Online-PDH©

Online Continuing Education for Professional Engineers
Since 2009

Guide to GIS Technology

PDH Credits:

8 PDH

Course No.:

GIS101

Publication Source:

Original Courseware

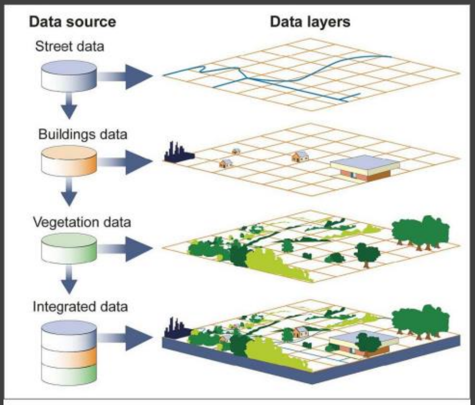
by Donald W. Parnell, PE

Release Date:

2018

DISCLAIMER:

All course materials available on this website are not to be construed as a representation or warranty on the part of Online-PDH, or other persons and/or organizations named herein. All course literature is for reference purposes only, and should not be used as a substitute for competent, professional engineering council. Use or application of any information herein, should be done so at the discretion of a licensed professional engineer in that given field of expertise. Any person(s) making use of this information, herein, does so at their own risk and assumes any and all liabilities arising therefrom.



Source: GAO.

Guide to GIS Technology

Credits: 6 PDH

Course Description

This course provides a comprehensive review of GIS technology.

Topics

- Functions of a GIS and Geoprocessing Operations
- Raster preprocessing, GIS “Features”, Shapefiles, Metadata
- US GIS Spatial Data Standards, and GIS Spatial Data Standardization
- GIS Hardware and Software; ArcGIS and other GIS applications
- Spatial and Attribute Data Models
- What is the Data Structure; What is Spatial Data
- Raster and Vector Data Model and Data Structure
- The TIN Data Model and Data Structure, Digital Terrain Modeling
- The Attribute Data Model and Data Structure
- The Topological Data Model and Data Structure
- Sources of Datasets
- Data Input – COGO, Digitized, Scanned
- Converting and Modifying Existing Data
- Remote Sensing, Data Verification
- The Data Storage and Retrieval System
- Spatial Databases or Geodatabases
- Data Compression vs Compaction; Organizing Data for Analysis
- Spatial Data Layers - Vertical Data Organization
- Spatial Indexing - Horizontal Data Organization
- Editing, Updating, and Conversion of Data
- Types of GIS File Formats, Spatial Data Relationships
- Spatial Data Errors,Spatial Editing (Data Cleanup)
- Manipulation and Transformation of Spatial Data
- Integration and Modeling of Spatial Data
- Data Quality and Standards

Online-PDH©

Online-PDH
1265 San Juan Dr.
Merritt island, FL 32952

Original Course by:
Donald W. Parnell, PE
All Rights Reserved

Chapter 1: GIS Technology and Geoprocessing

What is GIS?

GIS Systems

For the simplest definition of a GIS system; they are basically a marriage between computerized geographical mapping and database management, geographical data processing, and analytical applications.

This course introduces the basic GIS concepts used to visualize real-world features, establish patterns, analyze and obtain feedback information, and output that information to others visually through the use of maps, or textually through tabular data.

Using a GIS software platform, you can create and publish GIS maps, examine the data within the maps, analyze patterns within the maps, and query data from those maps.

GIS is a widely encompassing concept that can refer to a variety of geographical data visualization and analysis technologies, processes, and methods. It is attached to many operations and is related to many policy-making, engineering, planning, management, logistical, and statistical applications.

Functions of a GIS System

The Functions of a GIS system

A GIS system functions as a computerized network used to represent maps as data layers that can be studied in order to perform complex analyses in a visual manner.

It is a system of computer hardware, software, and geographic data that operators interact with, in order to perform the following operations:

Integrate Geographic Data

Aerial photos, USGS topo maps, soil maps, TIGER maps, CAD vector data, COGO digital surveyor points, cadastral data, etc.

Organize geographic spatial data

Create spatial data layers for overlaying (soil types, hydrology, roads, utilities, land features, boundaries, etc.)

Integrate and link with attribute (textual) data

Add labels, descriptors, other data fields, categories, classes, ranks, etc.)

Visualize the data

Combine various data layers, turning on some features, and turning off others in order to create a collage of visual data which represents a particular scenario to be analyzed.

Identify relationships, trends, and patterns

Create boundaries which can be overlaid to create sets and subsets, in order to recognize patterns and draw conclusions.

Data query

Find solutions to complex problems easily found through visual analysis of overlaid spatial and attributtal data sources.

Analyzing Spatial Data

It’s important to acknowledge, that the critical and primary function for a GIS system is the analysis of spatial data. GIS is not a recent innovation. It is actually a tool which has been utilized within a variety of scientific disciplines for many years.

Natural resources and environmental specialists, in particular, have been processing geographic data and honing these techniques since the 1960's.

More Powerful GIS Systems

As the art of geo-processing has continued to evolve, the GIS of today stands apart from the geographical processing procedures used in the past.

Through the use of technological innovations in computer processing and digital datasets, geographic data processing tools are able to explain and display spatial data in a far more effective, comprehensive, and visually-appealing manner.

The Functional Components of a GIS

A GIS system consists of four main functional components:

1.) Data input

This component consists of the capturing, collection, and transforming of spatial and thematic data into digital form.

Data input is usually derived from a combination of hard copy archival documents (maps and planimetrics), aerial photographs, remote sensed images, technical reports (soil, environmental, geotechnical, etc.), surveyor records and documents, etc.

2.) Data storage and retrieval

Within this component is the organization of the spatial and attribute data, into a form which will permit quick retrieval by the GIS user for analysis. It provides the means for efficient and accurate updating of the database.

This component usually involves use of a database management system (DBMS) for maintaining attribute data. Spatial data is usually encoded and maintained within various proprietary file formats.

3.) Data manipulation and analysis

This component allows the user to define and execute spatial and attribute procedures to derive tabular and graphical analytical information.

This component is considered the primary functional component of the GIS, distinguishing it from other database informational systems.

4.) Data output and display

This component allows the GIS user to create graphic presentations, such as maps, and empirical data such as tabular reports, which summarizes and displays the analytics derived through the previously mentioned component.

What is Geoprocessing?

Geoprocessing

Geoprocessing is a series of GIS operations for manipulating spatial data. It allows for the defining, managing, and analysis of assorted forms of data, in order to make informed decisions.

Geoprocessing Operations

The typical geoprocessing operational steps involve:

- Receiving an input dataset
- Performing the operation on that dataset
- Processing the result of the operation as an output dataset

Common geoprocessing operations include:

- *Raster preprocessing* - preparing images for viewing and analysis
- *Feature selection* – formatting and classification of feature elements
- *Feature analysis* – analysis of points, lines, polygons, cells
- *Data conversion* - The process of translating data from one format to another
- *Overlaying geographic features* – converting raw data sources into themes and layers
- *Topology processing* – the conversion of raw data to comply with the topological data structure

Raster Preprocessing

Simple raster editing is a geoprocessing operation that prepares images for viewing and analysis.
It includes:

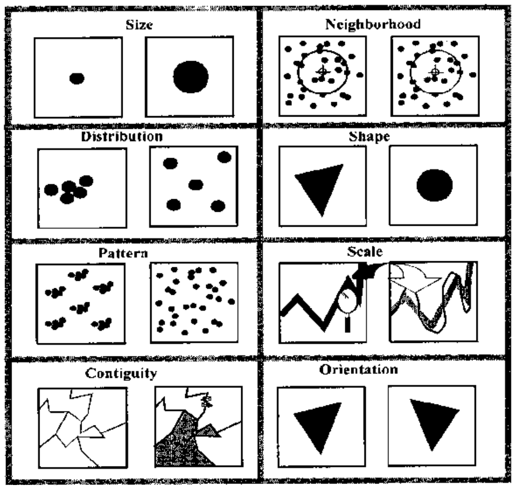
- Clipping
- Positioning
- Resizing
- Enhancing
- Georeferencing
- Mosaicking

What are GIS “Features”?

Features (image) are representations of real-world objects on a map.

Features can be represented in GIS as vector data (points, lines, or polygons) or as cells in a raster data format.

To be displayed in a GIS, features must have geometry and locational information.



Properties of GIS features
Image source: colfa.utsa.edu

Feature Class

This is a collection of geographic features with the same geometric type (such as a point, line, or polygon), the same attributes, and the same spatial reference. Feature classes can stand alone within a geodatabase or be contained within shape files, coverages, or other feature datasets.

Feature classes allow like features to be grouped as a single unit for data storage. Features such as highways, urban streets, and rural secondary roads can be grouped into a line feature class labeled *Roads*.

In a geodatabase, feature classes are also capable of storing annotation and dimensions.

Feature Dataset

This is a collection of feature classes stored together that share similar spatial reference; meaning they align with the same coordinate system, and their features fall within a common geographic area.

Feature classes with different geometry types may be stored in a feature dataset, as well.

Shapefiles

Part of the process of geoprocessing is creating *shapefiles*. These are geospatial vector data storage formats for storing the location, shape, and attributes of geographic features. A shapefile is stored in a set of related files and contains one feature class.

The shapefile format can spatially describe the vector features (points, lines, and polygons), representing items such as power poles, routes, and lakes.

Each item will usually have attribute data linked with it to describe the features within, such as title, name, temperature, etc.

Required files which make up a "Shapefile":

- **shp** - the shape format; feature geometry
- **shx** - a positional index of the feature geometry
- **dbf** - columnar attributes for each shape, in dBase IV format

Optional file types used with a Shapefile:

- **prj** (projection format) - plain text file describing the projection using well-known text format
- **sbn** and **sbx** - spatial index of the features
- **fbn** and **fbx** - spatial index of the features (read-only)
- **ain** and **aih** - attribute index of the active fields within a table
- **ixs** - geocoding index for read-write datasets
- **mxs** - geocoding index for read-write datasets (ODB format)
- **atx** - attribute index for the .dbf file
- **shp.xml** - geospatial metadata in XML format
- **cpg** - used to specify the code page (only for .dbf)
- **qix** - alternative quadtree spatial index

What is Metadata?

In GIS software, metadata is information which describes the characteristics of GIS spatial data or datasets.
This includes data parameters such as:

- an overview of the data
- the associated coordinate system
- its attributtal information

- the origin and accuracy of the data
- the title
- the type
- the source
- author
- last modified date
- thumbnail
- tags

Metadata can also include additional information such as:

- summary and description
- how accurate and recent the item is
- restrictions associated with using and sharing the item
- credits, etc.

Metadata can help others to query, parse, and validate the usefulness of the GIS data. Metadata is saved with the item it describes, and can be copied, moved, and deleted with the item, and is typically hidden from view, being accessible on a code level.

Metadata Standards - (FGDC) Federal Geographic Data Committee

This is an organization established by the US Federal Office of Management and Budget, which is responsible for coordinating the development, use, sharing, and dissemination of surveying, mapping, and related spatial data.

The FGDC defines spatial data metadata standards for the US in its "Content Standard for Digital Geospatial Metadata" publication, and manages the development of the National Spatial Data Infrastructure (NSDI).

GIS Hardware

Typical hardware used in by GIS personnel includes:

Desktop Computer Workstation

The standard workstation for a GIS operator is the common desktop or laptop computer. Due to the large amounts of data which need to be processed by the GIS software, it's best to use stations with the fastest CPUs affordable.

Use of dual or three monitor video graphics cards with large processors and memory are recommended as well.

Data Server or Cloud Server based GIS (Storage)

Depending upon the particular application, GIS systems designed for public access are normally found on data servers or cloud based servers, which are accessible over the internet.

Digitizers (Vector Data input)

These are electronic grid input devices which allow vector data points to be manually input into CAD or GIS drawings and maps through the use of a stylus pen, or "puck" style cross-haired, handheld stylus.

They come in all sizes from small tablets of 9"x 9" to large tables (image) of 36"x48" or larger. These are used for Paper to CAD conversions.

Digitizer Table

Image Source: Calcomp.com

Plotters (Printing - Output)

These are large format printers which can create large maps, plans, etc. in either raster or vector based data files.

Modern plotters tend to be mostly inkjet style printers. Some plotters are dual purposed printers and large format scanners.

Image Scanners (Raster Data Input)

This is a device that optically scans images, printed text, handwriting, and converts it to a digital rasterized image.

Small format scanning is the desktop flatbed scanner where the document is placed on a glass window for scanning.

All in One copier-printer-scanners work well for automatically inputting multiple text documents. Hand-held scanners, where the device is moved by hand work well for quick scanning of notes.

These scanners can be used for converting CAD drawings as well. Many CAD programs have vectorizing features which can convert a rasterized "line" drawing into vector elements.

These vectorized drawings tend to need a lot of cleanup afterward to improve the positional accuracy.

Large Format Scanner

Image Source: spatialvision.com

Large format scanners (Raster Data Input)

(image) These work on the same principle as large format plotters are the most important form of paper-to-raster or CAD conversions. These allow for scanning of decades-old plans and maps.

Mobile Devices

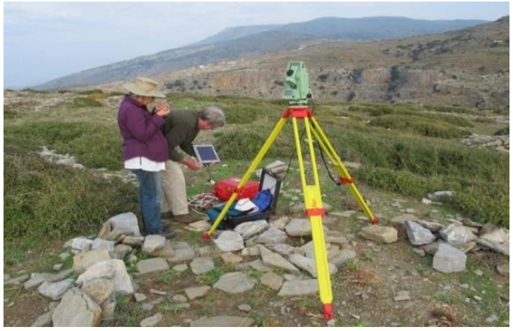
Most GIS systems are accessible by mobile device, these days. A GIS is the basis for the global navigational systems found in phones, devices, and automobiles.

When routes are automatically computed for your next trip, it will be accessing the contiguous, vector-line topological data structure to create the shortest route, and analyze the road conditions.

Field Hardware:

Total Stations (Coordinate Data Input)

These are devices (image) used by surveyors for electronic distance and angle measuring.



