smallest discernable physical object or detail in image
- Spectral resolution: “width” of bands in micrometers (minimum and maximum wavelength sensed)
- Spectral coverage: number of bands.
- Radiometric resolution: difference between minimum and maximum reading or measurement. Number of steps between minimum and maximum reading or measurement
- Temporal resolution: time between repeated sensing

Spatial resolution of sensors

Remote sensing as a multi-scale data capture system, provides data at global, continental, regional and local levels. The detail distinguishable in an image is dependent on the spatial resolution of the sensor and refers to the size of the smallest possible feature that is detected. Spatial resolution of passive sensors depends primarily on their Instantaneous Field of View (IFOV). The IFOV is the angular cone of visibility of the sensor (A) and determines the area on the Earth's surface that is "seen" from a given altitude at one particular moment in time (B). The size of the area viewed is determined by multiplying the IFOV by the distance from the ground to the sensor (C) (see Figure 22).

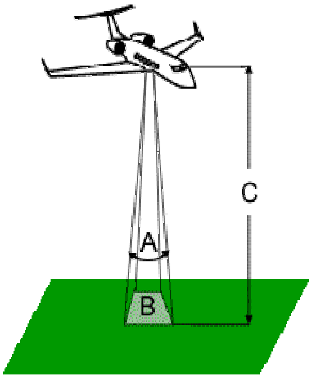


Figure 22. Spatial resolution

Most remote sensing images are composed of a matrix of picture elements, or pixels (picture elements), which are the smallest units of an image. Image pixels are normally square and represent a certain area on the ground or target. We usually say a sensor's spatial resolution is equivalent to its pixel size. This can be different to "image resolution" where we may have degraded the pixel size of an image. For a homogeneous feature to be detected, its size generally has to be equal to, or larger than, the resolution cell. Identification may still not be possible. If the feature is smaller than this, it may not be detectable as the average brightness of all features in that resolution cell will be recorded. However, smaller features may sometimes be detectable if their reflectance dominates within a particular pixel allowing for sub-pixel detection.

Spectral resolution

Broad classes of objects on the Earth, such as water and vegetation cover, can usually be separated using very broad wavelength ranges - the visible and near infrared. Other more specific classes, such as different rock types, may not be easily distinguishable using either of these broad wavelength ranges and would require analysis at much finer wavelength ranges to separate them. Thus, we would require a sensor with higher spectral resolution. Spectral resolution describes the ability of a sensor to discriminate the finer wavelength intervals (Figure 23).

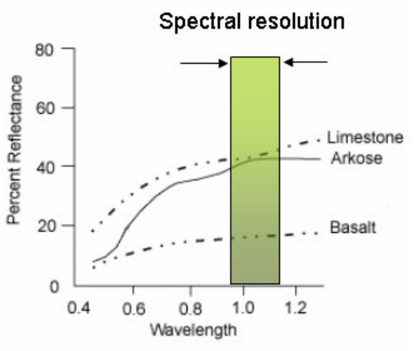


Figure 23. Spectral resolution

The finer the spectral resolution, the narrower the wavelength range for a particular channel or band. Spectral signatures characterize the reflectance of a feature or target over a variety of wavelengths. Different classes of features and details in an image can often be distinguished by comparing their responses over distinct wavelength ranges.

Spectral resolution should be distinguished from spectral coverage. Sensors record reflected light over a range of wavelengths, gathered and stored in channels, sometimes referred to as bands (see Topic 2). The number and distribution of these bands determines the spectral coverage. For example, Landsat 4 Band 1 has a spectral resolution of 0.1 micron, but its spectral coverage is 0.5 – 1.1 microns, registered in 4 channels (bands).

Many remote sensing systems record energy over several separate wavelength ranges at various spectral resolutions. If a sensor has acquired data using less than 10 bands, these systems are referred to as multi-spectral sensors. Advanced multi-spectral sensors, called hyper-spectral sensors, detect hundreds of very narrow spectral bands throughout the visible, near-infrared and mid-infrared portions of the electromagnetic spectrum. Their very high spectral resolution facilitates fine discrimination between different targets based on their spectral response in each of the narrow bands.

