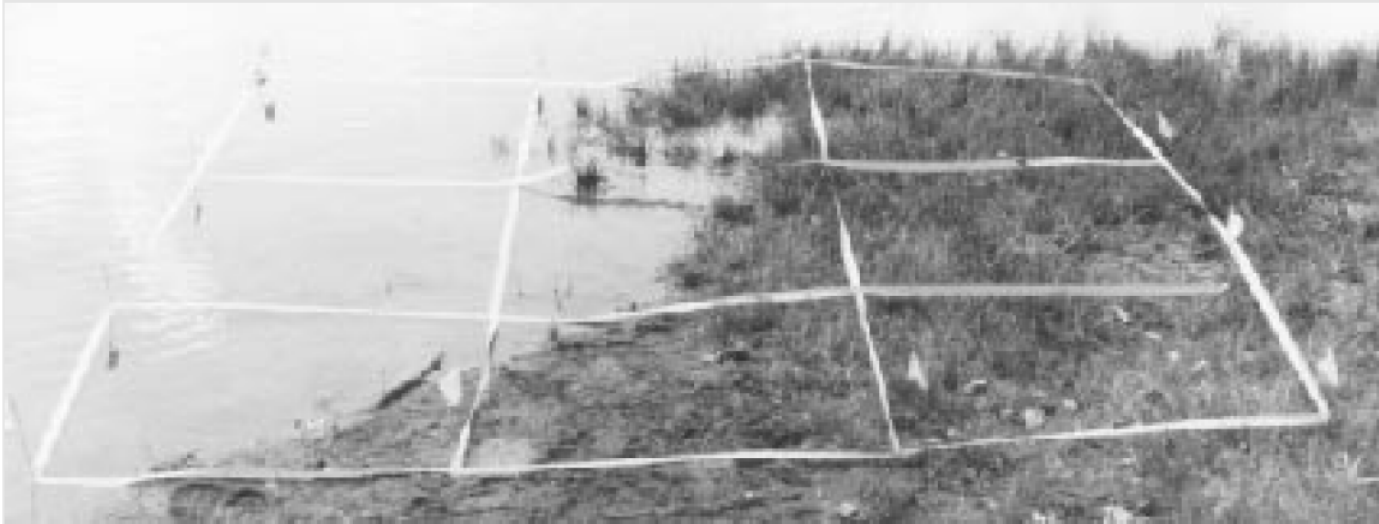


# CHAPTER 5

## THE NATURE AND SOURCE OF GEOGRAPHIC DATA



# SPATIAL DATA FORMATS – RASTER AND VECTOR

## ❖ Raster Data Format

- Raster data represents a graphic object as a pattern of dots, whereas vector data represents the object as a set of lines drawn between specific points.

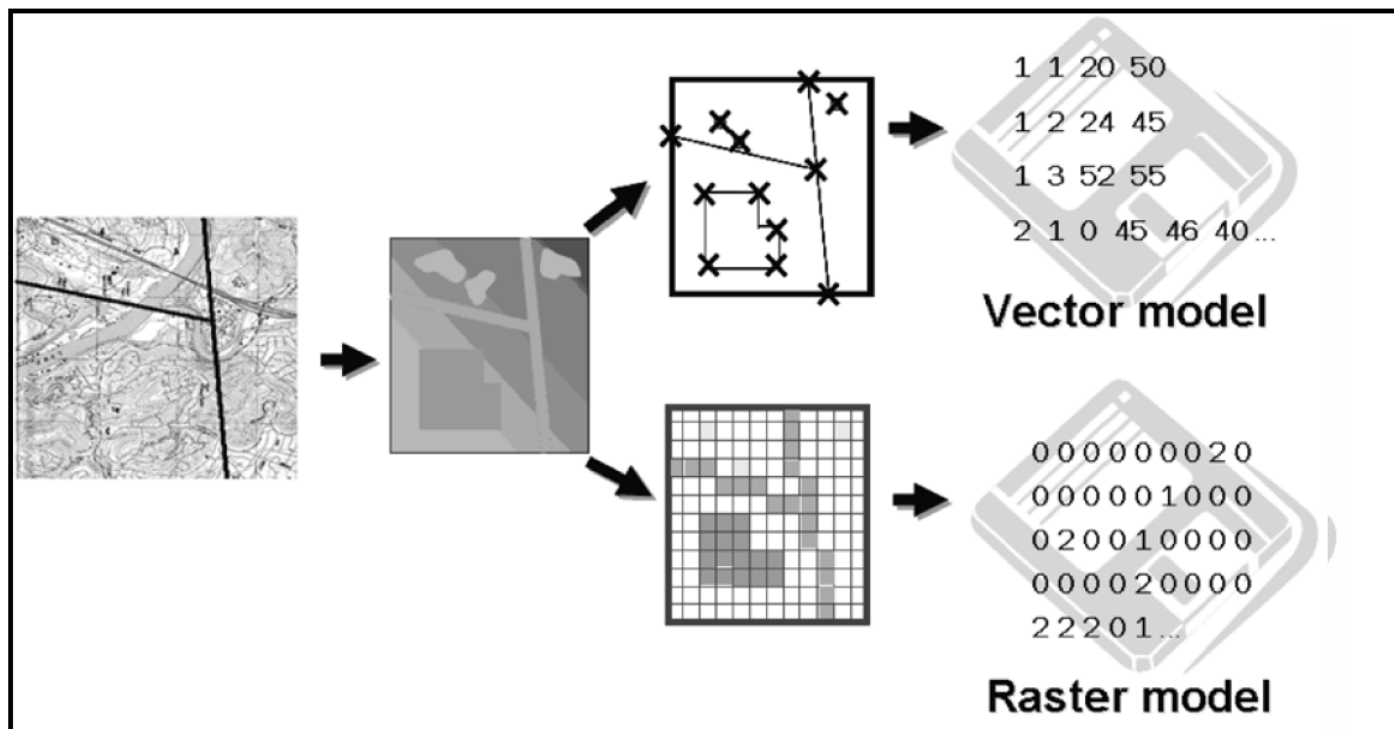


Figure 5.1: Modelling the real world.

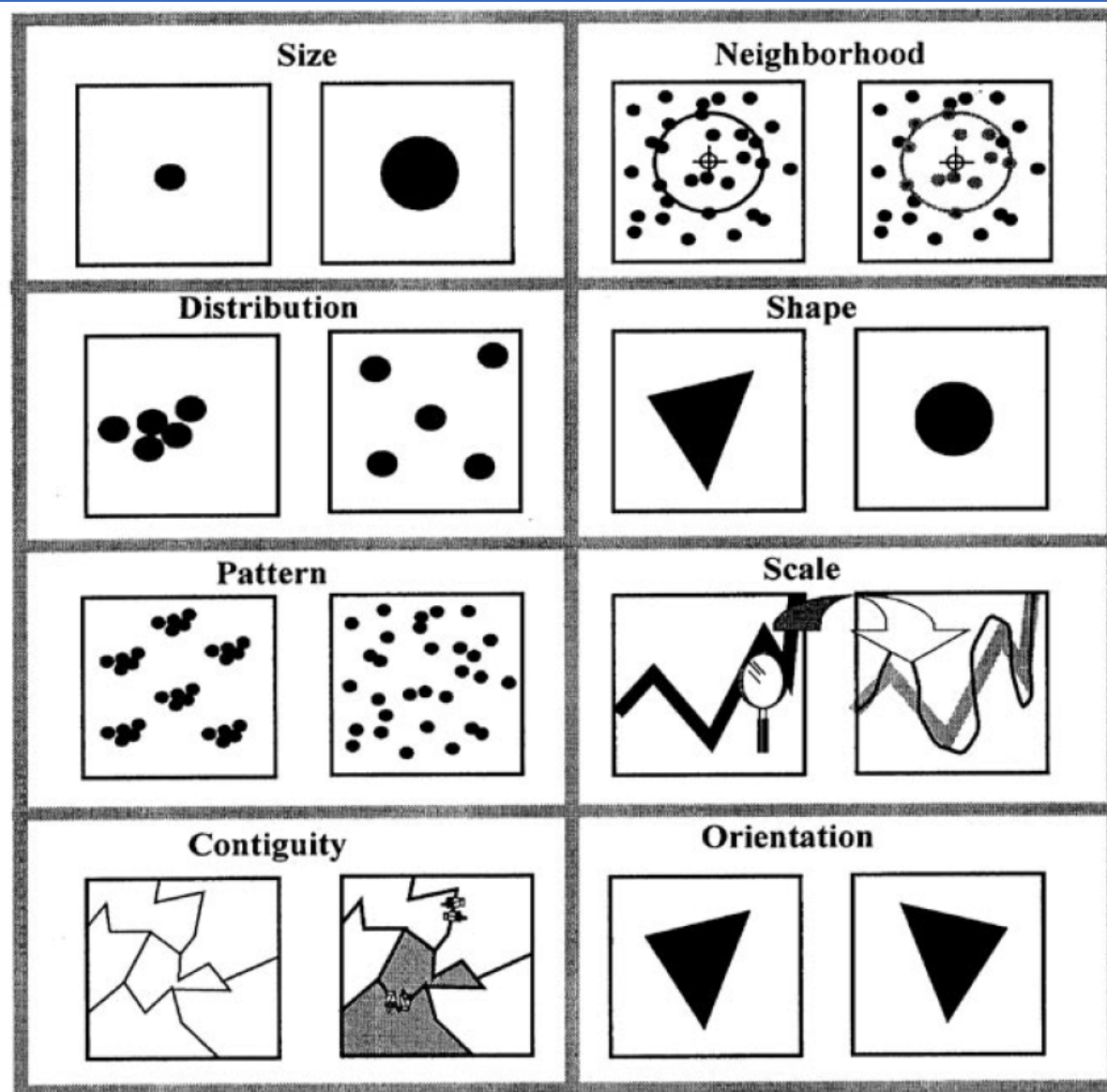


Figure 5.2: Basic properties of geographic features.