Total Station

Image source: powerhousemeusem.com

Total Station Prism

This is the other component used in total station surveying. The laser from the TS unit interacts with the infrared reflective prism (image) to compute distance and angle.

TS Prism

Robotic Total Stations

This is the same as a regular total station, but also has tracking, which allows for a single operator.

Data Collectors

This is a handheld device (image) which gathers the COGO data (or other data formats) from the TS units or LIDAR units for transference to the computer workstation.

Handheld Data Collector

Image source: aliexpress.com

3D LIDAR Scanners

This is a scanner (image) which can capture thousands of points to create a "point-cloud" for use with 3D topo applications.

LIDAR scanner unit
Image source: bgs.ac.uk

GPS Transceivers (Coordinate Data Input)

This is equipment used for gathering satellite signals for location positional data. It outputs COGO coordinate data, similar to a Total Station.

GIS Software

There is more than one software program which goes into the total GIS system. (Many are proprietary programs, while others are open source).

- *Proprietary* - This is purchased software or closed-source software which typically contains restrictions on use, analysis, modification, or distribution.
- *Open Source* - Open-source software has its source code made available with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose.

The following is a list of some of the more popular software platforms utilized in GIS development; however this course provides no endorsement of any given product!

Proprietary GIS framework

The industry or de facto standard for GIS is the ESRI suite of software programs released under the trademarked name, *ArcGIS*.

- *ArcGIS* - is a multi-scale architecture, with a desktop product released at three licensing levels:
- *ArcView (Basic Level)* - provides a basic set of GIS capabilities suitable for many GIS applications.
- *ArcEditor (Standard Level)* - at added cost, allows more extensive data editing and manipulation, including server geodatabase editing
- *ArcInfo (Advanced Level)* - The ArcInfo license allows users the most flexibility and control in all aspects of data building, modeling, analysis, and map display providing full, advanced analysis and data management capabilities, including geostatistical and topological analysis tools.

Open Source GIS - GRASS

GRASS - This is a free open source GIS framework developed by the Army Corp of Engineers. GRASS (Geographic Resource Analysis Support System) is an alternative to ESRI's premium-priced ArcGIS platform.

Geographical Resources Analysis Support System (GRASS) is one of the largest free software GIS projects released under the GNU General Public License (GPL). It combines powerful raster, vector, and geospatial processing engines into a single software package with tools for spatial analysis, modeling, image processing and sophisticated visualization.

GRASS GIS is highly utilized in academic institutions, with a large number of open source and proprietary add-ons or plugins for analyzing GIS datasets.

Other Open Source GIS Tools:

- *Capaware* - 3D GIS Framework with geographic graphical analysis and visualization features.
- *FalconView* - A mapping system created by the Georgia Tech Research Institute for Windows.

- *Kalypso*- Used on numerical simulations in water management.
- *TerraView* - Handles vector and raster data stored in a relational or geo-relational database.
- *Whitebox GAT* - Cross-platform, free and open-source GIS software.

GIS/Remote Sensing

- *Idrisi* - This is a GIS/remote sensing software package developed in the 80's. It is widely accepted for use within educational systems. It has a variety of tools such as image classification, restoration, and enhancements, as well as temporal analysis and object-based image analysis.
- *Earth Resources Data Analysis System (ERDAS) Imagine* - This is a GIS and remote sensing processing software owned by Hexagon Geospatial. ERDAS Imagine is a leading remote sensing software package with a range of classification, NDVI and image processing tools for satellite, hyperspectral, radar, LiDAR and other remote sensing data.
- *ENVI* - This is a software application used to process and analyze geospatial imagery, which is commonly used by remote sensing professionals and image analysts.

Graphical and Modeling

3D Civil

This is the basis of Land Development TIN modeling. Site and corridor (route) modeling datasets are very useful for spatial analysis. These programs are good for editing of TIN datasets.

They are capable of inputting surveyor COGO data, and converting it into terrain models (TIN). They can also process and edit point cloud data (LIDAR).

Vector Editing Programs

Common vector editing programs such as Adobe Illustrator, Corel Draw, or one of several open source programs, such as Inkscape are used for small scale editing.

Once again though, these are not the best choice for editing vector files are CAD programs such as Microstation and Autocad.

Raster Editing Programs

To edit raster data requires the use of raster editing software such as a proprietary program by Adobe, called Photoshop, or the open source program GIMP.

However, these programs do not work well for the types of large scale raster files encountered in GIS.

For this type of raster data, programs such as IRAS/B and Descarte which work inside of the Microstation CAD platform, are recommended. Within the Autocad platform, raster design is a good plug-in.

GDAL Geospatial Data Abstraction Library (GDAL)

This is a C++ library for reading and writing raster geospatial data formats, implementing common GIS operations (unions, intersections, joins, clipping, etc.) with command line utilities.

It supports old hardware and operating systems and requires minimal amounts of memory.

Engineering and Scientific Modeling and Analysis Programs

There are various programs which can produce a wide variety of file formats for importation into a GIS project.

These programs come in 2D and 3D environments, and can provide a model for nearly any scenario such as hydrological analysis, geotechnical analysis, power grid analysis, water and wastewater network analysis, and so many more.

Building Information Modeling (BIM) Programs

Building Information Modeling platforms such as Autodesk's Revit are used for outputting 3D models for architectural, MEP, structural, and facility purposes.

Although BIM and GIS are apples and oranges, when it comes to their basic core functions, BIMs are still capable of outputting datasets which can be of niche uses within the GIS geodatabase.

They are very useful for building AM/FM (Automated Manufacturing and Facilities Management) system models, which are similar in function to a GIS.

Standards for GIS Spatial Data

(FGDC) Federal Geographic Data Committee

This is an organization established by the US Federal Office of Management and Budget, which is responsible for coordinating the development, use, sharing, and dissemination of surveying, mapping, and related spatial data.

The FGDC defines spatial data metadata standards for the US in its “Content Standard for Digital Geospatial Metadata” publication, and manages the development of the National Spatial Data Infrastructure (NSDI).

(SDI) Spatial Data Infrastructures

A spatial data infrastructure (SDI) is a data infrastructure which is set up as a framework of geographic data, metadata, users, geoprocessing tools, and analysis tools. It is a combination of technologies, policies, standards, human resources, and related activities needed in order to facilitate the acquisition, processing, distribution, use, maintenance, and preservation of spatial data.

It’s also a means in which to coordinate technological standards agreements, institutional arrangements, and policies that will enable the innovation of geospatial resources and uses.

Most countries have their own specific Spatial Data Infrastructures. The International SDI is the GSDI Association which is an organization of academics and researchers, government agencies, firms, NGOs and individuals from around the world.

(NSDI) National Spatial Data Infrastructure

This is the SDI within the US, which was established by the FGDC organization. The goal of the NSDI is to reduce the duplicated efforts made among agencies, improve data quality and reduce the costs associated with geodata.

They are also tasked with the goal of making geographic data more accessible to the public, and to establish key partnerships with states, counties, cities, tribal nations, academia and the private sector to increase data availability.

GIS applications use a number of recurring themes of data which could be shared collectively amongst these users.

The framework is a collaborative effort in which these commonly needed data themes are developed, maintained, and integrated by public and private organizations within a geographic area.

The framework is one of the key building blocks and forms the data backbone of the NSDI.

This framework has three aspects:

- Data
- Procedures
- Technology

List of federal agency leads for framework theme development:

- *Geodetic Control* - National Geodetic Survey
- *Cadastral* - Bureau of Land Management
- *Ortho-imagery* - U. S. Geological Survey
- *Elevation* - U. S. Geological Survey
- *Hydrography* - U. S. Geological Survey
- *Administrative units* - U. S. Census Bureau
- *Transportation* - State DOTs, Federal Highway Adminstration (FHWA)

The Open Source Geospatial Foundation (OSGeo)

Established as a GIS Software Developmental Support Organization, OSGeo was formed in February 2006, OSGeo is a non-profit, non-governmental organization established to support, and promote the collaboration and development of open sourced geospatial software, technologies and datasets.

Chapter 2: Data Models used in GIS

Types of Spatial and Attribute Data Models

What are Spatial and Attribute Data Models?

A *data model* organizes data elements and standardizes how the data elements relate to one another. (The *data structure* is how a particular type of data model's coding is constructed).

A spatial data model is a means in which to store the spatial location of geographic objects or surfaces within a database.

The vector data model represents geography as groups of points, lines, and polygons, raster data models represent geography as cell matrices that store numeric values, and TIN data models are represented by irregularly triangulated vector surface facets.

There are spatial data models (graphical features and elements) and attribute data models (database textual data).

The coordinate location of a telephone pole would be its spatial data, while the characteristics of that pole: height, ratings, material, etc. would be its attribute data.

Selection of a Data Model

The selection of a particular data model can provide advantages when performing spatial analysis.

For example, the vector data model handles continuous data poorly (e.g. elevation), while the raster data model is better suited for this type of analysis.

While the raster data structure does not handle linear data analysis particularly well (e.g. shortest path for a computed route), vector systems can.

It's important to understand the advantages and disadvantages of each data model. The selection of a particular data model is dependent on the source and type of data, as well as the intended use.

Certain analytical procedures require raster data while others work best with vector data.

Spatial Data Model components:

- **Vector data models** – consists of point, line and polygon data
- **Raster data models** - consists of pixel and cell data
- **TIN data models** – consists of sets of contiguous, non-overlapping triangulated surface facets

Attribute Data Models:

Tabular or attribute data models

Attributes refer to the properties of spatial entities. Attribute data models consist of character, numeric, and temporal data.

They are commonly referred to as non-spatial data as they do not represent locational information.

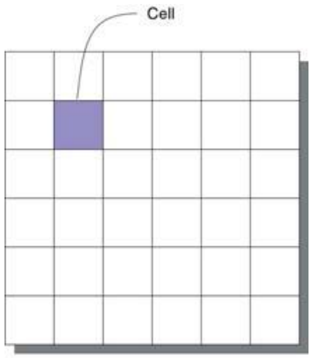
This type of data describes the quantitative or qualitative characteristics of the spatial features.

The Raster Data Model

The Raster Data Structure

A raster image is basically an array of cells (image), with each cell having a value which represents a specific portion of an object or feature. Raster images are a matrix of individual pixels, with associated pixel attributes, such as color, elevation, or an ID number.

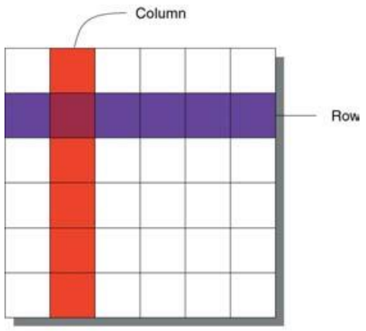
One grid cell is one unit or holds one attribute, with every cell having a value, even if it is empty. A cell can hold a number or an index value standing for an attribute. A cell has a resolution, given as the cell size in ground units.



Raster Cell
Image source: Gitta.info

All cells in a raster must be the same size, determining the resolution. The cells can be any size, but they should be small enough to accomplish the most detailed analysis. A cell can represent a square foot or a square inch.

Cells are arranged in a grid of rows (x-axis), and columns (y-axis), producing a Cartesian coordinate system arrangement (image).



Raster Cells, Columns and Rows
Image source: Gitta.info

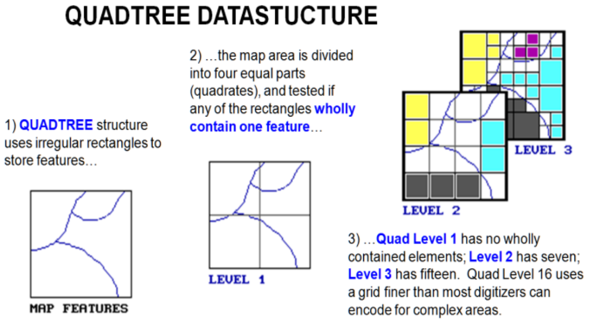
Each cell has a unique row and column address.

For raster data converted from vector data sources:

- *Points* – are converted to a single cell, or by a group of clustered cells
- *Lines* – are converted as a sequence of neighboring and contiguous cells
- *Polygons* – are converted to a collection of contiguous, staggered cells.

Raster Data (Image Formats and Image Data Encoding):

- *Tessellations* – A tessellation is a pattern of plane figures that fills the plane with no overlaps and no gaps. It is usually in a square grid pattern but can be other shapes as well. Tessellated data structures are constructed using this pattern.
- *Run length encoding* **(commonly used for image compression)** – is a very simple form of lossless data compression in which runs of data (or sequences with the same data occurring in many consecutive data elements) are stored as a single data value and count, rather than as the original run. This is useful on data that contains these runs such as simple graphic images.
- *Quad tree representation* **(commonly used for image compression)** – A quadtree (image) is a tree data structure in which each internal node has exactly four children or subnodes. Quadtrees are most often used to partition a 2-D space by subdividing it into quadrants or regions of fours (Quad).


QUADTREE DATASTUCTURE
1) QUADTREE structure uses irregular rectangles to store features…
2) …the map area is divided into four equal parts (quadrates), and tested if any of the rectangles wholly contain one feature…
3) …Quad Level 1 has no wholly contained elements; Level 2 has seven; Level 3 has fifteen. Quad Level 16 uses a grid finer than most digitizers can encode for complex areas.
MAP FEATURES
LEVEL 1
LEVEL 2
LEVEL 3

Quadtree Data Structure
Image Source: innovativegis.com

- *Band sequential (BSQ)* – is one of three primary means for encoding image data for *multiband raster* images in GIS, (such as images obtained from satellites). BSQ is not an image format, but is a method for encoding the actual pixel values of an image.
- *BIP (Band Interleaved by Pixel)* – Images encoded in BIP format have the first pixel for all bands in sequential order, followed

by the second pixel for all bands, followed by the third pixel for all bands, etc., interleaved up to the number of pixels.