The concept of spectral resolution is also applied to photographic cameras that use analogue films to record EM energy. Black and white (panchromatic) film records wavelengths extending over most, if not all, of the visible portion of the electromagnetic spectrum. Its spectral resolution is fairly coarse, as the various wavelengths of the visible spectrum are not individually distinguished and the overall reflectance in the entire visible portion is recorded. Color film is also sensitive to the reflected energy over the visible portion of the spectrum, but has higher spectral resolution than panchromatic photography. It is individually sensitive to the reflected energy at the blue, green and red wavelengths of the spectrum. Thus, it can represent features of various colors based on their reflectance in each of these distinct wavelength ranges.

Radiometric resolution

The sensitivity of film or a sensor to the magnitude of the electromagnetic energy determines the radiometric resolution. This describes the ability of an imaging system to discriminate very slight differences in reflectance. The finer the radiometric resolution of a sensor, the more sensitive it is to detecting small differences in reflected or emitted energy.

The radiometric resolution of digital sensors can be expressed by the number of brightness levels in any one band. Image data are represented by positive digital numbers that can vary from 0 to a selected power of 2. If a sensor used 8 bits to record the data, there would be $2^8 = 256$ digital values available, ranging from 0 to 255. However, if only 4 bits were used, then only $2^4 = 16$ values ranging from 0 to 15 would be available. Therefore, the radiometric resolution is less. Image data are generally displayed in a range of gray tones, with black representing a digital number of 0 and white representing the maximum value (for example, 255 in 8-bit data).

Temporal resolution

Temporal resolutions refer to the length of time it takes for a satellite to return and image the same point on the Earth. The revisit period of a satellite sensor is usually several days, however, because of some degree of overlap in the imaging swaths of adjacent orbits for most satellites and the increase in this overlap with increasing latitude, some areas of the Earth tend to be re-imaged more frequently. Some satellite systems are able to point their sensors to image the same area between different satellite passes separated by periods from one to five days. Thus, the actual temporal resolution of a sensor depends on a variety of factors, including the satellite/sensor capabilities, the swath overlap and the latitude.

In many cases the time factor in remote sensing can be important and defined by many factors. For example, persistent clouds offer limited clear views of the Earth's surface (often in the tropics); frequency and duration of an event for short-lived phenomena (floods, oil slicks, etc.)

which need to be imaged; multi-temporal comparisons (e.g. the spread of a forest disease from one year to the next); and, the changing appearance of a feature over time can be used to distinguish it from near-similar features (cropping versus pasture seeding).

5.3.3 Remote Sensing: from data to information

Most remote sensing systems supply the imagery in a digital form which makes it possible to use different computer algorithms to process such data. Analogue images, acquired by cameras and stored as films, can be converted into a digital form using desktop scanners, and thus, can be also processed in computers.

For discussion purposes, most of the common image processing functions available in image analysis systems can be categorized into the following four categories:

- Image Pre-processing
- Image Enhancement
- Image Transformation
- Image Classification and Information Extraction

Pre-processing functions involve those operations that are normally required prior to the main data analysis and extraction of information:

- Radiometric corrections include correcting the data for sensor irregularities and unwanted sensor or atmospheric noise, and converting the data so they accurately represent the reflected or emitted radiation measured by the sensor.
- Geometric corrections include correcting for geometric distortions due to sensor-Earth geometry variations and conversion of the data to real world coordinates – geo-referencing (e.g. latitude and longitude) on the Earth's surface.

The objective of the second group of image processing functions, grouped under the term of image enhancement, is solely to improve the appearance of the imagery to assist in visual interpretation and analysis. Examples of enhancement functions include contrast stretching to increase the tonal distinction between various features in a scene and spatial filtering to enhance (or suppress) specific spatial patterns in an image.

Image transformations are operations similar in concept to those for image enhancement. However, unlike image enhancement operations that are normally applied only to a single channel of data at a time, image transformations usually involve combined processing of data from multiple spectral bands. Arithmetic operations (i.e. subtraction, addition, multiplication, division) are performed to combine and transform the original bands into "new" images that better display or highlight certain features in the scene. Some of these operations include various methods of spectral or band rationing and involve a procedure called “principal components analysis” which is used to more efficiently represent the information in multi-channel imagery.

Image classification operations are used to digitally identify and classify pixels in the data. Classification is usually performed on multi-channel data sets and this process assigns each pixel in an image to a particular class or theme, based on statistical characteristics of the pixel brightness values. There are a variety of approaches that can be taken to perform digital classification. We will briefly describe the two generic approaches used most often, namely, supervised and unsupervised classification.