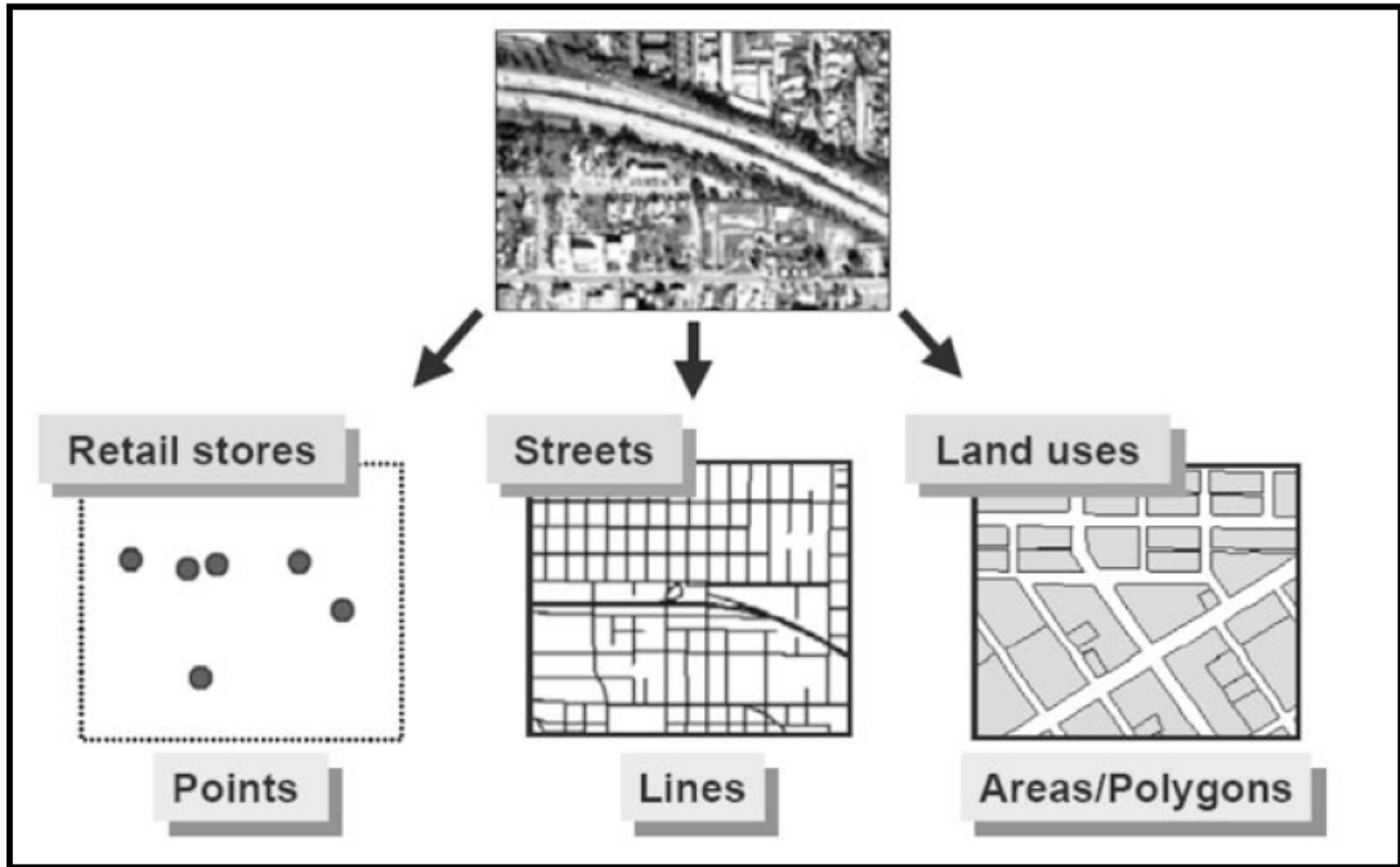


Figure 5.3: Representation of geographic details, point, line and area features.

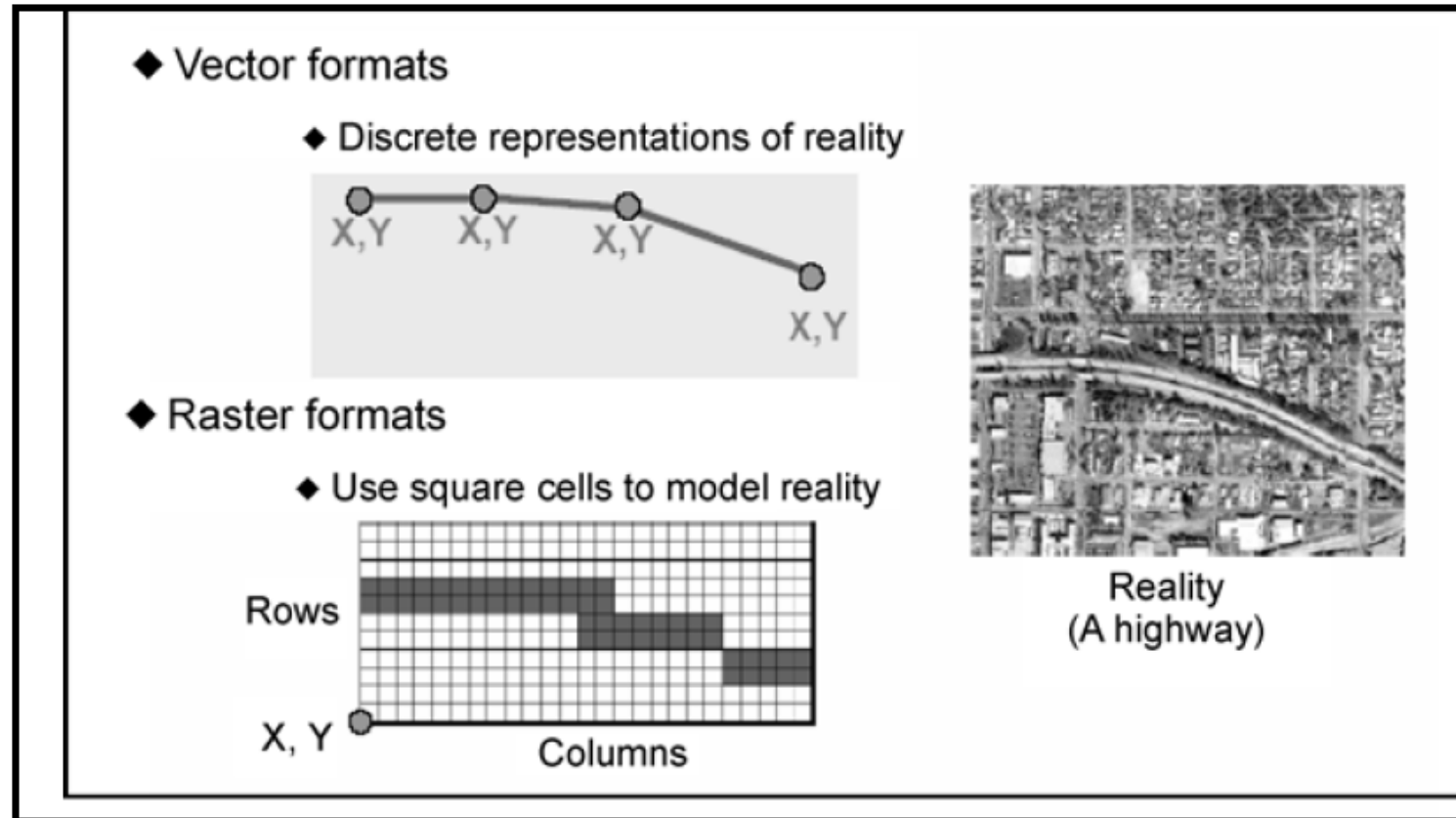


Figure 5.4: Storing of spatial data, vector and raster data formats.

- Raster files are most often used:
  - For digital representations of aerial photographs, satellite images, scanned paper maps, and other applications with very detailed images.
  - When costs need to be kept down.
  - When the map does not require analysis of individual map features.
  - When 'backdrop' maps are required.

- The relationship between cell size and the number of cells is expressed as the RESOLUTION of the raster.
  - A finer RESOLUTION gives a more accurate and better quality image.