- *BIL (Band Interleaved by Line)* – This encoding format stores the actual pixel values of an image in a file band by band for each line, or row, of the image. For example, given a three-band image, all three bands of data are written for row one, all three bands of data are written for row two, and so on. The BIL encoding is a compromise format, allowing fairly easy access to both spatial and spectral information. The BIL data organization can handle any number of bands, and thus accommodates black and white, grayscale, pseudo-color, true color, and multi-spectral image data.

Raster Input

Raster images are converted to digital form by the use of optical scanners, digital CCD cameras, and other raster imaging devices. Spatial resolution of a raster image is determined by the resolution of the input device and the quality of the original scanned data source.

Increasing the Raster Resolution

(2x the resolution, results in 4x the pixels) – One thing to keep in mind when enlarging the resolution on a raster image, the raster image has to have pixels for all of its spatial locations. Therefore when increasing the spatial resolution of an image by 2 times, the total file size of a 2-D raster image will increase by 4 times. This occurs because the number of pixels is doubled in both X and Y dimensions.

Grid-Cell Data Structure

Raster data models use a grid-cell data structure where the geographic area is divided into cells identified by row and column.

Tessellated Data Structures

Raster Structure

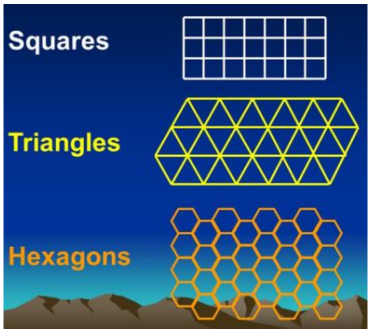
The most popular grid-cell structure is the regular-spaced matrix or raster structure. This data structure involves a division of spatial data into regularly spaced cells, with each cell being of the same shape and size.

Square grids are the most commonly used tessellated structure (Tessellated – tiled plane using one or more geometric shapes).

A raster data structure is a matrix where any coordinate can be calculated if the origin point and the size of the grid cell is known.

As grid-cells can be processed as 2-D arrays in computer encoding, many analytical operations are easy to program, making tessellated data structures a usual choice for many GIS software programs.

While the term raster typically refers to a regularly spaced grid, other tessellated data structures do exist in grid based GIS systems (image).



Types of Tessellated Data Structures
Image Source: Robyn Thornton

Resolution

This is a measure of the accuracy or detail of a raster image, which is expressed as dots per inch, or pixels per inch.

Each pixel is a sampling of an original image, with more samples (denser pixilation) providing more accurate representations of the original.

Spatial Resolution is the accuracy associated with the capturing of ground information, which is reproduced in a digital format, for example, 10-foot pixels or 100-foot pixels.

Minimum Mapping Unit

The minimum mapping unit is the smallest resolution area when interpreting remotely-sensed satellite or aerial imagery.

Some Raster Data Sources

The following are some sources and types of raster data used in a GIS system:

- *Satellite Data* – This is remotely sensed data obtained from extraterrestrial image capturing devices.
- *LANDSAT* – The satellites, Landsat 1 through 8 have provided millions of digital images, since the 70's.
- *SPOT* – Spot 1 through 7 are French satellites which are missioned with creating scientific, ecosystem and humanities based imagery of high resolution.
- *Scanned aerial photography* – This is typically archived analog (film) photography which has been scanned into digital form, and rectified, as aerial photography tends to have a "fish-eyed" type of distortion which is inherent in this type of photo.
- *Scanned Ortho-photography* – This is scanned analog aerial photos which have already been through the rectification process.
- *Scanned Maps and Other Documents* – hardcopy paper, Mylar, or vellum archives which have been scanned into digital form.
- *Digital Raster Graphics (DRG)* – This is a digital version map of the USGS topographic maps. They include imagery (NAIP), roads, place names, hydrography, elevation contours, and boundaries.
- *TIGER (Census Maps)* – The TIGER Map Service of the US Census Bureau was retired in May 2010. In 2012 the Census Bureau released TIGERweb, which is a web mapping and streaming services application.
- *Digital Surface Model (DSM)* - This is a digital surface model is an elevation that includes the top of buildings, tree canopy, powerlines and other features above the bare earth. For example, the first return of LiDAR consists of a DSM.
- *Digital Elevation Model (DEM)* – This is a bare earth elevation model representing the surface of the Earth. DEMs come with filtered out non-ground points such as bridges and trees.

Common Applications for Raster Data Layers:

- Utility Corridor Siting
- Environmental Mapping
- Natural Communities Mapping
- Forest resource planning
- Spatial data variability decisions
- Forest inventory
- Wildlife habitat analysis
- Utility Corridor Siting
- Environmental Mapping
- Natural Communities Mapping
- Forest resource planning
- Spatial data variability decisions
- Forest inventory
- Wildlife habitat analysis

Limitations for Raster:

- Aesthetics
- Data storage requirements
- Overlay operations performed on every cell
- Sparse data sets require as much processing as dense ones
- Grids are poor at representing points, lines and areas, but good at surfaces.
- Grids are good only at very localized topology, and weak otherwise
- Grids must often include redundant or missing data
- Grids suffer from the mixed pixel problem

The Image Data Model (Differs from Raster Data)

In a GIS framework, image data differs significantly from raster data.

Image data is typically used for representing:

- Graphic or pictorial data
- As background display data (if the image has been rectified and georeferenced)
- As a graphic attribute

Remote sensing software uses this type of image data for image classification and processing.

Image data is usually stored as standard GIS proprietary file formats, or can be stored as other graphic image formats, such as JPEG, TIFF, GIF, PCX, etc.

Most often, image data is used to store remotely sensed imagery, e.g. satellite scenes or orthophotos, or ancillary graphics such as photographs, scanned plan documents, etc.

To use image data for analytical purposes, this data must be converted into a raster format (or vector).

The Vector Data Model

The Vector Data Structure:

The vector data structure consists of located points (nodes), lines (a connected series of points), and areas (a closed, connected series of points, also called polygons).

Attribute information can be appended to a point, line, or area and stored in a related database. A line standing for a road includes attributes such as name, width, surface, etc. Design characteristics can be appended to points, lines, and areas.

- *Points* - These are stored as coordinate pairs (X-Y)
- *Line (Arc)* – These are stored as sets of coordinate pairs (X1-Y1, X2-Y2); Note: Lines are called arcs in the topological data structure.
- *Polygons* – These are comprised of a series of coordinate pairs, which make up a contiguous series of arcs, which ends with the same coordinate pair as the starting pair. (This closes the loop of vector arcs creating a polygon). When polygons have shared geometry, boundaries that are in common with a neighboring polygon, neighboring polygons should precisely coincide.
- *TINs* – In a triangulated irregular network model, surfaces are represented as a network of linked triangles linked between irregularly spaced points with x, y, and z values. TINs are an efficient way to store and analyze surfaces.

Advantages of the vector data structure:

- *Smaller file size* – than raster imaging, as a raster image needs space for all pixels while only point coordinates are stored in vector representation. This is particularly true in the case when the graphics or images have large homogenous regions and the boundaries and shapes are the primary interest.
- *Easier to compute* – has fewer data items and is more flexible in scale adjustment. This makes the vector data structure the most appropriate choice for mapping, GIS, and CAD applications.
- *Topological graphical objects or items are easier to represent* - using vector form, as a commonly shared edge can be easily defined according to its left and right side polygons. With raster data, this is difficult to do with the pixelated structure.

The CAD Vector Data Structure

Another vector data structure that is commonly utilized within GIS software is the computer-aided drafting (CAD) data structure.

This structure consists of listing “elements”, rather than features, which are defined by strings of vertices, to define geographic features, such as the points, lines, or areas elements.

Redundancy is an issue with this data model, as the boundary segment between two polygons can be stored twice, one for each feature.

The CAD vector data structure was developed for drafting purposed rather than the specific purpose of processing GIS features.

Since features, such as polygons, are self-contained in the CAD structure, queries about the adjacency of these features can be problematic.

The CAD vector model lacks the spatial relationship between features that is a vital part of the topologic data model.

Advantages:

- Graphic output is usually more visually pleasing
- Spatial data can be represented at its original resolution
- The accurate geographical location of data is preserved
- As most data, such as hard copy maps, are provided in vector form, data conversion is not needed
- Allows topology to be more efficiently encoded, resulting in more efficient operations that use topological information

Disadvantages:

- Locations of each vertex need to be explicitly stored
- To be analyzed properly, vector data must be converted into a topological structure
- Processing intensive, requiring excessive data cleanup.
- Topology is static, thus any editing of the vector data requires the topology to be re-built
- Analysis functions are complex
- Limits the functionality for larger data sets
- Continuous data, such as elevation data, is not properly structured in vector form, requiring data interpolation for these data layers
- Spatial analysis and filtering within polygons is not possible

The (TIN) Data Model

A TIN file (Triangulated Irregular Network)

This is another form of digital data structure used in GIS, which is used to represent a surface such as physical land surfaces or the sea bed.

A TIN is vector-based and comprised of a triangular network of irregularly distributed nodes with vertices, known as mass points, with associated 3-D coordinates, connected by edges to form a triangular tessellation. TINs are also called irregular triangular mesh or irregular triangular surface model.

3-D visualizations

3D visualizations can be created by rendering of these triangular facets. In areas with little variation in elevation, the points can be widely distributed, while in areas of more extreme change in elevation, the point density is increased.

Surfaces in TINs

A surface is represented by irregular spaces, a sample point and break line features. The tin dataset includes a topological relationship between points and their neighboring triangles.

Points in TINS

Each sample point has an x, y coordinate and a surface or z value. These points are connected by edges to form a set of non-overlapping triangles used to represent the surface.

TIN’s and DEM’s

TINs are often processed from elevation data of a rasterized digital elevation model (DEM). When using a TIN rather than a raster DEM in mapping and analysis, the points of a TIN are distributed variably based on an algorithm that determines which point priority in order to create an accurate terrain.

Data input

The inputting of data is flexible and fewer points are needed to be stored than in the raster DEM, with its regularly distributed points. A TIN may not work as well as a raster DEM for particular GIS applications, such as analyzing surface slopes and aspects.

Vertices

The vertices (points) of each triangle have unique X, Y, and Z (height) values, but each group of three vertices creates a triangular plane whose surface has a unique angle and direction.

The advantageous thing about the TIN model is that you can use it to predict or interpolate missing values, create cross sections through surfaces and volumes, draw contour lines, and create 3-D visualizations, just like its raster equivalent data model.

Digital Terrain Modeling (DTM)

These are the 3D CAD models usually generated automatically from COGO data imported from total station and GPS data collectors. This is a bare earth representation of the Earth's surface that has augmented natural features such as ridges and breaklines.

DEM or Digital Elevation Model

The representation of continuous elevation values over a topographic surface by a regular array of z-values, referenced to a common datum. These file types are typically used to represent terrain relief.

The Attribute Data Model

A separate data model is used, for storing and maintaining attribute data for GIS. These data models may exist internally within the GIS software, or may be reflected in external locations.

What are Attributes?

These are nonspatial data about a geographic feature in the GIS. They are usually stored in a database table and linked to the feature by a unique identifier.

For example, attributes of a river might include its name, high and low water levels, and stream gauging data, or a road segment (route section from node to node) would have attributtal data which might contain the road name, who maintains it, surface type, and maintenance schedules.

Attribute Domain

In a geodatabase, this is a mechanism for enforcing data integrity. Attribute domains define which values are allowed in a field, in a feature class, or nonspatial attribute table. If the features or nonspatial objects have been grouped into subtypes, different attribute domains can be assigned to each of the subtypes.

Attribute table

This is a database or tabular file containing the information about a set of geographic features, usually arranged so that each row represents a feature and each column represents one feature attribute.

In raster datasets, each row of an attribute table corresponds to a certain region of cells having the same value.

In a GIS, attribute tables are often linked or related to spatial data layers, and the attribute values they contain can be used to locate, query, and symbolize the GIS features or raster cells.

Types of Attribute Data

Attribute data can be store in a table or database, as one of five different field types:

- Character
- Integer
- Floating
- Date
- BLOB

Character Data

The character property or string is used for text-based values (such as the name of a street or descriptive values such as a street's surface type).

- Character attribute data is stored as a series of alphanumeric symbols.
- In addition to descriptors, character fields may contain other attribute values such as *categories* and *ranks*. *For example, a character field may contain the categories for a street: Drive, Court, or Highway*.
- Character fields can also contain the *rank*, which is a rating system. *For example, a ranking of the surface condition of a street with "F" being the street with the lowest surface quality*.
- Character data can be sorted in ascending (A to Z) and descending (Z to A) order. *Since numbers are considered text in this field, those numbers will be sorted alphabetically which means that a number sequence of 1, 2, 9, 11, 15, 23 would be*

sorted in ascending order as 1, 11, 15, 2, 23, 9.

- Because character data is not numeric, calculations (sum, median, product, etc.) can't be performed on this type of field, even if the value stored in the field are numbers (to do that, the field type would need to be changed into a numeric type of field).
- Character fields can be summarized to produce counts (e.g. the number of features that have been categorized as "Drive").

Numeric Data

Integer (whole numbers) and floating (values that have potential decimal places) are numerical values.

- Within the integer type, there is a further division to short and long integer values.
- Short integers store numeric values without fractional values for a shorter range than long integers.
- Floating point attribute values store numeric values with fractional values. Therefore, floating point values are for numeric values with decimal points (numbers to the right of the decimal point as opposed to whole values).
- Numeric values are sorted sequentially, either in ascending (1 to 10) or descending (10 to 1) order.
- Numerical field values can have calculation operations performed such on them, such as sum or average values.
- Numerical field values can be a count (ex. – the total number of employees at a business) or can be a ratio (ex. – the percentage of employees that are male).

Date/Time Data (Temporal Data)

These date fields contains date and time values.

BLOB Data

BLOB (or Binary Large Object) is an attribute type used for storing information such as:

- Images
- Multimedia
- Bits of code in a field

This field stores object linking and embedding (OLE) which are objects created in other applications such as images and multimedia and linked from the BLOB field.

Attribute Data Errors

Finding errors within attribute data is not as straightforward as identifying the spatial errors, especially if the errors are related to data quality or reliability.