5.3.4 Image pre-processing

Pre-processing operations, sometimes referred to as image restoration and rectification, are intended to correct for sensor- and platform-specific radiometric and geometric distortions of data. As any image involves radiometric errors, as well as geometric errors, these errors should be corrected. Radiometric correction avoids radiometric errors or distortions, while geometric correction removes geometric distortion.

Radiometric corrections may be necessary due to variations in scene illumination and viewing geometry, atmospheric conditions and sensor noise and response. When the emitted or reflected electro-magnetic energy is observed by a sensor on board an aircraft or spacecraft, the observed energy does not coincide with the energy emitted or reflected from the same object observed from a short distance. This is due to the Sun's azimuth and elevation, atmospheric conditions, such as fog or aerosols, and the sensor's response, etc. Therefore, in order to obtain the real irradiance or reflectance, those radiometric distortions must be corrected.

All remote sensing imagery is inherently subject to geometric distortion. These distortions may be due to several factors: the perspective of the sensor optics; the motion of the scanning system; the motion of the platform; the platform altitude, attitude, and velocity; the terrain relief; and, the curvature and rotation of the Earth. Geometric corrections are intended to compensate for these distortions so that the geometric representation of the imagery will be as close as possible to the real world. Many of these variations are systematic, or predictable, in nature and can be accounted for by accurate modeling of the sensor and platform motion and the geometric relationship of the platform with the Earth. Other unsystematic, or random, errors cannot be modeled and corrected in this way. Therefore, geometric registration of the imagery to a known ground coordinate system must be performed.

There are three main geometric transformations, applied to remotely sensed images:

- Resampling: change pixel size and orientation
- Rectification: change pixel position in image
- Geo-referencing: assign world coordinates for each pixel in image

Geo-referencing is the process of scaling, rotating, translating and de-skewing the image to match a particular size and position. This process involves identifying the image coordinates (i.e. row, column) of several clearly discernible points, called ground control points (GCPs), in the distorted image, and matching them to their true positions in ground coordinates (e.g. latitude, longitude). The true ground coordinates are typically measured from a map, either in paper or digital format. This is image-to-map registration. Once several well-distributed GCP pairs have been identified, the coordinate information is processed by the computer to determine the proper transformation equations to apply to the original (row and column) image coordinates to map them into their new ground coordinates. Geometric registration may also be performed by registering one (or more) images to another image, instead of two geographic coordinates. This is called image-to-image registration and is often done prior to performing various image transformation procedures, which will be discussed later, or for multi-temporal image comparison.

Image transformation, when geo-referencing, involves linear (affine) and non-linear (second and third order polynomials) coordinate transformations. Once cells have been moved as a result of such transformation, it is necessary to determine what value each cell should have. This can be done using a procedure called re-sampling. Re-sampling can be implemented using several algorithms. The three common one's are: nearest neighbour, cubic convolution and bilinear interpolation.

The nearest neighbour method matches the output cell centre to the nearest input cell centre and transfers the input cell value. This method is appropriate for discrete data and in some situations for continuous data. It is primarily used in land use classifications where data is categorized and where values within cells does not change. The maximum spatial error for this method of re-sampling will be one half of the cell size.

Bilinear interpolation determines the output cell value with a weighted distance average of the four nearest input cell centres. This method is appropriate for continuous data, but not for discrete data because values are averaged, and hence the cell values may be altered. If the input grid is integer, then the output values are truncated to integer. This option will cause some smoothing of the data.

Cubic is similar to bilinear interpolation, except that the nearest 16 cells are used. Like bilinear, cubic is appropriate for continuous data, but not for discrete data. This technique will generate a slightly sharper grid than through bilinear interpolation. The grid will be geometrically less distorted than the grid achieved by running the nearest neighbour re-sampling algorithm.

5.3.5 Image enhancement

Enhancements are used to make it easier for visual interpretation and understanding of imagery. The advantage of digital imagery is that it allows us to manipulate the digital pixel values in an image. Although radiometric corrections for illumination, atmospheric influences and sensor characteristics may be determined prior to distributing data to the user, the image may still not be optimized for visual interpretation. Remote sensing sensors, particularly those operated from satellite platforms, must be designed to cope with levels of target/background energy – a condition that is commonly encountered in routine use. With large variations in spectral response from a diverse range of targets (e.g. forest, deserts, snowfields, water, etc.) no generic radiometric correction could optimally account for, and display, the optimum brightness range and contrast for all targets. Thus, for each application and each image, a custom adjustment of the range and distribution of brightness values is usually necessary. These operations are referred to as contrast enhancements.