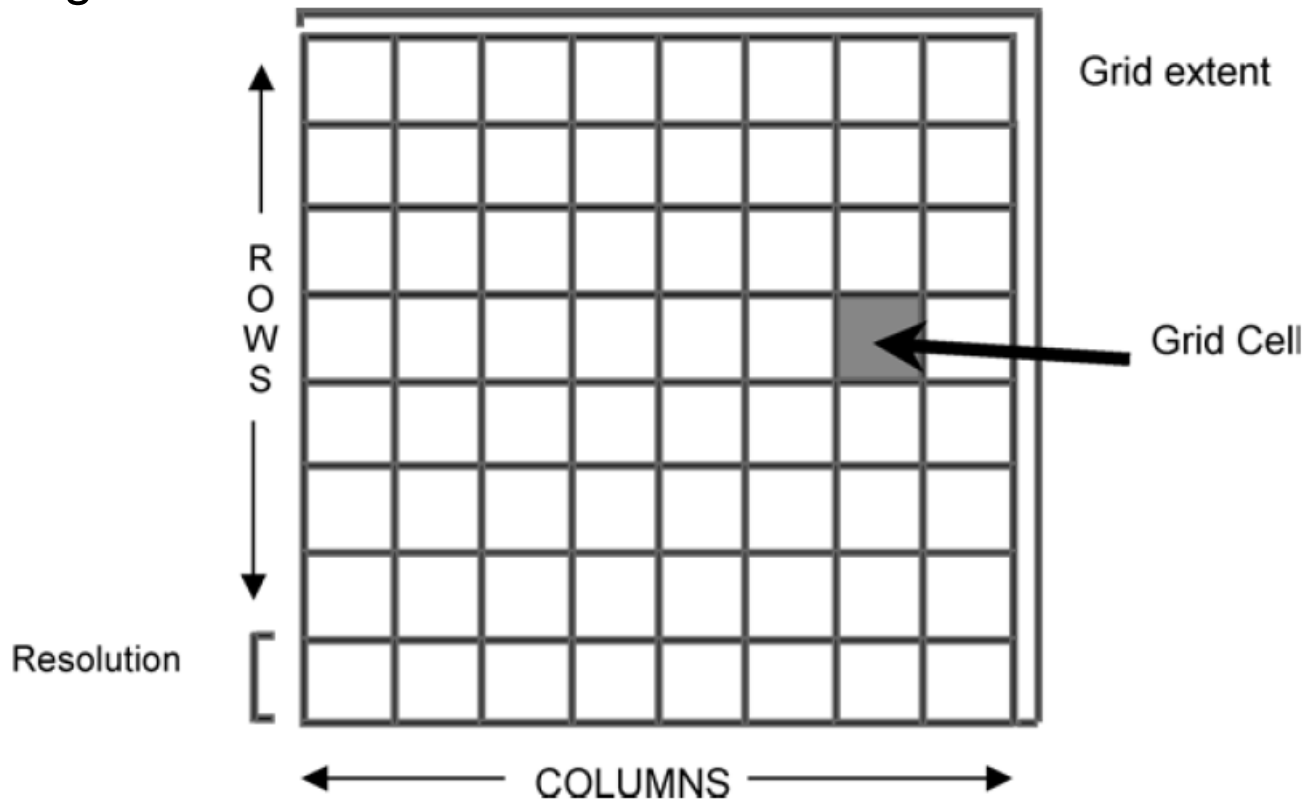
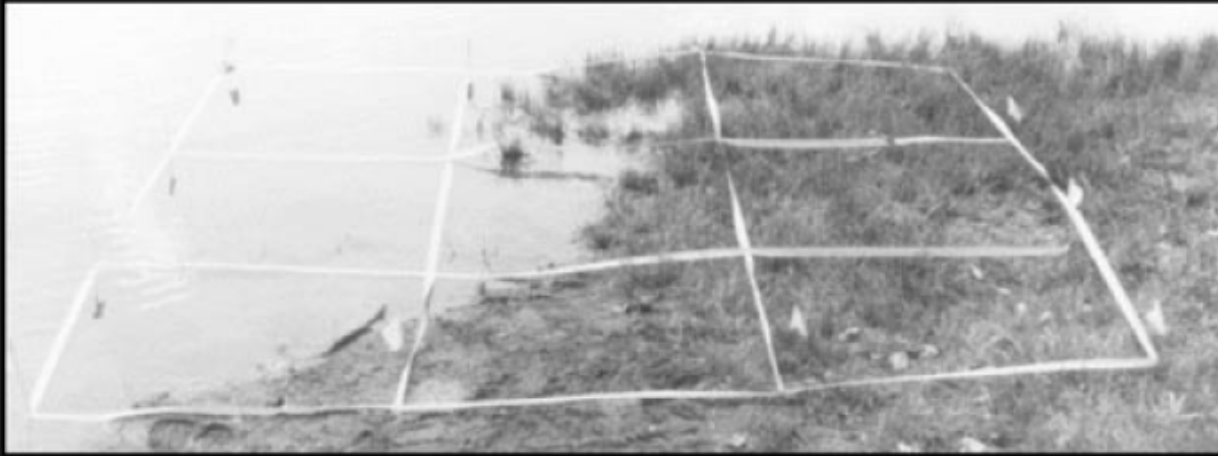


Figure 5.5: Generic structure for a grid.



**Water dominates**

W	W	G
W	W	G
W	W	G

**Winner takes all**

W	G	G
W	W	G
W	G	G

**Edges separate**

W	E	G
W	E	G
E	E	G

Figure 5.6: The mixed pixel problem.



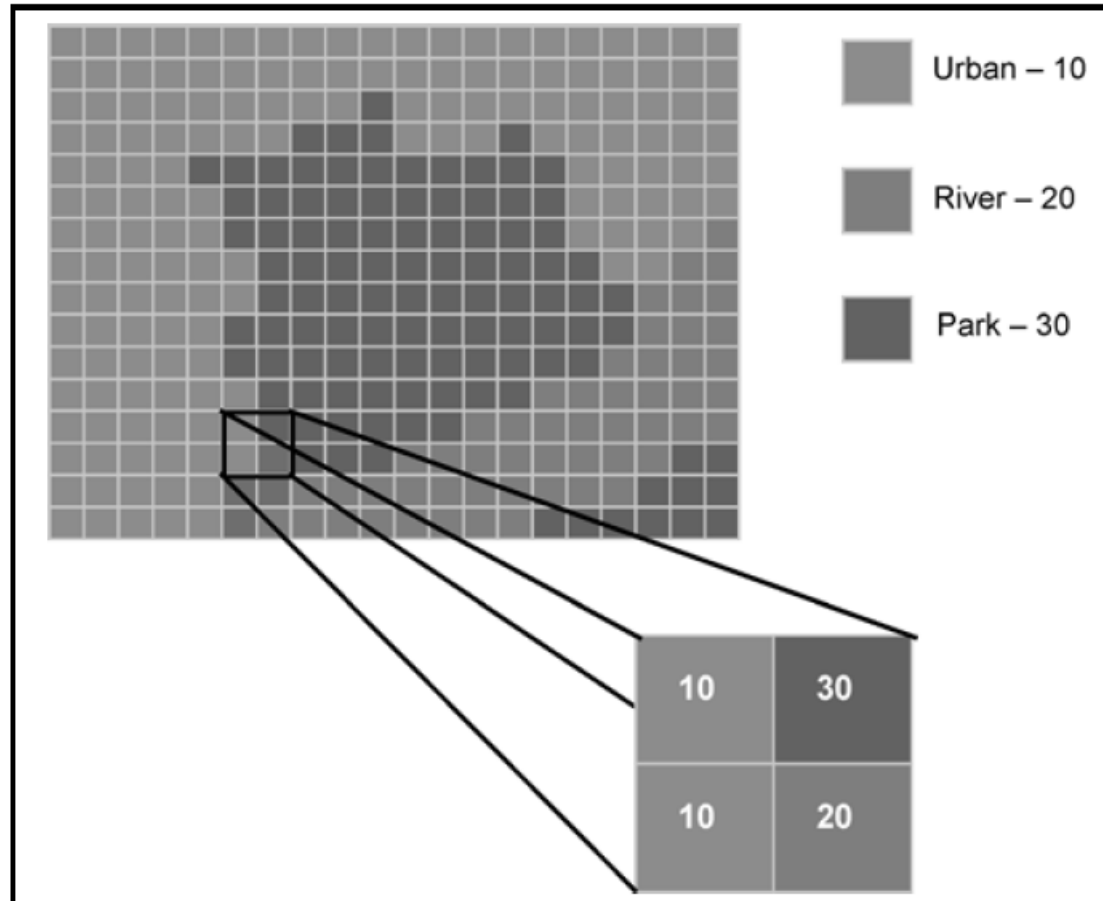


Figure 5.7: Attribute handling in raster data. Each pixel is assigned a single value which represents a real world object. Pixels can only hold numeric data; each pixel value in the raster here represents a feature class.