This type of error isn't usually found until later in the GIS processing. It is unlikely that an error in attribute data will be caught, if the values are correct in syntax, but incorrect in value.

Data linkage errors such as missing or duplicated records will become evident when linking spatial and attribute data. Most GIS software contains functions that check for and identify these types of linkage issues during operational functions.

Types of Attribute Database Management Systems:

Tabular ADMS

The tabular model was used in early GIS software packages to store attribute data. The simple tabular model stored attribute data as sequential data files with fixed formats (or comma delimited for ASCII data), for the location of attribute values within the predefined record structure.

This type of data model is outdated for GIS use, lacking a method of checking the integrity of data, as well as being inefficient with respect to data storage, (limited indexing capability for attributes or records, etc.).

Hierarchical ADMS

This model organizes data in a tree structure. Data is structured downward in a hierarchy of tables. Any level in the hierarchy can have unlimited children, but any child can have only one parent. They are oriented for data sets that are very stable, where primary relationships among the data change infrequently or never at all.

Also, the limitation on the number of parents that an element may have is not always conducive to the GIS application.

Network ADMS

This model organizes data in a network or plex structure. Any column in a plex structure can be linked to any other. Like a tree structure, a plex structure can be described in terms of parents and children.

This model allows for children to have more than one parent. While the more powerful structure of this DBMS in representing data relationships allows a more realistic model for GIS, network databases tend to become overly complicated quickly.

Relational Database Management System (RDBMS)

RDBMS

The RDBMS organizes data in tables, with each table identified by a unique table name, and organized by rows and columns.

- *Columns* **-** Each column within the table has a unique name, and store the values for a specific attribute, such as power pole group and height.
- *Rows* **-** Rows represent one record in the table, with each row usually linked to a separate spatial feature, such as a bridge, route segment, or power pole. Each row would be comprised of several columns, each column containing a specific value for that geographic feature.
- *Tables* **-** Data is often stored in several tables. Tables can be joined or referenced to each other by common columns (relational fields). Usually the common column is an identification number for a selected geographic feature. This identification number acts as the primary key for the table. The ability to join tables through use of a common column is the essence of the relational model. Such relational joins are usually ad hoc in nature and form the basis of for querying in a relational GIS product.

The relational DBMS is widely used as a commercial data management tool for managing the attributes of geographic data.

Most GIS software will provide an internal relational data model, in addition to support for commercial off-the-shelf (COTS) relational DBMS, or external DBMS.

This supports users with small data sets, where an internal data model will suffice, while also supporting users with larger data sets who utilize a DBMS for multiple data storage requirements.

With an external DBMS, the GIS software can connect to the database, and the user can make use of the inherent capabilities of the DBMS.

External DBMS tend to have much more extensive querying and data integrity capabilities than the GIS internal relational model.

The use of the external DBMS is a trend that has resulted in the expansion and growth of GIS, into a more traditional form of data processing environment.

Advantages of the relational DBMS are:

- **Simplicity** – in organizing and data modeling
- **Flexible** – as data can be manipulated in an ad hoc manner by joining tables
- **Efficient with storage** – when properly designed, data tables can minimize redundant data
- **Non-procedural** – queries on this database do not need to account for internal organization of the data

SQL for querying RDBMSs

The term, which is typically pronounced as “sequel”, stands for Structured Query Language. Developed by IBM in the 70’s, this is syntax for retrieving and manipulating data from a relational database. SQL has become an industry standard query language in most relational DBMS.

Object Oriented

This is a data management structure that stores data as *objects*, or instances of a class, instead of as rows and tables such as in a relational database.

An object being a collection of data elements and operations that combined, are considered a single entity. This approach is desirable in that querying is less complicated, as features can be bundled with attributes.

The Topological Data Model

Topology

Topology is a mathematical approach which allows us to structure data based on the principles of feature adjacency and feature connectivity. It's the mathematical method used to define spatial relationships.

Without a topologic data structure in a vector based GIS, most data manipulation and analysis functions would not be practical or possible.

Topology refers to the relationship between the GIS features, which refers to these characteristics:

- Connectivity of line features
- Directionality of line features
- Adjacencies of polygons
- Containment of features within polygons

Other characteristics:

- Some implementation of vector data structures is topological, such as coverages, while some are not, such as shapefiles.
- Lines cannot overlap without a node in a topological data structure.
- Lines can overlap without nodes in a non-topological data structure, such as a spaghetti structure.

Arc/Node Topological Data Model

The arc/node data model (consisting of an arc and a node), is the most common topological data structure.

(Note: Lines are considered as "arcs" within the topological data structure).

The Arc

This is a series of points, which are joined by straight line segments, starting and ending at a node.

The Node

This is an intersection point where two or more arcs are joined. Nodes also occur at the end of a dangling arc, (an arc that doesn't connect to another arc such as a cul-de-sac). Isolated nodes which are not connected to arcs represent point features.

A Polygon

This feature is made up of a closed chain of arcs.

Topological Tables

GIS software records the topological definitions in three tables, which are similar to relational tables. They represent the different types of features such as points, lines, areas, with a fourth table containing the coordinates.

The Arc, Node, and Polygon Tables

These store information about the node and the arcs that are connected to it. This table contains topological information about the arcs including the starting and ending node, and the polygon to the left and right that the arc is an element of. This table defines the arcs that make up each polygon.

Building the Topological Structure

Since most input data does not exist in a topological data structure, topology must be built with the GIS software. Depending on the data set this can be a CPU-intensive process.

This "topology-process" involves the creation of the topological tables and defining of the arc, node, and polygon entities.

To correctly define the topology, there will be specific rules with respect to graphic elements, such as no duplication of lines, no gaps in arcs that define the polygon features, and so forth.

The topological model is utilized because it effectively models the relationship of spatial entities, and it is well suited for operations to deal with contiguity, adjacency, proximity, and connectivity analytical processes.

Contiguity involves the evaluation of *feature adjacency*, such as when features contact one another, and *proximity*, as when features are close to one another.

Advantages of the Topological Data Model

The primary advantage of the topological model is that spatial analysis can be done without using the coordinate data. Many operations can be performed mostly, if not completely, by use of the topological definition alone.

This is a huge advantage over the CAD or spaghetti vector data structures that require the derivation of spatial relationships from the coordinate data before analysis can be performed.

Disadvantages of the Topological Data Model

The major disadvantage of the topological data model is its static nature. It can be a time consuming process to properly define the topology depending on the size and complexity of the data set.

For example, 100 soil type polygons will take much longer to construct the topology, than 1,000 parcel lot boundaries. This is due to the feature complexity.

Parcels of land are typically constructed of less than 10 arcs (lines), while soil type boundaries could consists of thousands of arcs. This should be taken into consideration when evaluating the topological building process.

The static nature of the topological model also means that every time editing is required, such as changes in the geomorphology of coastal boundaries, the topology must be rebuilt.

Feature Adjacency and Connectivity

Topology has its basis in the principles of feature adjacency and connectivity. The most common method for retaining spatial relationships among GIS features is to record adjacency information within a topologic data model.

The topological data structure allows spatial relationships between geographic features to be easily derived. Due to this, the topologic model is the primary choice of a vector data structure currently used in GIS technology.

Most complex data analysis functions cannot be undertaken effectively, without the topologic vector data structure.

Chapter 3: Spatial Data Acquisition and Input

Datasets

Compiling Data

Assembling a properly compiled spatial database is essential to success in GIS implementation. The digital data will be the most expensive and critical component of the GIS, therefore attention should be given to the quality of the data acquired and the processes by which it is prepared.

Digital data which is acquired from federal and state government agencies are an excellent digital source of “paper to digital” converted data. Attribute data has a wide variety of potential data sources.

Any text-based or tabular data that can be referenced into a geographic feature, e.g. a point, line, or area, can be input into a GIS system’s database.

Sources of Datasets

There are numerous sources of datasets available throughout the Internet. Nearly any type of data can be found in digital format now, for importation into a GIS model.

Most are free or of a minimal charge. The amount of datasets available to the public is endless. There are a number of file formats, which can pose a challenge, but most are in convertible formats.

Below are just a few of the places to look for data:

FEMA Datasets

The FEMA website has “declared disaster relief” (DDR) datasets available in Shapefile, KMZ, WMS, and Excel tabular formats. Also available are metadata, and lyr formatted maps free for download.

The current system includes a full range of GIS services that provide sophisticated geospatial analytics through the Mapping and Analysis Center (MAC) and deployable GIS technology through the Deployable Emergency GIS program (DEGS).

Army Geospatial Center (AGC)

This is a sub-agency of the US COE, which coordinates, integrates, and synchronizes geospatial information and standards within the armed forces.

They provide direct geospatial support and products to the various branches of defense.

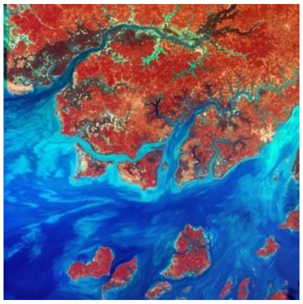
The AGC’s Geospatial Research and Engineering Division (GRED), is an engineering asset, which conducts R&D into geospatial data collection, processing, exploitation, and dissemination in support of both civilian and military missions.

This was formerly known as the Topographic Engineering Center (TEC).

Landsat Datasets

Beginning in 1972, until the most recent satellite launch (Landsat 8), in 2013, the Landsat program is the longest running project for acquiring satellite imagery of the Earth’s surface. The instruments on the eight Landsat satellites have acquired millions of images (image).

The images from Landsat are a unique resource for global change research and applications in agriculture, cartography, geology, forestry, regional planning, surveillance and education, and can be viewed through the USGS 'EarthExplorer' website.



Landsat 7 image of Guinea-Bissau
Image source: NASA.gov

MODIS Datasets

MODIS (on Terra and Aqua Satellites)

MODIS, or the *Moderate Resolution Imaging Spectroradiometer** is a key instrument aboard NASA's Terra and Aqua satellites.

Terra MODIS and Aqua MODIS view the entire Earth's surface every 1 to 2 days. As they pass over the earth, they acquire data in 36 spectral bands, or groups of wavelengths. This data plays a key role in the development of validated Earth system models used to predict global environmental changes.

- *Terra* **-** the orbit around the Earth is timed so that it passes from north to south across the equator in the morning
- *Aqua* - passes south to north over the equator in the afternoon.

*** Spectroradiometers** - These are devices designed to measure the spectral power distribution of a source. From the spectral power distribution, the radiometric, photometric, and colorimetric quantities of light can be determined in order to measure, characterize, and calibrate light sources for various applications. Spectroradiometers typically take measurements of spectral irradiance and spectral radiance.

Obtaining MODIS Data

MODIS data is freely available over the Internet, but prior to ordering data, the product descriptions and metadata should be reviewed thoroughly, in order to identify the most appropriate and accurate data.

- **Land** - A selection of land products are available at the LAADS site.
- **Ocean** - Oceanic products derived from MODIS, SeaWIFS, and other sensors and are available at the NASA OceanColor site.
- **Cryosphere** - The National Snow & Ice Data Center (NSIDC) is the site to learn about MODIS snow and ice products.

(SPOT) Systeme Pour l'Observation de la Terre Satellites

These satellites are a group of French high resolution optical imaging Earth observation satellite which has been in operation for over 20 years.

GNIS

"The Geographic Names Information System (GNIS), developed by the USGS in cooperation with the U.S. Board on Geographic Names (BGN), contains information for almost 2 million physical and cultural geographic features in the United States and its territories.

The Federally recognized name of each feature described in the database is identified, and references are made to a feature's location by State, county, and geographic coordinates. The GNIS is our Nation's official repository of domestic geographic names information."

Other Data Sources

Below are some more of the sources available for compiling the various layers for a given model:

- *(NLCD) National Land Cover Dataset* - The NLCD is a Landsat Thematic Mapper (TM) based classification of land cover in the US.
- *(NOAA) National Oceanic and Atmospheric Administration* - NOAA is the US government agency that oversees the development of national datums and several weather and ocean satellites.
- *(NWI – US Fish and Wildlife Services) National Wetlands Inventory* - The NWI is a dataset compiled by the US Fish and Wildlife Services that describes the type and extent of wetlands in North America.
- *(STATSGO) State Soil Geographic* - STATSGO is a coarse resolution digital soil dataset derived from more detailed soil survey maps. Areas without surveys were compiled using geological, topographic, vegetative, climatic, and Landsat-based data.
- *(TIGER – US Census Bureau) Topologically Integrated Geographic Encoding and Referencing* - TIGER is the central hub for US census data. TIGER also is an inventory of other spatial data related to rivers, lakes, buildings, cities, political areas and roads.
- *(USGS) United States Geological Survey* - USGS is a United States agency responsible for

Landsat satellites, nationwide map-making and spatial data development.

- *(FIPS) Federal Information Processing Standard* - FIPS is a federal code used to define political or physical features in the US. It was created as a standard for creating unique identifiers in data processing (similar to a social security number or 9 digit zip code).
- *(USGS - NHD) The National Hydrography Dataset* - Hydrological information located on the USGS website, this source provides datasets for lakes, ponds, streams, rivers, dams, stream-gauging, and flow modeling for the US.
- *(USGS) The Libre Map Project* - The Libre Map Project was begun to provide free access to various digital maps and related GIS data. It has an online collection of all digital USGS 1:24K scale topographic maps, as well as various other GIS data sources covering the US.
- *NPScape* – from the US Department of the Interior National Park Service, this is a landscape dynamics monitoring project that provides landscape-level data, tools, and evaluations for natural resource management, planning, and interpretation.
- *(USDA) NRCS soil data* – these are datasets available from the USDA Natural Resource Conservation Service.
- *(USGS) US National Atlas* - all raw data from the National Atlas. Everything from agricultural census data, election results, airports, railways, glaciers, arsenic content in groundwater, etc.
- *(USDA) ERS data Products* – this is data which comes from the Economic Research Service, including the Atlas of Rural and Small-Town America, Farm Program Atlas, Food Access Research Atlas, and Food Environment Atlas.
- *Public Land Survey System (PLSS)* - The PLSS is a land measurement system used in the western United States to define parcel boundaries and locations.