In raw imagery, the useful data often populates only a small portion of the available range of digital values (commonly 8 bits, which is 2^8 or 256 levels). Contrast enhancement involves changing the original values so that more of the available range is used, thereby increasing the contrast between targets and their backgrounds.

Image histograms and histogram manipulations are key concepts in image enhancement. A histogram is a graphical representation of the brightness values that comprise an image. The brightness values (i.e. 0-255) are displayed along the x-axis of the graph. The frequency of occurrence of each of these values in the image is shown on the y-axis (see Figure 24).

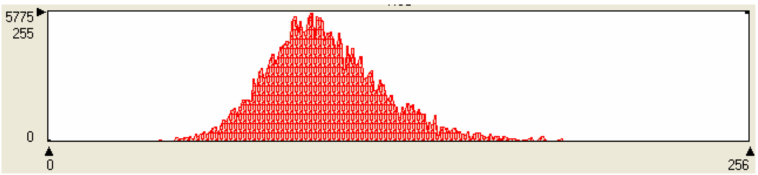


Figure 24. Image histogram

Contrast enhancement expands the original digital values of the remotely-sensed data into a new distribution. By expanding the original input values of the image, the total range of sensitivity of the display device can be utilized. Linear contrast enhancement also makes subtle variations within the data more obvious. These types of enhancements are best applied to remotely-sensed images with Gaussian or near-Gaussian histograms, meaning, all the brightness values fall within a narrow range of the histogram and only one mode is apparent.

There are three methods of linear contrast enhancement:

- Minimum-maximum linear contrast stretch: The original minimum and maximum values of the data are assigned to a newly specified set of values that utilize the full range of available brightness values.
- Piecewise linear contrast stretch: When the distribution of a histogram in an image is bi- or tri-modal, an analyst may stretch certain values of the histogram for increased enhancement in selected areas
- Percentage linear contrast stretch: Is similar to the minimum-maximum linear contrast stretch except this method uses specified minimum and maximum values that lie in a certain percentage of pixels from the mean of the histogram.

Histogram equalization is another popular contrast stretching technique which makes most efficient use of grey levels. As a result of equalization, all pixel values of the image are redistributed so there are approximately an equal number of pixels to each of the user-specified output grey-scale classes and contrast is increased at the most populated range of brightness values of the histogram (or "peaks"). Equalization automatically reduces the contrast in very light or dark parts of the image associated with the tails of a normally distributed histogram.

5.3.6 Image transformations

Image transformation is what gives digital remote sensing its edge. Much of the remote sensing literature concerns spatial, spectral and radiometric enhancement and transformations of images. At the extreme, enhancement allows us to fit the relevant information from hundreds of spectral bands into the three bands available to our eyes, and transform non-spectral information (spatial or temporal) data into those same channels.

Major image transformations include spectral transformations (Image Algebra, Principal Component Analysis (PCA), Tasseled Cap) and spatial transformations, implemented with different filtering techniques. Image transformations typically involve the manipulation of multiple bands of data, whether from a single multi-spectral image or from two or more images of the same area acquired at different times (i.e. multi-temporal image data). Either way, image transformations generate "new" images from two or more sources that highlight particular features or properties of interest, better than the original input images.

Basic image transformations in Image Algebra apply simple arithmetic operations to the image data. Image subtraction is often used to identify changes that have occurred between images collected on different dates. Typically, two images which have been geometrically registered are used with the pixel (brightness) values in one image being subtracted from the pixel values in the other. Scaling the resultant image by adding a constant to the output values will result in a suitable 'difference' image. In such an image, areas where there has been little or no change between the original images, will have resultant brightness values around mid-grey tones, while those areas where significant change has occurred will have values brighter or darker depending on the 'direction' of change in reflectance between the two images . This type of image transformation can be useful for mapping changes in urban development around cities and for identifying areas where deforestation is occurring.

Image division or spectral ratioing is one of the most common transforms applied to image data. Image ratioing serves to highlight subtle variations in the spectral responses of various surface covers. By ratioing the data from two different spectral bands, the resultant image enhances variations in the slopes of the *spectral signature curves* between the two different spectral ranges that could otherwise be masked by the pixel brightness variations in each of the bands.