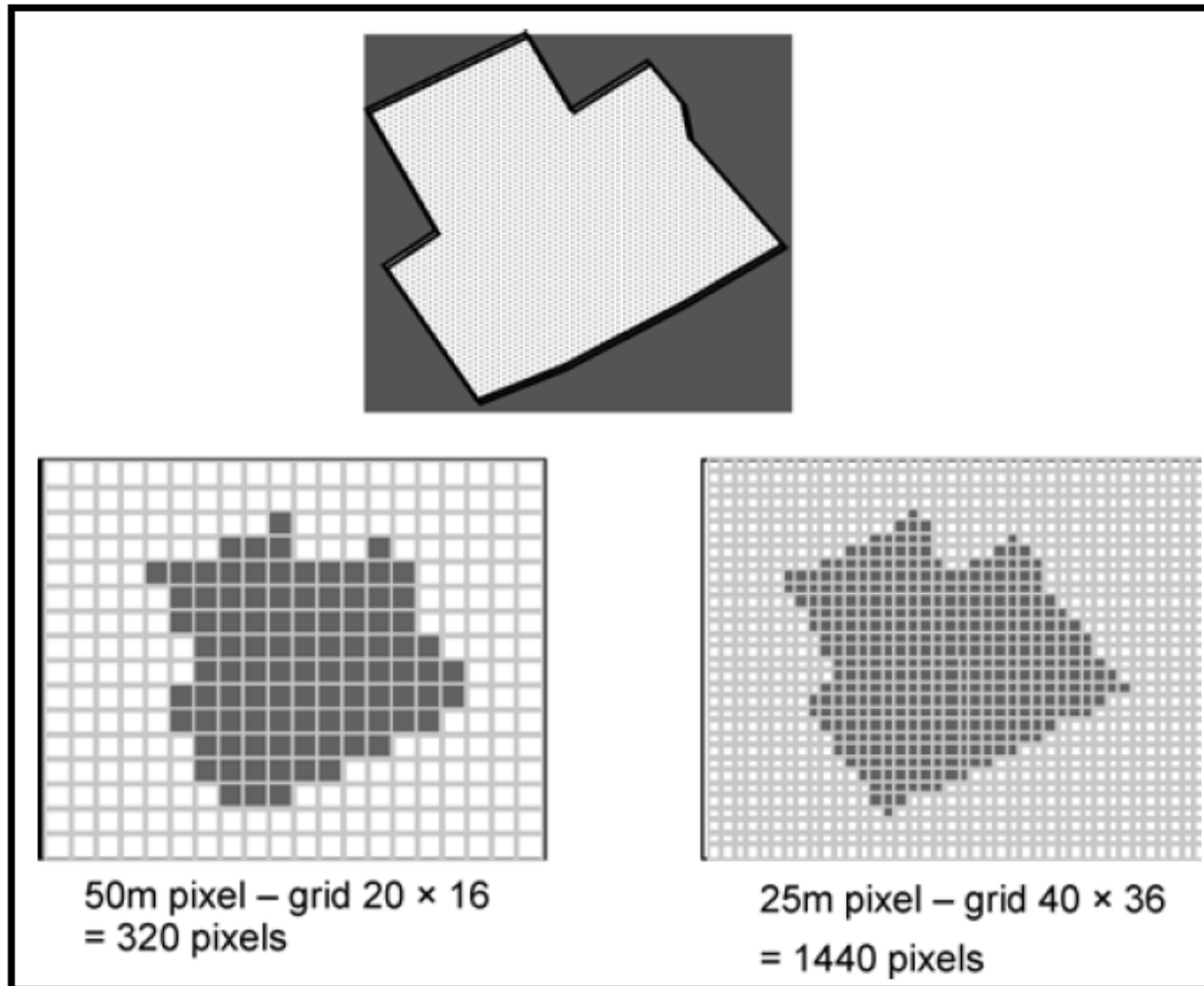


Figure 5.8: Effect of grid size on data in raster format.

## ❖ Vector Data Format

- A vector data model uses points stored by their real (earth) coordinates.
  - Lines and areas are built from sequences of points in order.
  - Lines have a direction to the ordering of the points.
  - Polygons can be built from points or lines.

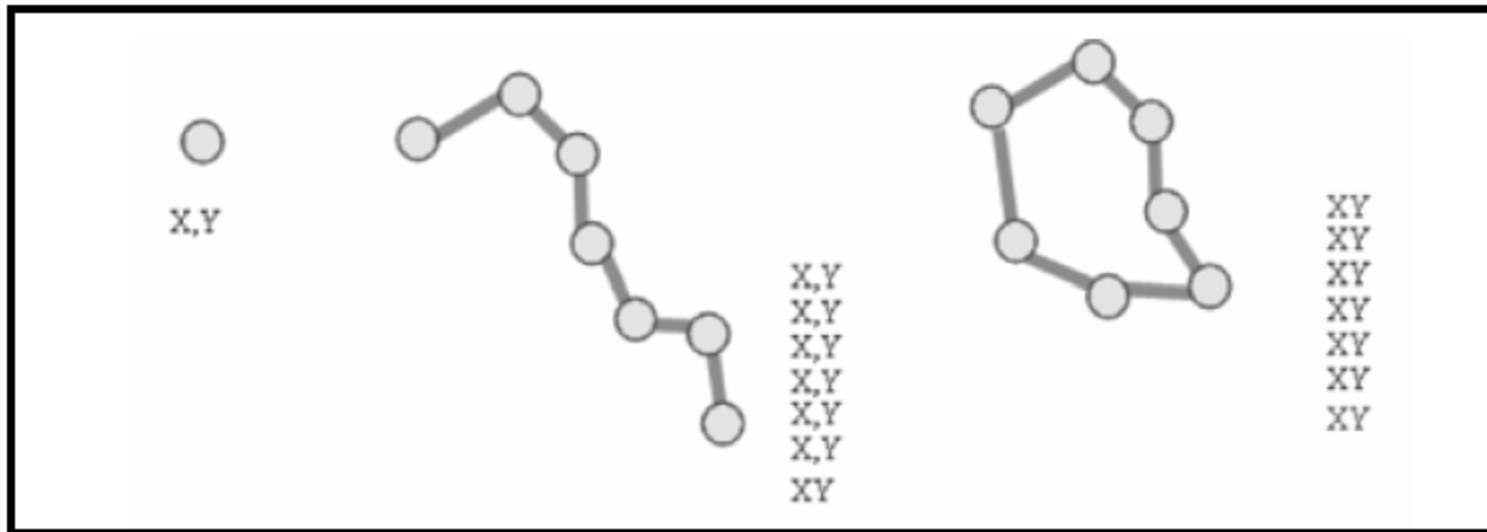


Figure 5.9: The vector data model is based around the storage of coordinate pairs.

- Vectors can store information about topology.
- Vector files are most often used:
  - Highly precise applications.
  - When file sizes are important.
  - When individual map features require analysis.
  - When descriptive information must be stored.

Raster Model	Vector Model
<p><b>Advantages</b></p> <ul style="list-style-type: none"> <li>• Simple data structure</li> <li>• Easy and efficient overlaying</li> <li>• Compatible with RS imagery</li> <li>• High spatial variability is efficiently represented</li> <li>• Simple for own programming</li> <li>• Same grid cells for several attributes</li> </ul> <p><b>Disadvantages</b></p> <ul style="list-style-type: none"> <li>• Inefficient use of computer storage</li> <li>• Errors in perimeter, and shape</li> <li>• Difficult network analysis</li> <li>• Inefficient projection transformations</li> <li>• Loss of information when using large cells Less accurate (although interactive) maps</li> </ul>	<p><b>Advantages</b></p> <ul style="list-style-type: none"> <li>• Compact data structure</li> <li>• Efficient for network analysis</li> <li>• Efficient projection transformation</li> <li>• Accurate map output</li> </ul> <p><b>Disadvantages</b></p> <ul style="list-style-type: none"> <li>• Complex data structure</li> <li>• Difficult overlay operations</li> <li>• High spatial variability is inefficiently represented</li> <li>• Not compatible with RS imagery</li> </ul>

## ❖ Vector data model vs. raster data model

- Vector data model
  - The method of representing geographic features by the basic graphical elements of points, lines and polygon is said to be the *vector method* or *vector data model* and the data are called *vector data*.
  - Related vector data are always organized by themes, which are also referred to as *layers* or *coverages*.
    - Examples of themes: geodetic control, base map, soil, vegetation cover, land use, transportation, drainage and hydrology, political boundaries, land parcel and others.
  - For themes covering a very large geographic area, the data are always divided into *tiles* so that they can be managed more easily.

- A collection of themes of vector data covering the same geographic area and serving the common needs of a multitude of users constitutes the *spatial component of a geographical database*.
- The vector method of representing geographic features is based on the concept that these features can be identified as discrete entities or objects. This method is therefore based on the *object view of the real world* (Goodchild, 1992).

- Raster data model
  - The method of representing geographic features by pixels is called the *raster method* or *raster data model*, and the data are described as *raster data*. The raster method is also called the *tessellation method*.
  - Raster data are organized by themes, which is also referred to as layers.
    - Examples of themes: bed rock geology, vegetation cover, land use, topography, hydrology, rainfall, temperature etc.
  - Raster data covering a large geographic area are organized by *scenes* (for remote sensing images) or by *raster data files* (for images obtained by map scanning).
  - The raster method is based on the concept that geographic features are represented as surfaces, regions or segments. This method is therefore based on the *field view of the real world*.