Data Input Procedures

The two forms of data input for GIS are attribute data input, and spatial data input. While the input of attribute data is fairly basic, as it deals with data in text form, spatial data input can be quite involved.

There are a number of methods that can be used to bring in spatial data to a GIS system. The choice of method depends on the application, project economics, and the type and complexity of the source data.

Four of the basic procedures used for inputting spatial data into a GIS are:

- Digitizing
- Scanning
- Coordinate Geometry (COGO)
- Conversion of existing digital data

Digitizing

Most of GIS spatial data entry is accomplished by manual digitizing (Figure 3.2).

GTCO Surface-Lit Digitizing Table
Image source: interworldna.com

The standard digitizer is an electronic grid device which consists of a table upon which the paper document is taped, and a cursor device which the operator uses to trace the spatial feature points electronically.

The operator traces the features on the tablet in much the same way a CAD operator would draw using a mouse with onscreen cursor, to input points along lines, polylines, etc.

In order to begin the process, the operator calibrates the table to the digital vector file by linking positional control points around the outer paper corners, with control points set up in the vector file (CAD for example).

This is usually done by using coordinate system or map projection points, in the horizontal and vertical planes to eliminate skewing within the process.

Manual digitizing has many advantages. These include:

- Low capital cost, as digitizing tables are reasonably priced
- Labor costs are low as digitizing is a easily learned skill
- Adaptable to different data types and sources
- Quality of data is high
- Typically offers a higher precision than the data need requires
- Ability to easily register and update existing data

Raster based GIS software data is occasionally digitized in a vector format and converted to a raster structure after the building of a clean topological structure. This procedure differs somewhat from the vector based software digitizing.

Automatic Scanning

A number of scanning devices exist for automatic capturing of spatial data. While several approaches exist for scanning of hardcopy documents, all have the advantage of being able to capture spatial features from a map quickly when compared with digitizing. Large format scanners are generally expensive to acquire and operate.

Scanning devices have limitations when it comes to the capturing of certain features, such as text and symbol recognition. Almost all scanned data requires a serious amount of manual editing and cleanup to create a clean and accurate data layer.

Other limitations of scanners include:

- Hard copy maps are often archived, preventing removal from the archiving facility to where a scanning device is available, as many companies or agencies cannot afford the scanning equipment and must outsource the scanning services
- Hard copy data may be in a form that is of poor quality for scanning capture, causing a great deal of noise in the final scan
- Geographic features may be too sparse on a single map to make scanning cost-effective
- Scanner may be unable to distinguish highly congested data features
- It is difficult to read unique labels (text) for a geographic feature effectively
- Scanning may be more expensive than manual digitizing, if excessive editing and cleanup is involved

Scanners work best when the information on the hard copy document is kept very clean, simple, and void of excessive graphic symbology.

Coordinate Geometry (COGO)

A third technique used for inputting spatial data, importing of coordinate points using coordinate geometry (COGO) procedures. This involves importing the point data collected from a total station, out of a survey data collector into the GIS model.

This method is useful for creating very precisely located cartographic definitions of property metes and bounds (and other miscellaneous parcel data), as well as utility features, and elevational data. This process is useful for cadastral or municipal types of data.

Converting and Modifying Existing Digital Data

A fourth technique for data input is the conversion of existing digital data sources. A good bit of spatial data, including digital maps, is available from a wide range of government agencies and private sources.

A number of data conversion programs exist, mostly from GIS software vendors, to transform data from CAD formats to a raster or topological GIS data format.

Given the wide variety of proprietary and non-proprietary data formats that exist, most GIS software and add-on providers have developed and provide data exchange or conversion software to go from their format to those considered common in the market place.

For those who are proficient in coding, most GIS software providers also provide an ASCII data exchange format specific to their product, and programming subroutine libraries that will allow users to write their own customized data conversion routines.

There are many other proprietary data formats that exist within the mapping and GIS industry. However, almost all provide data conversion to and from these formats:

- *IGDS* - Interactive Graphics Design Software (Intergraph)
- *DXF* - Drawing Exchange Format (Autocad)
- *DLG or Digital Line Graph* - (US Geological Survey)

Other Methods of Data Input (CAD-based)

- *Vectorizing of Raster Data* **-** If scanned raster data needs to be converted to a useable vector format, many higher end CAD programs, or 3rd party add-ons allow for hybridization of the file. This is accomplished by inserting the raster image into the vector file, and warping it to match vector control points. Then a vectorization operation is run to convert the raster data to vector elements. This requires a significant amount of element cleanup afterwards.
- *Using CAD Overlay for Tracing Vector Elements* **-** When the raster image has been properly inserted into the vector file and warped to the control points, vector elements can be traced over top of the read-only raster image, using grid snapping to maintain positional accuracy.
- *Hybrid CAD Procedures using Masking* **-** Some scanned images might only require a limited amount of modifying using vector type elements. Using the above process, regions of the raster images can be "masked-out" using polygon fence commands. Once the raster image is partially masked, vector elements can be drawn within the masked out area using standard CAD vector elements. This is a process which is mainly applied to CAD drafting, but has limited applicability for image data models within GIS.

Remote Sensing

Remote Sensing

This involves imagery of the earth's surface, taken from above, as in airplanes satellites, or more recently drones (remote sensing devices).

Imagery is available at different resolutions and can include non-visible energy such as ultraviolet.

Imagery can also be used to generate map-able data: roads can be traced or vegetation types delineated.

Satellite-based (Extraterrestrial Remote Sensing)

Satellite based remote sensing involves various processes of imaging from devices mounted on satellites in space, above the Earth's atmosphere.

These devices can take images of Earth at varying scales, using cameras and other devices which can capture data in a wide variety of light energy spectrums which the human eye cannot perceive, such as energy wavelengths or frequencies from x-rays, ultraviolet, visible, infrared, microwave to radio waves.

Aerial (Terrestrial Remote Sensing)

Aerial imagery is the capturing of the ground surface from an elevated/direct-down position. This usually doesn't include images captured from a ground-based structure.

Platforms for aerial photography include fixed-wing aircraft, helicopters, Drones (UAVs), balloons, blimps and dirigibles, and even kites or parachutes have been used.

Airplanes have been used for aerial remote sensing for well over a century, used heavily in defense reconnaissance in many wars, but UAV's are becoming the platform of choice, due to the low cost and flexibilities.

Remote Sensing by UAV (Unmanned Aerial Vehicle or Drone)

Satellite imagery is readily available, at low resolution from Landsat and MODIS for example, to higher resolution from sources such as WorldView and Quickbird.

However, these imagery sources are occasionally unable to offer sufficiently high enough resolution, cover a specific area of study, or capture the time series necessary to fulfill the entire purpose of a given project.

The relatively low-cost of high resolution image capture capability of UAVs creates the potential for them to fill the data gap between satellites, airplanes and ground surveying.

A UAV can do much more than image acquisition, occasionally making them advantageous over standard satellite or aircraft image acquisition, especially in situations where manned remote sensing may be dangerous.

Drones can fly in high-risk areas safer than manned aircraft, and can traverse areas, such as inaccessible shorelines or hurricanes.

Due to size and aerodynamics, UAVs are able to fly at lower altitudes, collecting more precise information than manned aircraft or satellites. This also means that they can fly below clouds making them advantageous in tropical areas where clouds can often impede satellite image collection.

UAV with a DSLR Digital Camera

Innovative Field Applications

UAVs have the potential to supplement data collection efforts and contribute to existing data inventory.

Specific environmental and ecosystem applications suitable for the use of a drone can range from precision agricultural management, to coastal geomorphological mapping, to tracking soil erosion, or species and habitat monitoring, or observing sea ice conditions and tally wildlife populations.

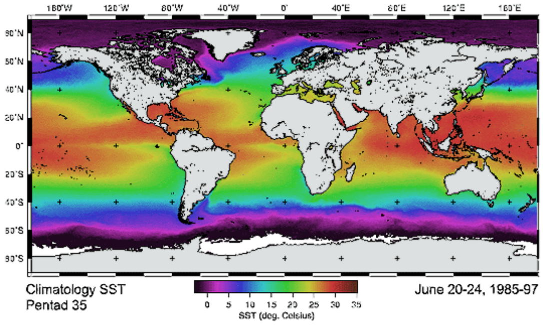
Environmental Uses of Drones

- Disaster Risk Management
- Disaster Risk Mitigation
- Illegal Activity Monitoring
- River erosion Flooding risk
- Map impacted areas
- Poaching
- Migration patterns
- Deforestation Landslide risk
- Broadcast messages
- Illegal fishing
- Endangered species status
- Urban expansion
- Volcano eruption risk
- Monitor forest fire spread
- Illegal trade Agriculture

The following is some terminology encountered in remote sensing:

- *Enhanced Thematic Mapper (ETM+)* – This is a sensor equipped on Landsat-7 which generates images of the Earth in 8 spectral bands. The blue, green, red, NIR and mid-infrared (MIR) have 30m resolution (bands 1-5, 7). The panchromatic (band 8) has 15 m resolution. The thermal band is 60 m resolution.
- *Electromagnetic (EM) Spectrum* – The EM spectrum refers to the range of energy wavelengths or frequencies from x-rays, ultraviolet, visible, infrared, microwave to radio waves.
- *Multispectral Imagery* – A multispectral image is a multi-channel raster consisting of several spectral bands of wavelengths. Example: Red, green, blue and NIR.
- *Nadir* - is the point directly below the aircraft which is usually near the center of the aerial image.
- *Noise* - In remote sensing, noise is any disturbance in a frequency band, an irregular, sporadic, or random oscillation in a transmission signal, or random or repetitive events that interfere with communication.
- *Wavelength* – An electromagnetic wave produce sinusoidal patterns with distinguishable shape and length. A wavelength is the distance between the peak of a wave and its successive wave.
- *Infrared Imaging* - An infrared image represents reflectance grid cells that are recorded in the near-infrared wavelengths of the light spectrum.
- *Active Sensors* - Active sensors illuminate a target and measure the reflected backscatter that returns back to the sensor.
- *Passive Sensors* - Passive remote sensing measure natural energy from the sun as reflected sunlight or thermal radiation. Passive sensor examples are Landsat, SPOT and GeoEye.
- *Atmospheric window* - An atmospheric window is the wavelength at which electromagnetic radiation (sunlight) from the sun will penetrate the Earth's atmosphere overall constricting these spectrum bands from reaching the Earth.
- *Orthoimage* - an air photo or satellite image that has been processed to remove perspective distortions. Distortions of tilt and relief are removed so that all features in an orthoimage are in their true orthographic positions.
- *Orthophoto* - an air photo that has been scanned, rectified, and reconstructed to represent its features in a map projection or at least in a flat rectangular form without the usual distortions of geometry and perspective. Usually orthophotos are prepared from very high resolution stereo pairs. an orthophotograph is a vertical photograph with an orthographic view using geometry and measurements to reduce tilt, terrain and perspective distortions.
- *Overlay* - (vector or CAD) a transparent layer placed on an underlying image. The overlay is where symbols, annotations, or image traces can be created or displayed without changing the underlying image.
- *Shaded Relief* – This is a map that displays the brightness and shadows of terrain reflection given a sun angle and direction of sunlight.
- *Photogrammetry* **-** The science of making reliable measurements of physical objects and the environment by measuring and plotting electromagnetic radiation data from aerial photographs and remote sensing systems against land features identified in ground control surveys, generally in order to produce planimetric, topographic, and contour maps. Parallax is term used in photogrammetry that describes the apparent shift in relative positions of Earth features when it viewed in different locations.

- **Planimetric map** - A map designed to portray the horizontal positions of features; vertical information is specifically ignored.
- *Quad, quadrangle, map quad or map quadrangle* - The geographic area covered by a map. One kind of map quadrangle is the 7.5' x 7.5' area that is covered by a standard USGS 7.5' topographic map. Referring to a 7.5' map quadrangle does not imply the presence of an actual paper map. The term may simply designate the area covered by electronically stored materials.
- *(CIR) Color-infrared image* **–** a form of image capture at the light spectrum between green through infrared. The photographic infrared radiation just beyond the range of human vision is then displayed as red. Normal red from the scene becomes green, and green becomes blue. Normal blue in the scene is filtered out and not recorded. CIR images are used to show the vigor of plant life. Healthy vegetation appears red, while distressed or damaged vegetation may look pink, tan, or yellow.
- *Sidelap (or side overlap)* - consists of the overlapping edge areas of photographs between adjacent flight lines.
- *Endlap* - This is the overlap in aerial photographs from end-to-end between flight lines. Flight lines are the paths that aircrafts take for complete coverage of an area. These flight lines are positioned to give endlap between succession photos.
- *Stereo Pairs* - Stereo pairs are overlapping photos taken at different positions but of the same area. Parallax is used to interpret height differences within the overlapping area.
- *IKONOS* – This is a high resolution commercial imaging satellite that provides one meter panchromatic and three meter multispectral (blue, green, red, near-infrared) imagery.
- *Advanced Very High Resolution Radiometer (AVHRR)* – AVHRR is operated by the National Oceanographic and Atmospheric Administration (NOAA) and collects infrared, visible and thermal images with an approximate 1 kilometer spatial resolution cell size. The primary purpose of these instruments is to monitor clouds and to measure the thermal emission of the Earth. These sensors have proven useful for a number of other applications, however, including the surveillance of land surfaces, ocean state, aerosols, etc. AVHRR data are particularly relevant to study climate change and environmental degradation because of the comparatively long records of data already accumulated (over 20 years). The main difficulty associated with these investigations is to properly deal with the many limitations of these instruments, especially in the early period (sensor calibration, orbital drift, limited spectral and directional sampling, etc.).