One widely used image transformation is the Normalized Difference Vegetation Index (NDVI) which has been used to monitor vegetation conditions on continental and global scales. The NDVI is based on the fact that healthy vegetation strongly reflects in the near-infrared zone (NIR) and strongly absorbs in the visible red. At the same time, other surface types, such as soil and water, equally reflect in both the near-infrared and red zones.

Normalized Difference Vegetation Index is calculated using the following formulae:

NDVI = (NIR-R)/(NIR+R),

that, for vegetation, gives ratio values significantly more than 1.0, and for soil and water approximately equal to 1.0. Thus, the NDVI enables the discrimination of vegetation from other types of surface covers. We can also better identify areas of unhealthy or stressed vegetation, which show low near-infrared reflectance, as the ratios would be lower than for healthy green vegetation.

5.3.7 Image interpretation

Acquired, pre-processed and enhanced, remotely-sensed data can be already very valuable and useful, but the user can benefit much more by implementing the further processing. Value-added information can be extracted by implementing the following forms of image analysis:

- *Classification* is a type of categorization of image data using spectral, spatial and temporal information
- *Change detection* is the extraction of change between multi-date images
- *Extraction of physical quantities* corresponds to the measurement of temperature, atmospheric constituents, elevation, and so on, from spectral or stereo information
- *Extraction of indices* is the computation of a newly defined index, for example, the vegetation index from satellite data
- *Identification of specific features* is the identification, for example, of disaster, lineament, archaeological, and other features, etc.

Information extraction can be made by human methods (visual analysis) or computer (digital analysis). Visual and digital analyses of remote sensing imagery are not mutually exclusive and both have their merits. In most cases, a mix of both methods is usually employed when analyzing imagery. In fact, the ultimate decision of the utility and relevance of the information extracted at the end of the analysis process must still be made by humans.

Image interpretation is defined as the extraction of qualitative and quantitative information in the form of a map, about the shape, location, structure, function, quality, condition, relationship of and between objects, etc. by using human knowledge or experience. In some approaches the image is used firstly to generate tabular data which characterize certain parameters of depicted objects, which then applied for actual delineation of objects in the image. As a narrow definition, "photo-interpretation" is sometimes used as a synonym of image interpretation. Manual image interpretation is often called “visual” and the term “image classification” is used when computer-assisted methods are employed. The typical process of visual image interpretation is outlined in Figure 25.

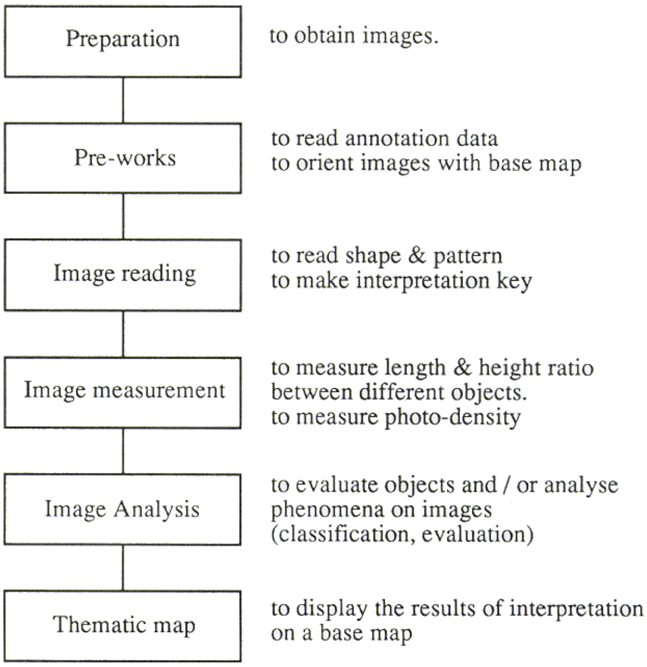

Preparation
to obtain images.
Pre-works
to read annotation data
to orient images with base map
Image reading
to read shape & pattern
to make interpretation key
Image measurement
to measure length & height ratio
between different objects.
to measure photo-density
Image Analysis
to evaluate objects and / or analyse
phenomena on images
(classification, evaluation)
Thematic map
to display the results of interpretation
on a base map

Figure 25. Visual image interpretation workflow

There are several commonly agreed to key elements of visual image interpretation: tone (or color), size, shape, texture, pattern, shadow, site and association.

Tone is the relative brightness or colour of objects in an image. Generally, tone is the fundamental element for distinguishing between different targets or features. Variations in tone also allow the elements of shape, texture and pattern of objects to be distinguished.