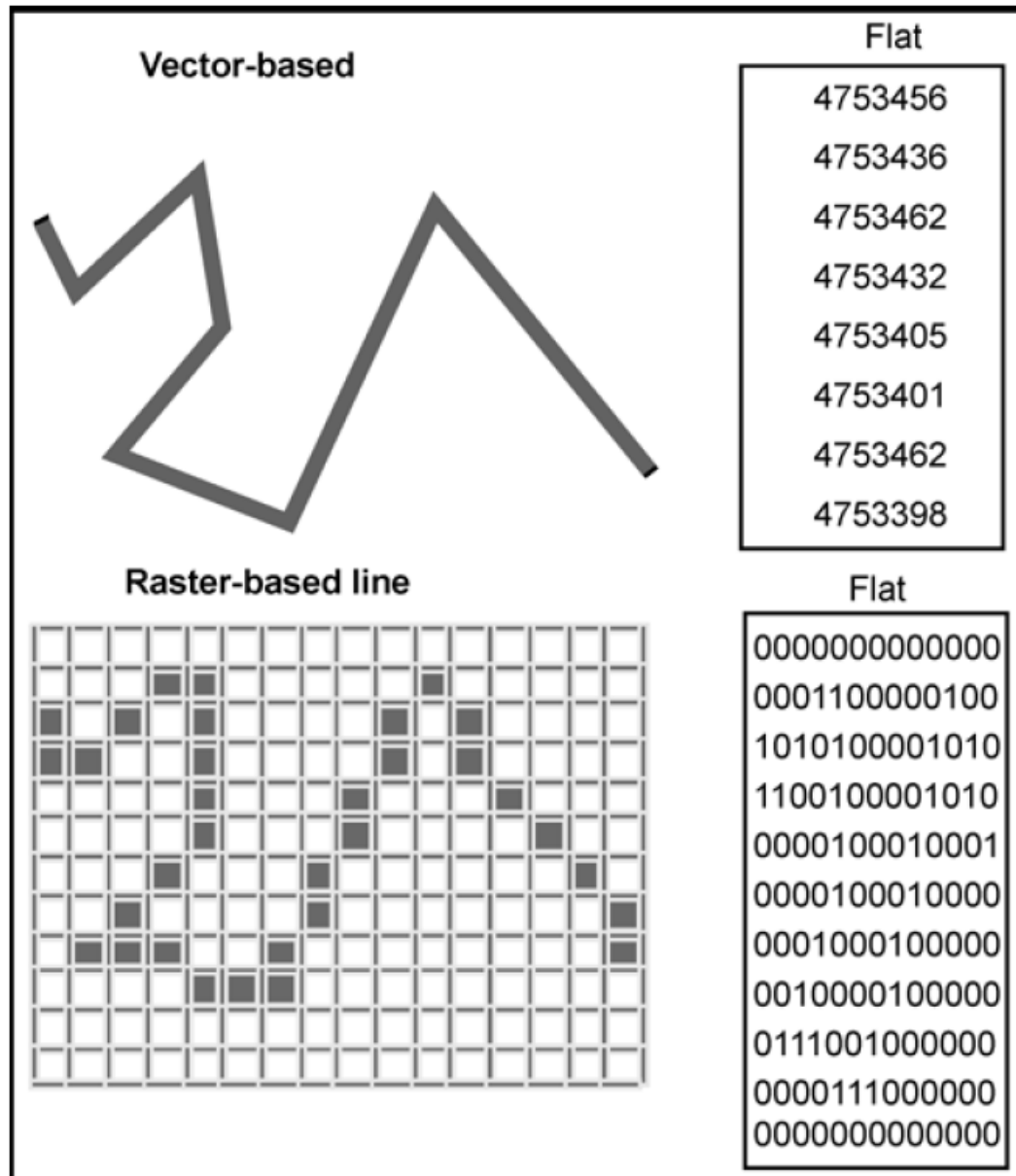
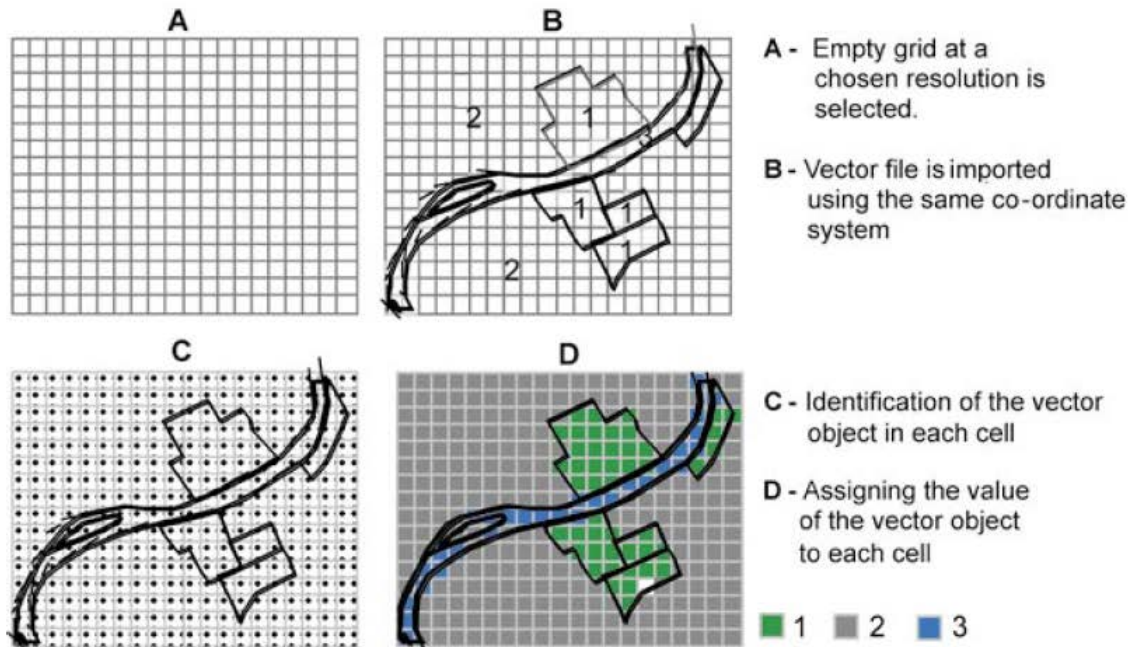
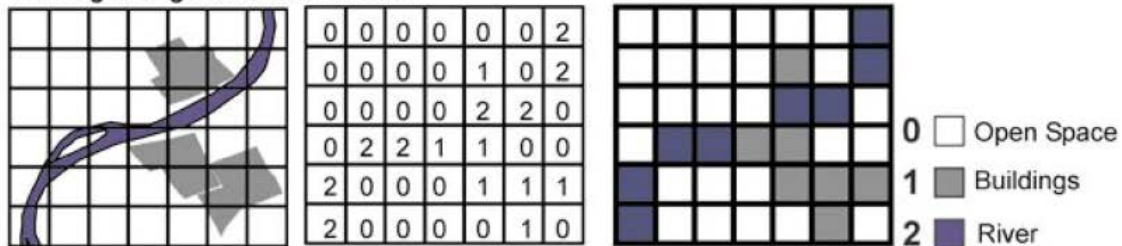


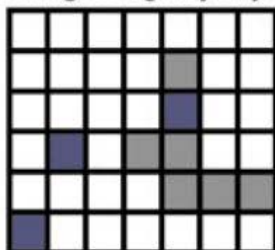
Figure 5.10: Rasters and vectors can be flat files...if they are simple.



**Coding Using Feature at Centre**



**Coding Using Majority Area**



**Coding Using Priority Rule**

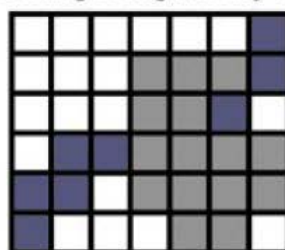


Figure 5.11: Raster data capture-rasterisation.

# CHOICE BETWEEN RASTER AND VECTOR

- An important current trend involves linking raster and vector systems, displaying vector data overlying a raster base.
- The question has evolved from ‘Which is best?’ to ‘Under what conditions is which best and how can we have flexibility to use the most appropriate approaches on a case by case basis?’
- Four issues to the discussions of raster versus vector:
  - coordinate precision
  - speed of analytical processing
  - mass storage requirements
  - characteristics of phenomena

Box 8: *Choice between raster and vector data*

	<b>Raster</b>	<b>Vector</b>
Data Collection	Rapid	Slow
Data Volume	Large	Small
Data Structure	Simple	Complex
Geometrical Accuracy	Low	High
Graphic Treatment	Average	Good
Area Analysis	Good	Average
Network Analysis	Poor	Good
Generalization	Simple	Complex

# DATA CAPTURE

- GIS can contain a wide variety of geographic data types originating from many diverse sources.
- From the perspective of creating geographic databases, it is convenient to classify raster and vector geographic data as primary and secondary (Box 10).

Box 10: *General classification of geographic data.*

Source	Raster	Vector
Primary	<ul style="list-style-type: none"> <li>• Digital aerial photographs</li> <li>• Digital remote sensing images</li> </ul>	<ul style="list-style-type: none"> <li>• Survey measurements</li> <li>• GPS measurements</li> </ul>
Secondary	<ul style="list-style-type: none"> <li>✦ Scanned maps</li> <li>✦ Photographs</li> <li>✦ DEM generated from maps</li> </ul>	<ul style="list-style-type: none"> <li>✦ Topographic maps</li> <li>✦ Toponymy databases (Place names)</li> </ul>

- Primary data sources
  - They are those collected specifically for use in GIS.
  - Typical primary GIS sources include raster IRS, SPOT and IKONOS Earth satellite images, and vector building survey measurements captured using a total survey station.
- Secondary sources
  - They are those that were originally captured for another purpose and need to be converted into a form suitable for use in a GIS project.
  - Typical secondary sources include raster scanned colour aerial photographs of urban areas, and USGS and IGN paper maps that can be scanned and vectorized.

- Geographic data may be obtained in either digital or analog format.
  - Analog data must always be digitized before being added to a geographic database.
  - Depending on the format and characteristics of the digital data, considerable reformatting and restructuring may be required prior to import.
- The processes of data collection are also variously referred to as data capture, data automation, data conversion, data transfer, data translation, and digitizing.
  - They essentially describe the same thing, *i.e.*, adding geographic data to a database.

Box 9: Possible encoding methods for different data sources.

Data source	Analogue or Digital source	Possible encoding method	Examples
Tabular data	Analogue	<ul style="list-style-type: none"> <li>• Keyboard</li> <li>• Text scanning</li> </ul>	<ul style="list-style-type: none"> <li>• List of school</li> <li>• Education board publications</li> </ul>
Map data	Analogue	<ul style="list-style-type: none"> <li>• Digitizing</li> <li>• Scanning</li> </ul>	<ul style="list-style-type: none"> <li>• Political maps</li> <li>• Historical maps</li> </ul>
Aerial photo	Analogue	<ul style="list-style-type: none"> <li>• Digitizing</li> <li>• Scanning</li> </ul>	<ul style="list-style-type: none"> <li>• Landuse maps</li> <li>• Water bodies</li> </ul>
Tabular data	Digital	<ul style="list-style-type: none"> <li>• Digital file transfer</li> </ul>	<ul style="list-style-type: none"> <li>• Census data</li> </ul>
Satellite image	Digital	<ul style="list-style-type: none"> <li>• Digital file transfer</li> </ul>	<ul style="list-style-type: none"> <li>• Landuse data</li> </ul>