AVHRR Image

Image Source: NASA

Data Verification

Six steps in the verification and editing of spatial data:

- *Visual review* - By running a check plot of the map
- *Cleanup of lines and junctions* - is usually done by software first and interactive editing second.
- *Weeding of excess coordinates* - involves the removal of redundant vertices in linear and/or polygon features
- *Correction for distortion and warping* - most GIS software has functions for scale correction and rubber sheeting. The specific rubber-sheet algorithm used will vary

depending on the spatial data model, vector or raster, employed by the GIS system. Some raster-based techniques may be more labor intensive than the vector-based

- *Construction of polygons* – Since much of data used in GIS is polygonal, the construction of polygon features from lines/arcs is necessary. Usually this is done in along with the topological building process
- *Addition of unique identifiers or labels* - this process is performed manually, with some systems providing the capability to automatically build labels for a data layer

These data verification steps occur following the data input stage and proceeding or during the linkage of the spatial data to the attributes.

Data verification ensures that the integrity between the spatial and attribute data is maintained.

The data verification process should include a brief querying of attributes and cross checking against known values.

Chapter 4: Spatial Data Storage and Retrieval

The Data Storage and Retrieval System

A major component of the GIS is the data storage and retrieval system. This subsystem organizes the data, (spatial and attribute), in a form which allows for quick retrieval for the purposes of updating, querying, and analysis.

Many GIS platforms use proprietary software for their spatial editing and retrieval system, while using an external database management system (DBMS) for their attribute data storage requirements. An internal data model is usually employed to store the primary attribute data associated with the topological definition of the spatial data.

These internal database tables contain primary columns such as:

- area
- perimeter
- length
- internal feature id number

Thematic attribute data is usually maintained in an external DBMS that is linked to the spatial data through the use of the internal database table.

Spatial Database (or Geodatabase)

A spatial database is a database which has been optimized for the storing and querying of data that represents objects defined in a geometric space. Most spatial databases are able to handle the processing of simple geometric objects such as points, lines and polygons, while others are able to handle more complex structures such as topological surfaces, linear networks, 3D objects, and TIN surfaces.

Common everyday databases are designed to manage various numerical and character data types. However additional functionality is required for the efficient processing of spatial data types (geometry or features). The Open Geospatial Consortium (OGC) created the “Simple Features” specification which establishes the standards for adding spatial functionality to database systems.

- **Simple Features** (officially known as “Simple Feature Access”) is both an OGC and ISO standard (ISO 19125). This standard specifies a common storage and access model used for 2-D geographical data features (point, line, polygon, multi-point, multi-line, etc.)

- The ISO 19125 is a two part standard:
    - The first section defines a model for 2-D simple features, with linear interpolation between vertices. This data model is a hierarchy of classes.
    - The second section of the standard, defines an implementation using SQL. The OpenGIS standards also cover the implementations in CORBA and OLE/COM, though these have waned in popularity behind the use of SQL, and are not standardized by ISO.

  The ISO/IEC 13249-3 SQL/MM Spatial extends the Simple Features data model, with circular interpolations (e.g. circular arcs), and other features such as coordinate transformations, and means for validating geometries, as well as providing support for the GML syntax.

- **The Geography Markup Language (GML)** – This is the XML syntax which was defined by the OGC to express geographical features. GML serves as an interchange language for GIS systems, as well as an open interchange format for geographic transactions over the Internet. The key to GML's utility is its capacity for integrating all types of geographic information, including not only conventional "vector" or discrete objects, but coverages, and sensor data. GML consists of a set of primitives which are used to build application specific schemas or application languages.

These primitives include:

- ○ Features
- ○ Geometry
- ○ Coordinate reference system
- ○ Topology
- ○ Time
- ○ Dynamic feature
- ○ Coverage (including geographic images)
- ○ Unit of measure
- ○ Directions
- ○ Observations
- ○ Map presentation styling rules

- **KML** – This is a markup language made popular by Google Earth. Whereas GML is a language for encoding geographic content for any application, by describing a variety of application objects with related properties (e.g. bridges, roads, buoys, vehicles etc.), KML is a language for visualizing geographic information created specifically for the Google Earth platform.

- **PostGIS Spatial Database -** This is an open source spatial database extension which adds the needed support for geographic objects to be managed within the PostgreSQL object-relational database server. PostGIS "spatially enables" a PostgreSQL based server, which allows it to be utilized as a backend spatial database for GIS use, much like the ESRI SDE or Oracle Spatial database extensions.

  PostGIS follows the OpenGIS "Simple Features Specification for SQL" standard and has been certified as being compliant with the "Types and Functions" profile. Development by Refractions Research as a project in open source spatial database technology, PostGIS has been released under the GNU General Public License.

Data Compression vs Data Compaction

Both of these encoding operations are generally the same, in that each one results in more compacted storage and reduced file sizes, however applied to file geodatabases, both use unrelated operations.

Compaction

- Involves cleaning up of the record storage within the files by reordering them and eliminating gaps or free space
- Is recommended for file where data is frequently added or deleted
- Can reduce file sizes and improve performance of the database
- Does not affect the read/write capabilities of the file
- Performed on a routine basis

Compression

- Reduces file storage requirements
- Improves the performance of the database
- Makes database or feature classes read-only within the file
- Performed on an as-needed basis

Organizing Data for Analysis

Most GIS platforms use the *thematic approach* to organizing the spatial data, which categorizes data within vertical layers.

Layers tend to have common titling based on their specific applications. For example, the typical layers used in the forestry industry may include timber species, soil types, elevation, roads, endangered species regions, hydrology, etc.

Spatial data layers are usually input into the system, one at a time. Attribute data is also entered one layer at a time. Depending on the attribute data model used by the data storage subsystem data must be organized in a format that will allow for the manipulation and analysis operations that will be needed.

Often times, the spatial and attribute data can be entered at varied stages, and linked at a later time though this is dependent upon the particular sources of data.

Spatial Data Layers - Vertical Data Organization

In most GIS systems, the data is organized in themes as data layers. This approach allows input of data as separate themes for overlaying based on analysis needs.

While the overlay/layer approach used in CAD systems is used for separating major classes of spatial features, this concept is also used to order data in GIS in a logical way.

Many different terms are used when defining data layers in GIS, such as themes, coverages, layers, levels, objects, and feature classes. Data layer and theme are the most commonly used and least proprietary to any given proprietary GIS software.

In a GIS project, a diverse variety of data layers will be needed. These must be established prior to beginning a project, with priority given to the input or digitizing of the spatial data layers. This is required,, as many time, one data layer will contain features that coincide with another. Data layers can be defined by the user, and are commonly established based upon the scope of the particular project.

In considering the physical needs of the GIS software, it should be understand that two types of data are required for each layer, attribute and spatial data.

Commonly, data layers are input into the GIS one layer at a time. As well, often a data layer is completely loaded, (graphic conversion, editing, topological building, attribute conversion, linking, and verification), prior to the next data layer beginning.

As there are a number of steps involved in completely loading a data layer, it can become very confusing if various layers are loaded all at once.

Properly identifying layers before beginning data input is an important consideration. This is often achieved through a user needs analysis.

The user needs analysis performs several functions including:

- identifying the users
- educating users with respect to GIS needs
- identifying information products
- identifying data requirements for information products
- prioritizing data requirements and products
- determining GIS functional requirements

Cost-benefits Analysis

Often a user needs assessment will include a review of existing operations, (sometimes called a situational assessment, and a cost-benefit analysis).

The cost-benefit process is well established in conventional data processing and serves as the means to justify the acquisition of hardware and software. It defines and compares costs against potential benefits.

The following list of sample data layers, to illustrate data layers that might be used for a typical operational forestry GIS project.

Sampling of the different forest project data layers:

- Thematic Data
- Forest Inventory
- Soil Types
- Wildlife Habitat
- Plant Species Delineation
- Wetlands Delineation
- Physical land classification
- Ecological land classification
- Elevation
- Positional Reference (Base Map) Data
- Survey / legal description
- Hydrography data (such as lakes, streams, and rivers)
- Transportation route network (access roads)

- Administrative boundaries (forest management units, fire control)

Most GIS projects integrate data layers to create derived themes or layers that represent the result of some calculation or geographic model, (timber pricing, land use requirements, etc). Derived data layers are completely dependent on the mission of the particular GIS project.

Each separate data layer would be input individually, and topologically integrated to create the combined data layers. Based on the data model, (vector, raster, TIN), and the topological data structure, selected data analysis functions could be performed. It should be noted that in vector based GIS software the topological structure defined can only be traversed by means of unique labels to every feature.

Spatial Indexing - Horizontal Data Organization

Spatial indexing refers to the proprietary organizing of data layers in a horizontal fashion within the GIS system. Spatial indexing is the method utilized by the software to store and retrieve spatial data. A variety of different approaches exist to speed up the spatial feature retrieval process within a GIS system, most involving the partitioning of the geographic area into manageable subsets or tiles.

Tiles are then indexed mathematically, to allow for a quick search and retrieval operation when querying is initiated by the user.

Spatial indexing can be compared to the definition of map sheets, except that specific indexing techniques are used to access data across map sheet (tile) boundaries. This is done to improve the query performance for large data sets that span multiple map sheets, and to ensure data integrity across map sheet boundaries.

The method and process of spatial indexing is usually transparent to the user, though it becomes very important especially when large data sets are utilized. The notion of spatial indexing has become increasingly important in the design of GIS software, as larger scale applications have been initiated using GIS technology.

Users have found that often the response time in querying very large data sets is tediously slow. This has remedied, as GIS developers by created sophisticated algorithms for the indexing and retrieval of spatial data.

Raster systems, by the configuration of their data structure, do not typically need a spatial indexing method. The raster approach creates regular, readily addressable partitions on the data, via its data structure. However, more sophisticated vector GIS does require a method to quickly retrieve spatial objects.

The horizontal indexing of spatial data within GIS software involves several issues, which concern the extent of the spatial indexing approach.

They include:

- the use of a library system to organize data for users
- the need for a formal definition of layers
- the need for feature coding within themes or layers
- requirements to maintain data integrity through transaction control, e.g. the locking of selected spatial tiles (or features) when editing is being performed by a user with the appropriate permissions

While all these issues need not be satisfied for spatial indexing to occur, they are important aspects to consider when deciding on a GIS platform. While the spatial indexing method is usually not the key point for deciding upon a GIS system, users should consider these requirements, especially if very large data sets, such as if thousands of polygons are to be typical in their applications, and a vector data model is to be employed.

Editing and Updating of Data

The primary function in the data storage and retrieval system is editing and updating of data. *The following data editing capabilities are commonly required:*

- The interactive editing of spatial data
- The interactive editing of attributtal data
- Ability to add, manipulate, modify, and delete both spatial features and attributes (independently or simultaneously)
- Ability to edit selected features with a batch processing mode

Dataset Updates

Updating involves more than just the editing of features; it is the resurvey and processing of new information and is of considerable importance during any GIS project. The life span of most datasets can range anywhere from 1 to 10 years, with the validity of the data being from 5 to 10 years. This long time span is a result of the lengthy and labor intensive task of capturing and inputting data.

Often routine data updates are needed that involve an increased accuracy and detail of the data layer. Changes in the classification standards and procedures may bring upon the need for these updates. For example, the changes and updates needed for a coastal shore line data layer to show the changes which occurred from an active hurricane season.

Many times data updates are needed based on the results of a derived GIS product. The generation of a derived product can show obvious errors within the data or inaccurate classifications for a given data layer.

Many times, the data update process will be a direct result of a physical change in the geographic landscape, such as the hurricane example. With this type of update new features are usually required for the data layer, such as a new erosion polygon layer. Additionally, existing features may need to be altered, (such as dune crossovers to be relocated). There is a strong need for a record keeping capability with this type of update process to show the history of updates and alterations to the data.

Users should take this requirement into consideration and design their database organization to accommodate these needs.

Data Retrieval and Querying

Xx

Being able to query and retrieve data, based upon user-defined criteria is a necessary function of the data storage and retrieval system. Data retrieval involves the ability to easily select data for graphic or attribute editing, updating, querying, analysis and display operations.

The ability to adequately retrieve data is based on the unique structure of the DBMS, and command interfaces are commonly provided with the software. Most GIS software also provides a macro language, so that users can write their own specific data retrieval routines when necessary.

Data Querying using SQL

Querying is the ability to retrieve data, such as a data subset, outlined by user-defined parameters. These data subsets are referred to as *logical views*. Often the querying is closely linked to the data manipulation and analysis component of the GIS system. Many GIS software programs have standardize their querying capabilities by using SQL, especially systems that use an external Relational Database Management System (RDBMS). The trend in the GIS industry is towards the use of a generic interface with external relational DBMS's. The use of an external DBMS, linked via a SQL interface, has becoming the industry standard for GIS querying.

By using SQL, GIS software can interface to a variety of different external DBMS systems, thus providing the user with the flexibility to select their own DBMS. This allows the organization to use an existing DBMS that is presently being used for other business needs.

The idea of integrating the GIS software to utilize an existing DBMS through standards is referred to as corporate or enterprise GIS. With the evolution of GIS technology, from it beginning as a research tool to being a decision support tool, there is a potential for it to be totally integrated within existing corporate functions, such as accounting, reporting, etc.

GIS File Formats

There are a wide variety of file formats used within the GIS environment. Raster and vector data files, ESRI files, database files, conversion or exchange files, GPS, COGO, etc. all have their own specific type of file extension. Some of the data formats common to the GIS marketplace are listed below. Note that most formats are only used for spatial data, as attribute data is usually handled as ASCII text files.

Data exchange or intechange

In order to convert datasets from one file type to another requires the use of data exchange or interchange syntax. This is the process of taking data which is structured under a “source schema” and transforming it into data structured under a “target schema”, so that the new data is an accurate representation of the original data. With data exchange, the data is actually restructured, which runs the risk of possible loss of content through the process.