Size is a function of scale. It is Important to assess size of a target relative to other objects in a scene, as well as the absolute size, to aid in the interpretation of that target. A quick approximation of target size can direct interpretation to an appropriate result more quickly. For example, large buildings, such as factories or warehouses, suggest commercial property; small buildings indicate residential use.

Shape characterizes general form, structure or outline of individual objects. Straight-edged shapes typically represent urban or agricultural (field) targets, while natural features, such as forest edges, are generally more irregular in shape, except where humans have created a road or clear cuts.

Texture is the arrangement and frequency of tonal variations in particular areas of an image. Rough textures would consist of a mottled tone where the grey levels change abruptly in a small area; smooth textures would have very little tonal variation.

Pattern represents spatial arrangement of visibly discernible objects. Typically, an orderly repetition of similar tones and textures will produce a distinctive and ultimately recognizable pattern.

Shadow may provide an idea of the profile and relative height of a target or targets which may make identification easier. However, shadows can also reduce or eliminate interpretation in their area of influence, since targets within shadows are much less (or not at all) discernible from their surroundings.

Site and associations show relationships between other recognizable objects or features in proximity to the target of interest. For example, a lake is associated with boats, a marina and adjacent recreational land.

In many cases in-house image interpretation requires some field work to implement field verification. Field verification can be considered a form of collateral material. Field verification is typically conducted to assist in the analysis of the data to be analyzed. Essentially, this is familiarizing the interpreter with the area or type of feature or object to be interpreted. This type of verification is done prior to the interpretation. After an interpretation, field verification can be accomplished to verify the accuracy of the interpretation conducted.

5.3.8 Image analysis: image classification

Images, represented in digital form, allow the use of sophisticated computer algorithms to extract many types of information. Automated image classification is one of common techniques used in remote sensing.

Image classification uses the spectral information represented by the digital numbers in one or more spectral bands, and attempts to classify each individual pixel based on this spectral information. The objective is to assign all pixels in the image to particular classes or themes (e.g. water, coniferous forest, deciduous forest, corn, wheat, etc.). The resulting classified image is comprised of a mosaic of pixels, each of which belong to a particular theme, and is a thematic "map".

There are two main approaches in image classification: *unsupervised classification* which is completely done through the use of statistics, and *supervised classification*, where the user “guides” the classification by providing examples for the computer to follow.

Unsupervised classification can be defined as the identification of natural groups, or structures, within multi-spectral data. It can be demonstrated that remotely-sensed images are usually composed of spectral classes that internally are reasonably uniform in respect to brightness in several spectral channels. Unsupervised classification is, therefore, the process of identifying, labelling and mapping these classes. In unsupervised classification, spectral classes are grouped first, based solely on the numerical information in the data and are then matched by the analyst to information classes (if possible).

Usually, the analyst specifies how many groups (or clusters) are to be examined in the data. In addition to specifying the desired number of classes, the analyst may also specify parameters related to the separation distance among the clusters and the variation within each cluster. The end result of this iterative clustering process may result in some clusters that the analyst will want to subsequently combine or clusters that should be broken down further - each of these requiring a further application of the clustering algorithm. Thus, unsupervised classification is not completely without human intervention. However, it does not start with a pre-determined set of classes, as in a supervised classification.

In contrast to unsupervised classification, supervised classification involves some form of supervision by the operator by specifying, to the particular algorithm, numerical descriptors of various land-cover types present in a particular scene. The analyst identifies in the imagery homogeneous representative samples of the different surface cover types (information classes) of interest. These samples are referred to as training areas (sites).

The selection of appropriate training areas is based on the analyst's familiarity with the geographical area and his/her knowledge of the actual surface cover types present in the image. Thus, the analyst is "supervising" the categorization of a set of specific classes. The numerical information in all spectral bands for the pixels comprising these areas is used to "train" the computer to recognize spectrally similar areas for each class. The computer uses a special program or algorithm (of which there are several variations), to determine the numerical "signatures" for each training class. Once the computer has determined the signatures for each class, each pixel in the image is compared to these signatures and labelled as the class it most closely "resembles" digitally. Thus, in a supervised classification we are first identifying the information classes which are then used to determine the spectral classes which represent them.

The last stage of supervised classification is the accuracy assessment, when the results of classifications are checked against certain criteria to statistically assess quality of the achieved results. Like in the case of visual image interpretation, field verification is often needed to make the final decision on correctness of performed image classification.