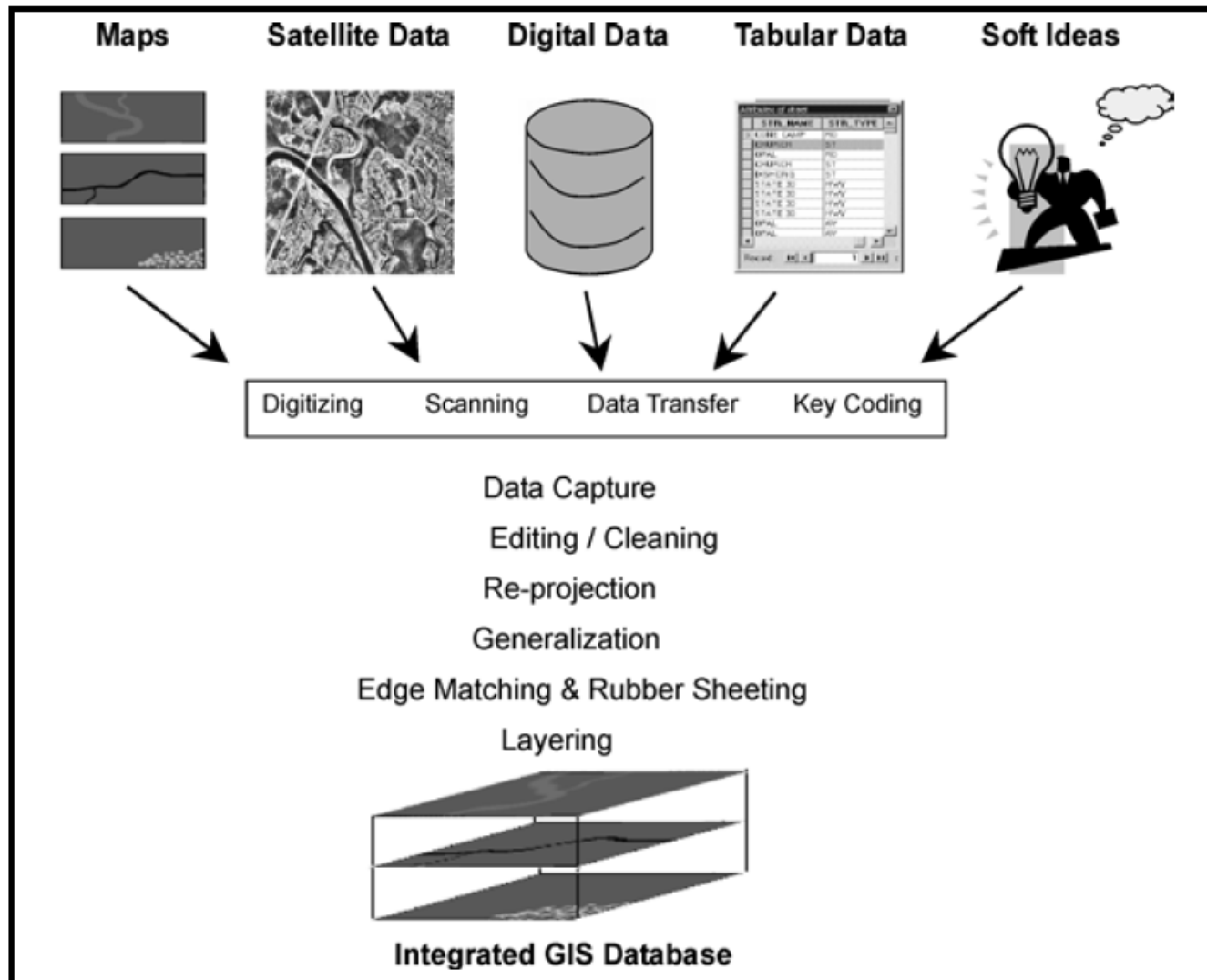


Figure 5.12: GIS data stream.

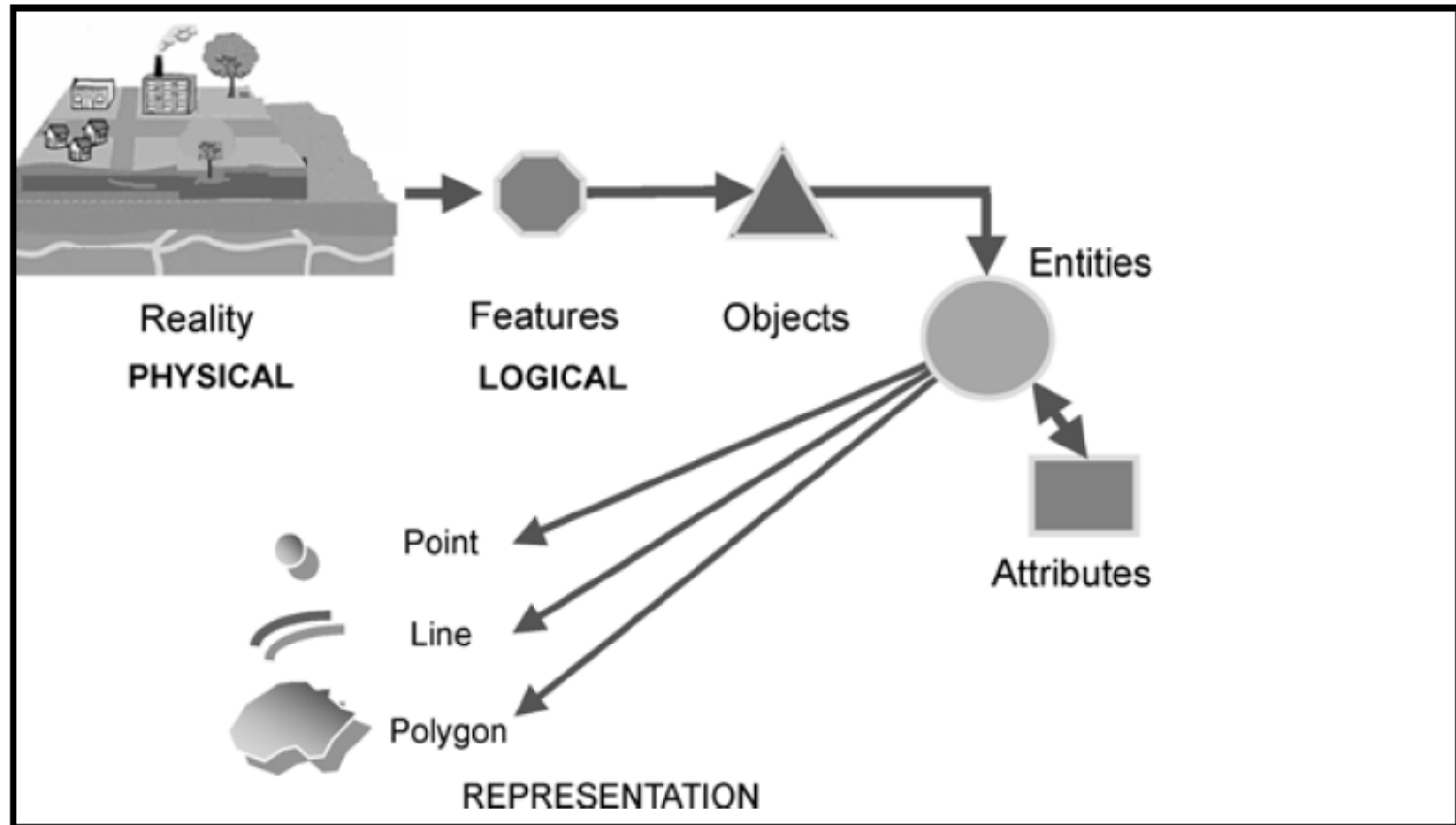


Figure 5.13: Modelling the real world data into GIS.

- In the early days of GIS, data collection was the main project task and it typically consumed the majority of the available resources.
- Data collection still remains a time consuming, tedious, and expensive process.
  - Usually it accounts for 15 – 50% of the total cost of a GIS project (Longley, et al., 2001).
- After an organization has completed basic data collection, their emphasis moves on to data maintenance.
- Data maintenance often turns out to be a far more complex and expensive activity than initial data collection.
  - This is because of the high volume of update transactions in many systems and the need to manage multi-user access to operational databases.

# DATA COLLECTION WORKFLOW

- Data collection projects involve a series of sequential stages (Figure 5.14).
  - The workflow commences with planning, followed by preparation, digitizing or transfer, editing and improvement and, finally, evaluation.

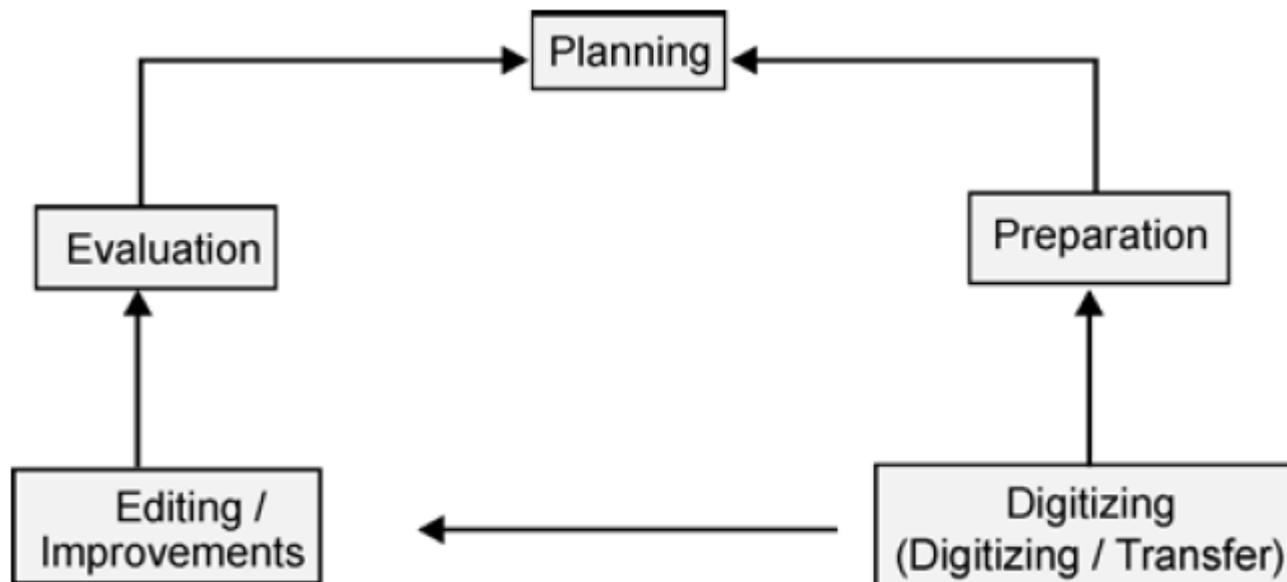


Figure 5.14: Stages in data collection.