Some file formats and codes used in data exchange:

- **MIF** – Mapinfo Data Interchange
- **GML** – Geography Markup Language
- **GPX** – GPS data exchange
- **IGDS** – Interactive Graphics Design Software (Intergraph) – This binary format is the standard for Intergraph proprietary software platforms such as Microstation (CAD) or GeoMedia (GIS). It has become the standard in Canada’s mapping industry, a majority of the US state DOT’s (transportation departments), and many of the electrical utility industry firms and municipalities. It is a proprietary format, however most GIS software vendors provide DGN translators.
- **E00** – ARC/INFO interchange
- **DXF** – Drawing Exchange Format (Autocad) – This ASCII format is used mainly to convert to and from the Autocad drawing format and is a standard within the engineering community. Most GIS software vendors provide a DXF translator.
- **XML** - Developed by the World Wide Web Consortium (W3C), XML is a standard for designing text formats that facilitates the interchange of data between computer applications. XML is a set of rules for creating standard information formats using customized tags and sharing both the format and the data across applications.
- **ASCII** - The term "ASCII file" is often used to mean a text-only or plain text file. Documents in most word processors are not text-only files, since they include header information and formatting characters. However, most word processors have an export or print-to-file utility that will convert a document into a text-only ASCII format. Converting to “plain text” is the best way to “filter” formatted textual data, such as text from MS Word.
- **GENERATE – ARC-INFO Graphic Exchange Format** – A generic ASCII format for spatial data used by the ArcGIS software suite to accommodate generic spatial data.
- **EXPORT – ArcGIS Export Format** – An exchange format that includes both graphic and attribute data. This format is used for transferring ArcGIS data from one hardware platform, to another. It is also often used for archiving ArcGIS data.

Spatial Data Relationships

Within the context of GIS, spatial data relationships are important concept to comprehend. In particular, the relationship between the ways geographic features interact is a complex one. This is of concern as the principle role of a GIS system is the manipulation and analysis of large amounts of spatial data.

The accepted theoretical solution is to “topologically structure” the spatial data. The topologic data model best reflects the geography of real world features, providing an effective mathematical basis for encoding spatial relationships, and providing a data model for manipulating and analyzing vector based data.

Most GIS software platforms divide spatial and attribute data into separate data management systems. Usually the topological or raster structure is used to store the spatial data, while a relational database structure is used to store the attributtal data. Data from both structures are linked together for use through unique identification numbers, (feature labels and DBMS primary keys).

The linking of the spatial features with an attribute record is usually maintained by an internal number assigned by the GIS software.

A label is needed so that the user may load the correct attribute record for a particular geographic feature.

Often a single attribute record is automatically created by the GIS software once a clean topological structure has been generated.

This attribute record normally contains:

- the internal number for the feature
- the user's label identifier, the area of the feature
- the perimeter of the feature (linear features have the length of the feature defined instead of the area).
-
- Spatial Indexing - Horizontal Data Organization
- Editing and Updating of Data
- Types of GIS File Formats
- Converting Data from One Format to Another - Data exchange or interchange
- Spatial Data Relationships

Chapter 5: Spatial Data Cleanup and Analysis

Overview

Data Cleanup and Analysis

The first part of this chapter deals with data “cleanup”; the manipulation and transformation processes used to prepare data sources for GIS system integration and spatial analytical functions.

The second part addresses some basic types of spatial analytical processes used once the data has been properly cleaned up and GIS formatted.

A major difference between GIS and CAD mapping is the ability within GIS to transform the original spatial data in order to perform analytical queries. Some transformation capabilities are common to both GIS and CAD systems.

However, the topological data structure in the GIS software and related geodatabase, provides a larger range of analysis capabilities used to operate on the spatial features and aspects of the geographic data, on the non-spatial attributes of these data, or on both.

The main criteria used to define a GIS are its capability to transform, and ability to integrate spatial data.

Spatial Data Editing (Cleanup)

Cleaning up the Data

After data has been input using the methods afore mentioned, the process of editing and verifying the newly input data is required to clean up the errors that occurred in the input process. This is a time consuming, combing-through process that can consume many hours of pain-staking labor.

Data Errors

Common errors that can occur during data input can be classified as:

- *Incomplete spatial data* - includes missing points, line segments, and/or polygons
- *Positional placement errors* - result from careless digitizing or poor quality of the original data source
- *Distorted spatial data* - caused by poorly rubber-sheeted aerial photographs, or paper documents stretched during the roller scanning process.
- *Incorrect linking between spatial and attribute data* - This type of error is commonly the result of mislabeled unique identifiers (labels) being assigned during the manual key-in or digitizing processes.
- *Attribute data is wrong or incomplete* - When attribute data does not exactly match up with the spatial data. Frequently from non-verified independent sources, or from different timeframes. Missing data records or too many data records are the most common problems of this type.

Identifying errors in spatial and attribute data is often difficult, with many errors becoming evident during the topological building process. The use of check plots to clearly determine where spatial errors exist is good practice. Most topological building functions in GIS software clearly identify the geographic location of the error and indicate the nature of the problem.

Some GIS software packages allow users to graphically perform a walk-through and edit the spatial errors, while others just identify the type and coordinates of the error.

As this is often labor-intensive, users should consider the error correction capabilities very important during the purchase of GIS software.

Conversion Errors

A variety of common data problems occur in converting data into a topological structure. These problems occur due to the quality of the original source data and the nature of the data capturing process.

A majority of time and effort in building a GIS framework is putting data into the proper structure in order to work well for GIS analysis purposes.

Cleanup Tools

Most GIS software has tools to clean the data and build a topologic structure, through automated, semi-automated, or manual means. If the data has never gone through any preparation or restructuring processes, then the cleaning process can be very time-consuming.

Topological errors only exist with linear and areal features, being most obvious with the polygonal features.

The most common problems that occur in converting data into a topological structure include:

- Slivers and gaps in the line work
- Dead ends, or ("dangling arcs"); ie. over-shoots and under-shoots in the linework
- Bowties or twisted polygons which arise from inappropriate closing of connecting features
- Duplicate lines

Slivers

This is one of the most common problems encountered when cleaning data. Slivers can occur when side by side boundaries are digitized separately, by following a polygon overlay, or when combining data from different sources.

By digitizing data layers with respect to an existing data layer, rather than attempting to match data layers later, slivers can be avoided.

Dead ends

These usually occur when data has been digitized in spaghetti mode, or without snapping to existing nodes.

Many GIS software programs have tools which will clean up the under-shoots and over-shoots based on a user defined tolerance, such as a *proximity distance*.

Bowties

During topological building, the definition of an inappropriate distance often leads to the formation of bow ties or twisted polygons. Setting a tolerance too high will force arcs to connect when they should not have been joined, resulting in small polygons called bow ties.

To set a proper tolerance for cleaning requires knowledge of the scale and accuracy of the data set. Most bow ties occur when incorrect tolerances are set during the automated cleaning of data, with too many over-shoots.

Most GIS software provides a tool to eliminate bow ties and slivers using a semi-automated feature-elimination command based on area.

Duplicate lines

Another problem that occurs when building a topologic data structure is the duplication of lines.

These usually occur when data has been digitized or converted from a CAD system. The lack of topology in CAD, allows for the accidental creation of elements that are duplicate. However, most GIS packages have a means for automatic eliminating of these duplicate elements during the topological building process.

One type of duplicate line to be aware of is, the duplicate element that retraces itself, such as a three verticed line where the starting point is also the ending point.

Some GIS programs are unable to identify these feature glitches and will build the feature as a valid polygon. This is because the topological definition is mathematically correct, though not geographically correct.

Manipulation and Transformation of Spatial Data

Maintenance and transformation of spatial data involves the ability to first input the data, then manipulate and transform data once it has been created.

Coordinate Thinning

Coordinate thinning involves reducing the coordinate pairs, (X-Y) from arcs. This is needed when data has been captured with excessive vertices in the linear features which results in redundant data and large data volumes. *Coordinate thinning is performed on various features, such as:*

- contours
- hydrography
- routes
- boundaries
- other linear and polygon features

Map Generalization

This process involves decreasing the amount of detail on a map, so that it will remain uncluttered following a scale reduction. The coordinate thinning of coordinate pairs is also required in the map generalization process of linear simplification.

Linear simplification is one component of map generalization that is required when data from one scale, such as 1:10,000, is to be integrated with data from another 1:50,000.

Geometric Transformations ("Rubber-sheeting")

Also known as "rubber-sheeting", this process deals with the registering of a data layer to a common coordinate scheme. This usually involves registering a new set of data layers to an existing standard data layer which has already been registered.

Rubber sheeting involves stretching one data layer to meet another based on predefined control points of known locations.

This procedure preserves the interconnectivity, or topology, between points and objects through stretching, shrinking, or reorienting their interconnecting lines.

- *Warping* - Two other functions may be categorized under geometric transformations. These involve warping a data layer stored in one data model (raster or vector), to another data layer stored in the opposite data model. For example, often satellite imagery may require warping to fit an existing data layer, or a poor quality vector layer may require warping to match a more accurate raster layer.
- *Rectification* **–** This process involves removing the geometric distortion from a raster or a vector data source. This is usually achieved by aligning raster features or vector coordinate positions with features in a base map or other coordinate reference framework. The rectification procedures can be used to bring multiple distorted image segments into a common framework so they can be integrated with a larger image or image series.

Map Projection Transformations

This operation involves transforming geographic coordinate data from an existing map projection to another map projection.

Most GIS analysis requires that data layers be in the same map projection as one another for proper spatial analysis. If the data is acquired in a map projection that differs from the other data layers, it will require transformation.

GIS software platforms can typically support twenty or more map projections.

Conflation - Sliver Removal

Conflation is a procedure of reconciling the positions of corresponding features in different data layers, referred to as "sliver removal".

This occurs when two layers contain the same feature, but with differing boundaries for that feature. This may be caused by lack of coordination or insufficient prioritization of data during the digitization stage, or by a number of different manipulation and analysis techniques.

When these two layers are combined, such as through polygon overlay, they will not match precisely and small sliver polygons will be created. Conflation concerns the process of removing these slivers and reconciling the common boundary.

The most common approach to sliver removal is for the user to define a priority for data layers in combination with a set tolerance value. After polygon overlay operation, if a polygon is below the size tolerance it is classified as a sliver.

To resolve this situation, the arcs of the data layer that have higher priority will be retained, while the arcs of the other data layer will be deleted.

Another approach is to simply divide the sliver down the center and collapse the arcs making up the boundary. It is generally cheaper in the long run, to reconcile maps manually in the map preparation and digitization stages, than afterwards.

Edge Matching

Edge matching is the process of adjusting the position of features that extend across map sheet boundaries. In theory, data from adjacent map sheets should precisely align at the map edges; however, in practice this rarely happens. Edge matching always requires some level of interactive editing, though GIS software is able to automate this process to some degree.

Misaligned features can be caused by several factors including:

- Digitizing error
- Paper shrinkage of source maps
- Errors in the original mapping

Interactive Graphic Editing

Interactive graphic editing functions involve: adding, deleting, moving, and changing of the geographical position of features.

Most graphic editing occurs during data compilation, which consumes the lion's share of the time required to complete a given GIS project.

Most of the editing that is undertaken involves the cleaning up of topological errors identified earlier. The ability to snap to existing elements, such as nodes and arcs, is vital.

The user interface and use-ability of the editing functions differs greatly among various GIS programs, though they all ultimately perform similar functions.

Even so, it may be cost-effective to invest in better GIS manipulation tools upfront, rather than spending thousands of wasted hours in data cleaning, by using cheap, cumbersome software programs.

Integration and Modeling of Spatial Data

The integration of data allows for the capability to perform complicated spatial queries that could not be otherwise be ascertained, (inventory or locational questions such as “how much” or “where at”).

These types of questions rely on the combination of several different data layers, in order to be able to provide a more complete and logical query result. Being able to combine and integrate datasets is vital to GIS querying.

- *Spatial modeling* **-** Applications often require a more sophisticated approach to answering complex spatial queries. The technique used to solve these questions is called spatial modeling. Spatial modeling uses spatial characteristics and methods in manipulating data. Methods exist to create a wide range of capabilities to analyze data by combining sets of primitive analysis functions. In areas such as resource planning and inventory compilation, the use of GIS spatial modeling tools has helped to quantify processes and lay out models for deriving analytical results.
- *Raster cell processing* **-** The raster data model is the primary spatial data source used for analytical modeling of quantitative analysis of numerous data layers. To accommodate these raster modeling techniques, most GIS software programs will use a separate module just for for cell processing.

Spatial Analysis

Spatial Analyzing

Spatial analysis is the application of statistical analysis and other analytic techniques to data which has a geographical or spatial aspect.

This type of analysis typically employs software which is able to render maps, process spatial data, and apply analytical methods to terrestrial or geographic datasets, including the use of GIS and geomatics (which is the gathering, storing, processing, and delivering of geographic or spatially referenced information).

Geospatial analysis, through GIS, was originally developed to analyze problems within the environmental, ecological, geological and epidemiological disciplines of science.

It is now used by almost all industries for gathering and processing large volumes of datasets on:

- Defense and Intelligence Gathering
- Electrical Power Distribution Grid
- Oil and Gas
- Forestry
- Sales and Demographical
- Geopolitical
- Insurance and Medicine
- Public safety
- Emergency Management (911 Dispatch)
- Crime Data
- Disaster Risk Reduction and Management (DRRM)
- Climate change adaptation (CCA)

Spatial statistics typically gains results primarily from the visual observations of mapped data rather than experimentation.

Vector-based Analysis

Vector-based analysis is typically related to operations such as:

- *Map overlay* - the combining of two or more maps or map layers based on a set of predefined rules
- *Simple buffering* - identifying analytical regions on a map, within a specific distance of one or more features, such as city boundaries, easements, right of ways, roads, lakes, etc.

Raster-based Analysis

Raster-based Analysis is generally used for environmental and remote sensing based analysis, consisting of numerous operations which can be applied to the grid cells of one or more maps (or images). These analysis operations involve filtering or algebraic forms of analytical operations.

These techniques involve processing one or more raster layers based on a set of parameters, which result in a new map layer.

For example:
replacing each cell value in a raster region, with a combination of neighboring values, or computing the sum or difference of specified attributtal values for each grid cell, in two matching raster datasets.

Descriptive Statistics

Statistical analysis operations such as cell counts, means, maxima, minima, variations, cumulative values, frequencies and a number of other measures and distance computations are considered as spatial analysis.