Summary

- The underlying basis for most remote sensing is measuring the varying energy levels of a single entity, the fundamental unit in the electromagnetic force field known as the photon
- Variations in photon energies are tied to the parameter wavelength, or its inverse, frequency
- EM radiation that varies from high to low energy levels comprises the electromagnetic spectrum
- EM radiation interacts with mediums and can be scattered, reflected, emitted by the substance, transmitted (refracted) or absorbed.
- Amount of the energy that is reflected by targets on the Earth's surface over a variety of different wavelengths, represented in numerical or graphical form, is called a spectral response for that object.
- Spectral signatures can be used to distinguish different Earth features.
- Remote sensing systems differ by platform, type of sensor and wavelengths of operation
- Majority of sensors deploy natural source of energy
- Most historical images were acquired by photographic cameras using analogue films
- Modern remote sensing systems deploy digital sensors
- Multi-spectral remote sensing systems register EM radiation in several spectral zones (channels or bands)
- Different channels might be composed using three principal colors to produce real and false-color image composites
- Images might be interpreted visually or with the aid of computers

- Visual interpretation requires skills and knowledge in applying several basic interpretation elements
- Automated image classification is possible with digital imagery and employs computer algorithms
- There are two basic techniques for automated classification: supervised and unsupervised classification
- Unsupervised classification is purely a statistical approach to classifying clusters of pixels into a certain number of classes
- Supervised classification involves an operator to select training areas, which then are used to calculate statistically consistent areas that correspond to certain classes

Module self-study questions:

- What are the fundamental characteristics of electromagnetic radiation and how they are used in remote sensing?
- How does electromagnetic radiation interact with objects on the Earth's surface?
- How do atmospheric windows differ from spectral signatures?
- Discuss features of satellite sensors and aerial photo cameras in remote sensing. Describe the main differences, advantages and disadvantages.
- What is image resolution? How do these resolutions define capabilities of imaging systems?
- What are spectral ratios and why they are useful? What is NDVI?
- What is the concept of multi-spectral remote sensing? Evaluate differences between panchromatic, color and multi-spectral images.
- List and describe the basic elements in visual (manual) image interpretation.
- Elaborate on the principles of automated image classification.

Required Readings:

- Lillesand, T.M., Kiefer, R.W. (2000), Remote Sensing and Image Interpretation, 4th Edition. John Wiley and Sons, Inc.

ESRI Virtual Campus Course:

- Understanding Geographic Data

Assignment:

- Exploring remote sensing data in ArcGIS
- Supervised and unsupervised classification

References

- Lillesand, T.M., Kiefer, R.W. (2000), Remote Sensing and Image Interpretation, 4th Edition. John Wiley and Sons, Inc.
- Pain, D., and Kiser, J. (2003). Aerial Photography and Image Interpretation. 2nd Edition. John Wiley and Sons, Inc.
- CCRS Tutorial: Image Analysis http://ccrs.nrcan.gc.ca/resource/tutor/fundam/index_e.php
- The "Short" Tutorial: Short, N., NASA, http://www.fas.org/irp/imint/docs/rst/Front/tofc.html
- Lo, C.P. and Albert K.W. Yeung, (2002). Concepts and Techniques of Geographic Information Systems, Pearson Education Canada, Inc., Toronto.

Terms used

- Photon
- Electromagnetic radiation
- Wavelength
- Frequency
- Electromagnetic spectrum
- Color
- Absorption
- Transmission
- Reflection
- Spectral signature
- Passive sensor
- Active sensor
- Platform
- Image formation: frame and scanning
- Image registration: film and digital array
- Panchromatic, color and multi-spectral image
- Channels (bands)
- Color composition / de-composition
- Aerial photographs
- Multi-spectral imaging systems
- RADAR
- LIDAR
- Image resolutions: spatial, spectral and temporal
- Image pre-processing
- Radiometric correction
- Atmospheric correction
- Geo-referencing
- Image enhancement
- Histogram stretching
- Histogram equalization
- Image transformations
- Spectral ratio
- NDVI
- Re-sampling
- Image visual interpretation vaizdų vizualusis dešifravimas, žmogaus atliekamas dešifravimas
- Elements of visual image interpretation
- Image classification – vaizdų klasifikavimas
- Unsupervised classification – nekontroliuojamas klasifikavimas
- Supervised classification – kontroliuojamas klasifikavimas