- Planning
  - Planning is obviously important to any project and data collection is no exception.
  - It includes establishing user requirements, garnering resources (staff, hardware, and software) and developing a project plan.
- Preparation
  - Preparation is especially important in data collection projects.
  - It involves many tasks such as obtaining data, redrafting poor-quality map sources, editing scanned map images, and removing noise.
    - Noise: unwanted data such as speckles on a scanned map image
- Digitizing and transfer
  - Digitizing and transfer are the stages where the majority of the effort will be expended.

- Editing and improvement
  - Editing and improvement follows digitizing / transfer.
  - This covers many techniques designed to validate data, as well as correcting errors and improving quality.
- Evaluation
  - Evaluation is the process of identifying project successes and failures.
- Since all large data projects involve multiple stages, this workflow is iterative with earlier phases (especially a first, pilot, phase) helping to improve subsequent parts of the overall project.

# PRIMARY GEOGRAPHIC DATA CAPTURE

- Primary geographic capture involves the direct measurement of objects.

## ❖ Raster data capture

- The most popular form of primary raster data capture is remote sensing.
- Remote sensing is a technique used to derive information about the physical, chemical, and biological properties of objects without direct physical contact.
- Information is derived from measurements of the amount of electromagnetic radiation reflected, emitted, or scattered from objects.
- As used here, the term remote sensing subsumes the fields of satellite remote sensing and aerial photography.

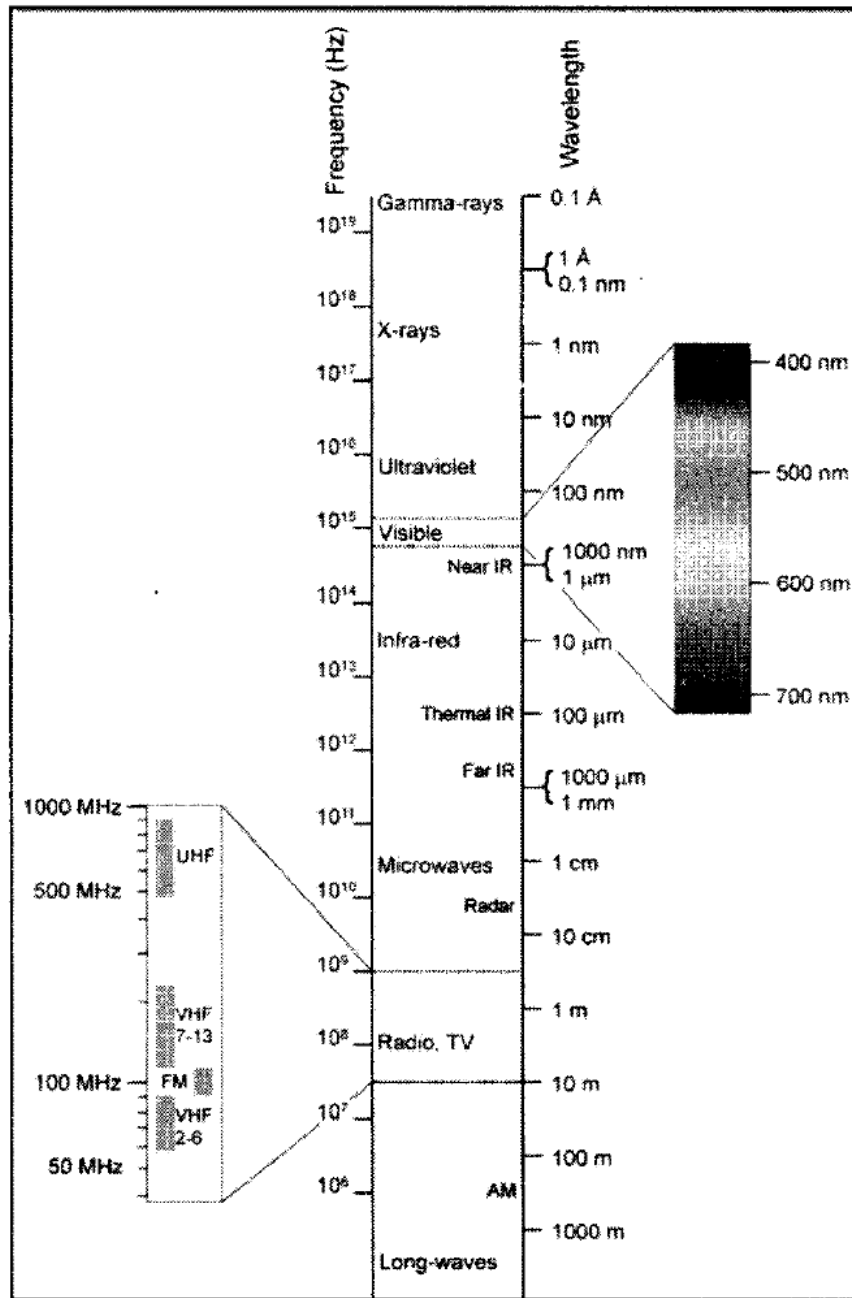


Figure. 5.15: Electromagnetic spectrum.



- From the GIS perspective, resolution is the key physical characteristic of remote sensing systems.
  - three basic aspects to resolution: spatial, spectral, and temporal.
  - Spatial resolution refers to the size of object that can be resolved and the most usual measure is the pixel size.
  - Spectral resolution refers to the parts of the electromagnetic spectrum that are measured.
  - Temporal resolution, or repeat cycle, describes the frequency with which images are collected for the same area.

- Aerial photographs
  - They are normally collected using analog optical cameras and then later rasterized, usually by scanning a film negative.
  - The quality of the optics of the camera and the mechanics of the scanning process both affect the spatial and spectral characteristics of the resulting images.
  - Aerial photographs are very suitable for detailed surveying and mapping projects.
- Satellite and aerial photography systems can provide stereo imagery from overlapping pairs of images.
  - These images are used to create a 3D model from which 3D coordinates, contours and digital elevation models can be created.

- Advantages of satellite and aerial photograph data
  - The consistency of the data and the availability of systematic global coverage make satellite data useful for large area projects and for mapping inaccessible areas.
  - The regular repeat cycles and the fact that they record radiation in many parts of the spectrum makes such data especially suitable for assessing the condition of vegetation.
  - Aerial photographs are very useful for detailed surveying and mapping of urban areas and those applications requiring 3D data.

- Drawbacks of satellite and aerial photograph data
  - The spatial resolution of commercial satellites is too coarse for many large area projects
  - The data collection capability of many sensors is restricted by cloud cover.
  - The data volumes from both satellites and aerial cameras can be very large and create storage and processing problems.
  - The cost of data can also be prohibitive for a single project or organization.

# VECTOR DATA CAPTURE

- The two main branches of vector data capture are ground surveying and GPS.

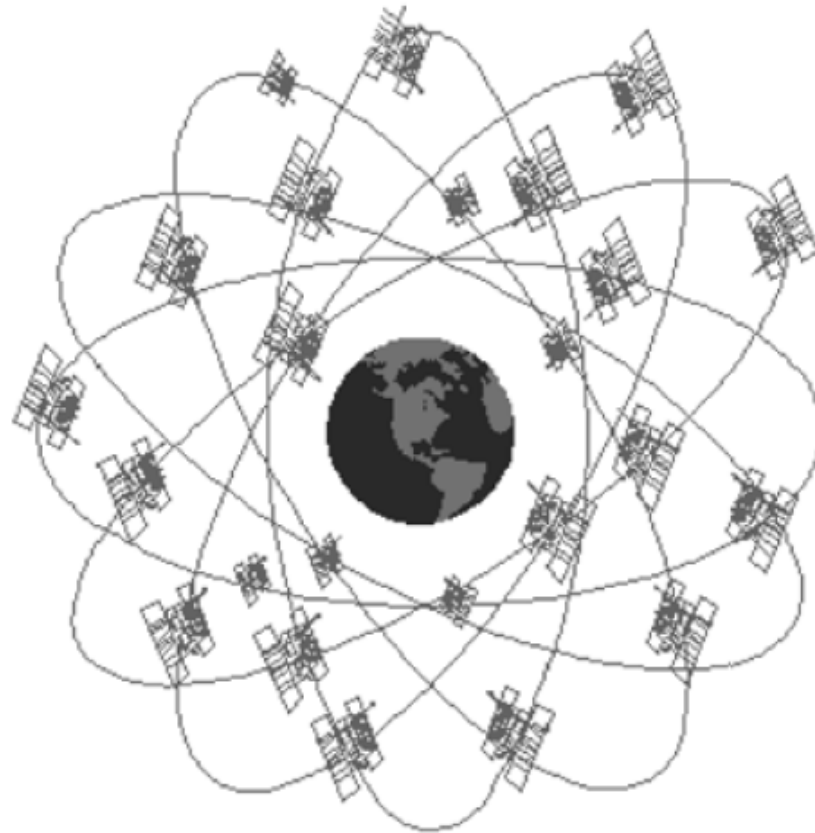
## ❖ Surveying

- Ground surveying is based on the principle that the 3D location of any point can be determined by measuring angles and distances from other known points.
  - Surveys begin from a benchmark point.
  - Since all survey points are obtained from survey measurements their locations are always relative to other points.
  - Any measurement errors need to be apportioned between multiple points in a survey.
- Total stations can measure both angles and distances to an accuracy of 1 millimeter; automatically log data and the most sophisticated can create vector point, line, and polygon objects in the field, thus providing direct validation.