Spatial analysis includes a large variety of statistical techniques to describe, explore, or explain that applies to data that varies spatially and temporally.

Advanced statistical techniques can be used to determine clustering patterns of spatially referenced data, as well.

Advanced Analysis Operations

Geospatial analysis can go beyond just 2D and 3D mapping operations, or spatial statistics.

Other forms of geospatial analysis:

- *Surface analysis* – the analyzing of physical surface properties; such as gradient, aspect and visibility, and analyzing surficial data fields.

- *Network analysis* – the examination of the properties of natural and man-made networks in order to comprehend the behavior of flow, in and around these networks.
- *Locational analysis* - network analysis within a GIS frame, may be used to address a variety of practical problems like route selection and facility location. Problems which are not specifically network constrained, such as new routings, business locations, point feature positioning (ie. power poles, transformers), or the siting of businesses, may be effectively analyzed without referencing existing physical networks.

Integrated Analytical Functions in a GIS

Most GIS systems have the ability to build complex analytical models through the combination of basic analytical functions, with most systems providing a standard set of these basic analytical tools.

Four main categories of GIS analysis functions are:

- Retrieval, Reclassification, and Generalization
- Overlay Techniques
- Neighborhood Operations
- Connectivity Functions

Analysis techniques within these categories are numerous; therefore we will focus on providing an overview of the fundamental primitive functions that are most commonly used in performing spatial analyses.

Retrieval and Reclassification

One of the first GIS analysis operations that a GIS user might undertake is the retrieval and reclassification of data.

Retrieval operations

Often data is selected through an attribute subset and viewed graphically. Retrieval consists of the selective search, manipulation, and output of data, without the need to modify the geographic location of the features involved. These occur on both spatial and attribute data.

Reclassification Operations

This involves the selection and presentation of a selected layer of data based on the classes or values of a specified attribute. It involves searching for an attribute, or an attribute grouping, for a single data layer and classifying that data layer based upon the range of values of the attribute.

Adjacent features which might have a common value (power poles) but differ in other characteristics (pole height or material), will be treated and appear as one class. In raster based GIS, numerical values are often used to indicate the class.

Reclassification is an attribute generalization technique, which typically utilizes polygon patterning techniques such as crosshatching or color shading for graphic representation.

Boundary Dissolving

Boundary dissolving is often performed for visual clarity when creating map compositions. Most GIS software will provide the ability to dissolve boundaries based on reclassification results.

Some systems allow the user to create a new data layer for the reclassification while others just simply dissolve boundaries during data output.

- *For vector data:* boundaries between polygons of common re-classed values should be dissolved to create a cleaner map of homogeneous continuity.
- *For raster data:* The dissolving of map boundaries is based upon a specific attribute value, which often results in a new data layer being created.

Topological Overlay Techniques

The primary analysis technique used in GIS applications, whether vector or raster, is the topological overlay of selected data layers.

Having the ability to overlay multiple data layers vertically is the most vital and commonly deployed technique in geographic data processing, being an initial reason why the topological data structure was adopted as the GIS data structure.

The development of mathematical topology polygon overlay has become the most popular geoprocessing tool, and the basis of any GIS program.

Topological overlay is concerned with the overlaying of one type of polygon data with another, such as a region of plant growth overlain with a region of drought conditions.

There are requirements for overlaying point, linear, and polygon vector data in combinations, such as point in polygon, line in polygon, and polygon on polygon.

Vector and raster based software have a number of differences in their approaches to topological overlay.

- *Raster-based software* - is geared towards arithmetic overlay operations, (addition, subtraction, division, multiplication of data layers). The nature of the one attribute map approach, typical of the raster data model, usually provides a more flexible and efficient overlay capability. The raster data model is very effective in numerically modeling (quantitative analysis). Most sophisticated spatial modeling is performed using raster.
- *Vector-based systems* – With vector systems, the topological overlay is performed by creating a new topological network from two or more existing networks. This requires rebuilding the topological tables, (arc, node, polygon), thus can be time consuming. The result of the overlay is a new topological network, containing attributes of the original input data layers. From this result, selected queries can be performed on the original layer.

Neighborhood Operations

These evaluate the characteristics of an area which surrounds a specific location. Nearly all GIS software will provide some type of neighborhood analysis, with varied ranges of neighborhood functionality.

The analysis of topographic features, such as surficial relief, is normally considered a neighborhood operation. This involves various point interpolation techniques including slope and aspect calculations, contour generation, and Thiessen polygons.

- *Interpolation* - is a means to calculate unknown values that lie between the known values of neighboring locations. This process is often used with point-based elevational data. (This is the estimation and approximation of surficial values at unsampled points derived from known surface values of surrounding points). Interpolation can be used to estimate elevation, rainfall, temperature, chemical dispersion, or other spatially-based phenomena. Interpolation is commonly a raster operation, but it can also be performed on a vector-based TIN surface model. Two interpolation techniques include the "inverse distance weighted" and "kriging" methods.
- *Elevational Data (3-D analysis)* – elevational data is stored in TIN data models, which consists of irregular spaced elevation points. This is the vector data model used for 3-D data analysis (using the z-axis coordinate data). Another model used in storing elevational data is the regular point Digital Elevation Model (DEM). DEM refers to a grid of regularly space elevation points. These points are usually stored with a raster data model.
- *Buffering* - The most common neighborhood function is buffering. This deals with the ability to create distance buffers offset a certain distance from specific features, such as points, lines, or areas. Buffers are

created as polygons as they represent a type of “aura” around a given feature. Buffering is also referred to as *corridor or zone generation* with the raster data model. The results of a buffering process are used routinely in a topological overlay, with another data layer.

Buffering is usually used around point or linear features, based on a set distance from that feature or on a specific attribute of that feature. Some features may have a greater zone of influence due to specific characteristics, such as an Interstate might have a greater influence than a dirt road. Also, different sizes of buffers can be created, based on selected attribute values or feature types.

Connectivity Analysis Functions

Connectivity operations use functions that accumulate attribute values over an area being traversed. Often these include the analysis of surfaces or networks.

Connectivity functions include:

- Proximity analysis
- Network analysis
- Spread functions
- 3-D surface analysis (such as visibility and perspective viewing)

Raster based systems may provide more complex analysis on surficial features, while vector based systems tend to work well with linear feature analysis.

Proximity Analysis

These are techniques used to analyze one feature’s “nearness” or proximity to another.

Proximity is the ability to identify any feature that is near another feature based on:

- Location
- Attribute value
- A specified distance

A simple proximity analysis would be to identify all the power poles that are within 20 feet of a highway, but not necessarily adjacent to it.

Neighborhood buffering is often categorized as being a proximity analysis capability. Depending on the GIS system, the data model used, and the operational functionality of the software it may be hard to distinguish between proximity analysis and buffering.

Adjacency identification

This is another proximity-based form of analytical operation.

Adjacency identification refers to the ability to identify features with particular attributes that exhibit adjacency with other selected features having a particular attribute, for example is the ability to identify all matured pine trees, within 150 feet of a road.

Network analysis

A characteristic of network analysis techniques are their use of feature networks.

Feature networks are comprised almost entirely of linear features. Hydrographic and route networks are good examples of feature networks.

Route optimization

This is a prime example of network analysis techniques used on feature networks.

The analysis is performed to determine the shortest path between connected points or nodes within the feature network based on attribute values.

Attribute values might be as simple as minimal distance, or as complex as defining traffic capacities, and capital improvement costs.

3-D analysis

This type of analysis involves a variety of capabilities. The most commonly used is the generation of perspective surfaces.

Perspective surfaces are usually represented by a wire frame diagram reflecting profiles of the landscape.

These profiles viewed together, with the removal of hidden lines, provide a 3-D perspective.

Other functions which are normally available for 3-D analysis are:

- User-specified vertical exaggeration, viewing azimuth, and elevation angles
- The ID of viewsheds (seen vs unseen areas)
- Feature draping, (point, lines, and shaded polygons onto the 3-D surface)
- Shaded relief model generation
- Cross section profiles generation
- Symbology display on the 3-D surface
- Line of sight perspective views

Chapter 6: Data QC, Output, and Display

Data Quality

Data Standards

The (GIFDS) *Geographic Information Framework Data Standard* establishes common guidelines for data exchange with seven themes of geospatial data that are of critical importance to the National Spatial Data Infrastructure (NSDI). These themes are fundamental to many different GIS applications.

These themes are known as NSDI Framework data themes:

- Cadastral data
- Digital ortho-imagery
- Elevation
- Geodetic control
- Government Units
- Hydrography
- Transportation

Throughout the years, there has been a great bit of concern as to the amount of error that may be inherent in GIS processing methodologies.

Several practical recommendations are identified which help to locate possible error sources, and define the quality of data.

Three distinct components of data quality are:

- Data accuracy
- Data quality
- Data error

Data Source Accuracy

Data accuracy is of fundamental importance in GIS datasets. Accuracy is the closeness of results of observations to the true values or values accepted as being true.

This implies that observations of most spatial phenomena are usually only considered to be estimates of the true value. The difference between observed and true (or accepted as being true) values indicates the accuracy of the observations.

Basically two types of accuracy exist: ***positional*** *and* ***attribute*** *accuracy.*

Positional accuracy

This is the deviation in the geographic location of an object from its true ground position, which is what we commonly think of when discussing accuracy.

The two components to positional accuracy are relative and absolute accuracy:

- *Absolute accuracy* - concerns the accuracy of data elements with respect to a coordinate scheme, such as the UTM.
- *Relative accuracy* - concerns the positioning of map features relative to one another. Often relative accuracy is of greater concern than absolute accuracy. For example, the fact that a particular survey's coordinates do not precisely coincide is of less importance than the absence of a parcel from a tax map.

Attribute accuracy

This is equally importance to positional accuracy. Errors within tabulated data can be very problematic and costly.

Data Source Level of Usefulness

Quality can simply be defined as the level of usefulness for a specific data set. Data that is allowable for one application may not be for another.

It is dependent upon parameters such as the scale, accuracy, and extent of the data set, as well as the quality of other data sets to be used.

Data Quality Parameters

Five data quality parameters to be aware of:

- *Lineage* – The history and the means in which data was compiled. The source and content of the data, data capture specs, geographic coverage of the data, compilation method (digitized or scanned), transformation methods applied to the

data, the use of an pertinent algorithms when compiled (linear simplification and feature generalization operations).
- *Positional Accuracy* – Is the positional accuracy maintained, including consideration of inherent error (source error) and operational error (introduced error by the operator).
- *Attribute Accuracy* – Is the attributtal data accurate; this can be a difficult quality to verify. This quality component deals with the verification of the reliability of the textual data within the dataset.
- *Logical Consistency* - Is the data structure maintained for a data set (connectivity and adjacency issues); such as spatial data inconsistencies, non-connecting line intersections (nodes), duplicated lines or boundaries, or gaps in lines, and other spatial or topological errors.
- *Incomplete or Corrupted Data* – Are there gaps in the data, missing areas, and any compilation procedures that may have caused data to have been corrupted or eliminated.

Data Source Errors

Two sources of errors, inherent and operational, reduce the quality of the output generated by GIS.

- **Inherent errors** – These are already present in source documents and data.
- **Operational errors** – These are created through the data capture and manipulation procedures usually by human error.

Six possible sources of operational errors include:

- mislabeled areas on thematic maps
- missing and misaligned horizontal (positional) boundaries
- human error in digitizing; poor digitizing skills and habits
- human bias
- classification errors
- GIS algorithm inaccuracies

As with any production process, quality control and quality assurance in data handling procedures help to maintain a consistence in data quality.

Map Composition

Mapping Standards

With the extent to which GIS is used for critical decision-making, there is a need for high quality cartographic output, with the importance of compiling maps using standard cartographic principles ever-increasing.

Most GIS platforms achieve map composition through the use of user-defined map templates, to generate plotted scales independent of the data set being utilized.

Map composition capabilities typically include the capability to define viewports or windows into the data set.

Often viewports are defined by a map's required output scale. These viewports, a standardized scaling feature popularized by Autodesk CAD software in the 80's, can be used for a variety of graphical displays and perspective views.

The primary map composition requirements include defining:

- color
- feature symbology for points and lines
- line width, line type, line pattern, point symbols
- annotation (text labeling and descriptions)
- map scale bar and legend
- polygon shading symbology (cross-hatching or color fill)

Tabular Reports

Tabular and DB Reports

Another form of output that's commonly performed is the tabular or database report. Output maps usually require some statistics to accompany the results of an analysis. Tabular reports are database listings of the results of a GIS analysis which coincides with the graphic display.

A GIS operator or analyst might generate a report based on a defined analytical process.

Statistical analysis functions which might be performed on geographical features are:

- Means
- Modes
- Averages
- Totals
- Regression analysis
- Correlation analysis
- etc.

Often, the user must export data from the GIS into another software package, such as those programs developed by SAS, in order to properly run a statistical analysis.

GIS is a specialized tool and depending on the software platform, may lack many of the standard information processing options found in non-spatially oriented systems.

Sources of Data

Data Input

As previously identified, two types of data are input into a GIS, spatial and attribute. The data input process is the operation of encoding both types of data into the GIS database formats.

The creation of a clean digital database is the most important and time consuming task upon which the usefulness of the GIS depends. The establishment of a robust spatial database is the cornerstone of a successful GIS implementation.

As well, the digital data is the most expensive part of the GIS. Yet, often, not enough attention is given to the quality of the data or the processes by which they are prepared for automation. T

he general agreement within the GIS community is that 60 to 80 % of the project costs during the implementation stage of GIS technology lie in data acquisition, data compilation and database development. A wide variety of data sources exist for both spatial and attribute data.

The most common general sources for spatial data are:

- hard copy maps
- aerial photographs
- remotely-sensed imagery
- point data samples from surveys
- existing digital data files

Graphic Output - Plotters

Large format inkjet printers (plotters) are the primary means of hardcopy output for GIS maps. Other outdated plotters use pens, and electrostatic for plotting.

While maps are the most common graphic output product, graphical reports can be generated also, including: histograms, scattergrams, profiles, and charts.