- Ground survey is a very time-consuming and expensive activity, but it is still the best way to obtain highly accurate point locational data.
- Surveying is typically used for capturing buildings, land and property boundaries, and other objects that need to be located accurately.
- It is also used to obtain reference marks for other data capture methods; for example, large-scale aerial photographs and satellite images are frequently georeferenced using points obtained from ground survey.

## ❖ GPS

- The Global Position System (GPS) is a collection of 27 NAVSTAR satellites orbiting the Earth at a height of 12,500 miles, five monitoring stations, and individual receivers.
- GPS has revolutionized primary data capture, especially since the development of Differential GPS (Box 11), the removal of selective availability, and the creation of low-cost, low-power receivers.
  - Selective availability was removed in May 2000, so that now users can fix the location of objects relatively easily to an accuracy of better than 10 m.
  - Today units costing less than \$100 can easily provide locational data at better than 10 m accuracy.



21 satellites with three operational spares, 6 orbital planes,  
55 degree inclinations, 20,200 kilometer, 12 hour orbit.

Figure 5.16: GPS.



GPS works according to a simple principle—the length of time it takes a signal to travel from a satellite to a receiver on the ground. The GPS satellites constantly transmit a coded radio signal that indicates their exact position in space and time. The receiver measures how long it takes the signal to travel from the satellites. By measuring the distance from three or more satellites, the location of the receiver can be obtained by triangulation. If a signal can be obtained from a fourth satellite, then the elevation of the receiver can also be determined.

Although standard GPS receivers can provide locations at accuracies of 5–10 m, it is important to understand that there are several possible sources of error inherent in these locations. Some of the errors are random in nature, while others are systematic and can therefore be corrected. Errors arise from signal degradation due to atmospheric effects, minor variations in the location of the satellites, inaccuracies in the timing clocks, errors in receivers, and variations in the reflection of signals from local objects.

A number of techniques are available to improve the accuracy of GPS measurements. Many GPS receivers perform averaging of measurements to improve apparent accuracy. Others snap measurements to map features. So, for example, in-car navigation systems snap the location of the vehicle to a road centerline.

The accuracy of measurements can also be improved by using Differential GPS. This technique uses two receivers. One is fixed and the other is used to collect measurements. If the location of the fixed (base) receiver is known accurately, comparing the exact location with the location reported by GPS will provide an estimate of error. This error can be used to correct measurements obtained from the roving receiver provided that it is within about 300 km. In some countries, the differential correction information is broadcast freely over airwaves and can be received using a standard radio receiver. Differential GPS can improve accuracy to allow locations to be determined to better than 1 meter.

- The drawback of GPS is that it is necessary to have three or more satellites in unobstructed view in order to collect measurements.
  - For example, in forests and urban areas with tall buildings.
- GPS is very useful for recording ground control points for other data capture projects, for locating objects that, and for direct capture of the locations of many types of objects.
- GLONASS is the Russian version of GPS offering similar coverage and accuracy; Galileo is the European Union's proposed equivalent.

# SECONDARY GEOGRAPHIC DATA CAPTURE

- Geographic data capture from secondary sources is the process of creating raster and vector files and databases from maps and other hardcopy documents.
  - Scanning is used to capture raster data.
  - Table digitizing, heads-up digitizing, stereo-photogrammetry, and COGO data entry are used for vector data.

There are three different types of scanner generally used for data entry.

Flat-bed scanner – A common PC peripheral, it is small and inaccurate.

Rotating drum scanner – It is expensive and slow but accurate.

Large-format feed scanner – most suitable for capturing data in GIS. It is quicker, cheaper and accurate.

All scanners work on the same principles, where a scanner has a light source, a background (source document) and a lens. During scanning the absence or presence of light is detected as one of the three components moves past the other two.

Precautions for map scanning in GIS:

**OUTPUT QUALITY:** The output quality of map is very crucial in GIS, it needs to be sharp and clear. Setting up the brightness and contrast levels can enhance the quality of images. In some cases *gamma correction* (a method which looks at histogram of the image and places points strategically along the histogram to isolate data types) or *filtering methods* (selectively removal of noise disturbance).

**RESOLUTION:** This is the density of the raster image produced by the scanning process. The resolution of scanners is usually measured in dots per inch (dpi) as a linear measurement along the scan line. Commonly, 150 dpi for text, 300 dpi for line maps and higher dpi scanning is done for high quality ortho-photos.

**ACCURACY:** The accuracy of the scanned image is important if the image needs to be used in GIS. It needs to fit for its intended use in terms of its physical and cartographic quality. That is why cleaning of scanned map is essential before using it in GIS because stains and folding marks in maps can affect the map accuracy.

**GEOREFERENCING:** The output of a map from scanner needs to be correctly referenced according to the coordinate system used in GIS. Generally, this process is controlled using linear transformation from the row and column number. Distortion across scanned image can create problem if the scanned image is of low quality.

**VECTORIZATION:** The output from scanned maps are often used to generate vector data. This involves, automatic or user controlled raster to vector conversion. Here the resolution of scanned map is very important because it affects the generalization of features in the map.

## ❖ Raster data capture using scanners

- A scanner is a device that converts hardcopy analog media into digital images by scanning successive lines across a map or document and recording the amount of light reflected from a local data source.
- Scanned maps and documents are used extensively in GIS as background maps and data stores.
- Most GIS scanning is in the range 400– 1000 dpi (16-40 dots per millimeter).
  - An 8 bit (256 grey levels) 400 dpi (16 dots per millimeter) scanner is a good choice for scanning maps for use as a background GIS reference layer.
  - For a colour aerial photograph that is to be used for subsequent photo-interpretation and analysis, a colour (8 bit for each of three bands) 1000 dpi (40 dots per millimeter) scanner is more appropriate.



Figure 5.17: Using a scanner.

- There are three reasons to scan hardcopy media for use in GIS:
  - Documents, such as building plans, CAD drawings, property deeds, and equipment photographs are scanned to reduced wear and tear, improve access, provide integrated database storage, and to index them geographically.
  - Film and paper maps, aerial photographs, and images are scanned and georeferenced so that they provide geographic context for other data (typically vector layers).
  - Maps, aerial photographs, and images are also scanned prior to vectorization.
- The quality of data output from a scanner is determined by
  - the nature of the original source material,
  - the quality of the scanning device,
  - And the type of preparation prior to scanning (e.g., redrafting key features or removing unwanted marks will improve output quality).

# VECTOR DATA CAPTURE

- Secondary vector data capture involves digitizing vector objects from maps and other geographic data sources.
- The most popular methods are manual digitizing, heads-up digitizing and vectorization, photogrammetry, and COGO data entry.

## ❖ **Manual digitizing**

- Manually operated digitizers are much the simplest, cheapest, and most commonly used means of capturing vector objects from hardcopy maps.
- They operate on the principle that it is possible to detect the location of a cursor or puck passed over a table inlaid with a fine mesh of wires.
- Vertices defining point, line, and polygon objects are captured using manual or stream digitizing methods.



- Manual digitizing involves placing the center point of the cursor cross hairs at the location for each object vertex and then clicking a button on the cursor to record the location of the vertex.
- Stream mode digitizing partially automates this process by instructing the digitizer control software automatically to collect vertices every time a distance or time threshold is crossed.
- Stream-mode digitizing is a much faster method, but it typically produces larger files with many redundant coordinates.

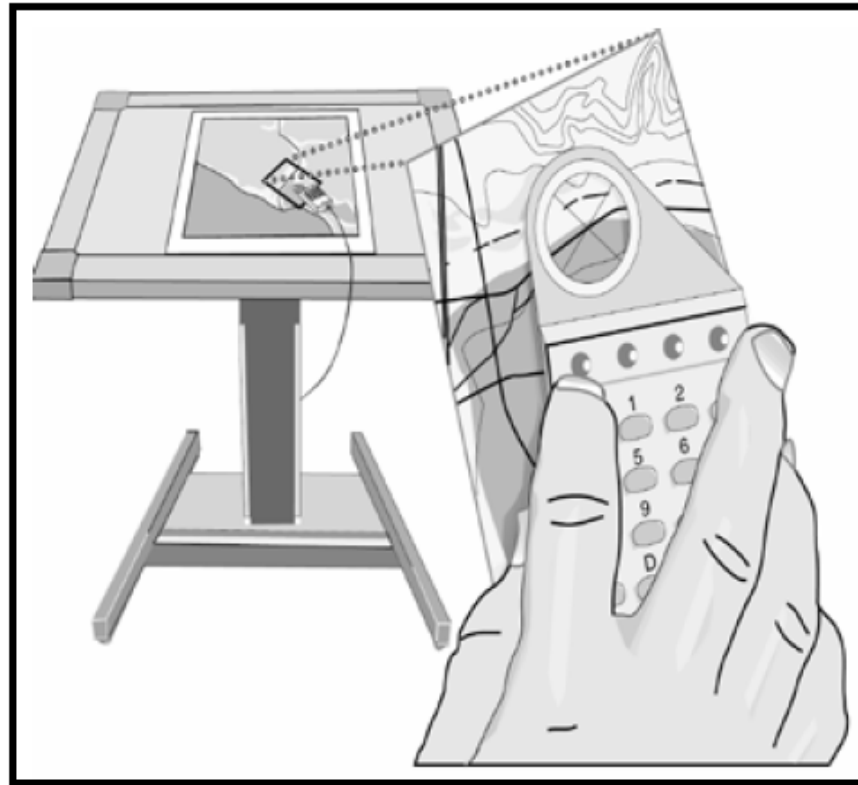


Figure 5.18: Digitizing table.

1. Digitizer cursor transmits a pulse from an electromagnetic coil under the view lens.
2. Pulse is picked up by nearest grid wires under tablet surface.
3. Result is sent to computer after conversion to x and y units.

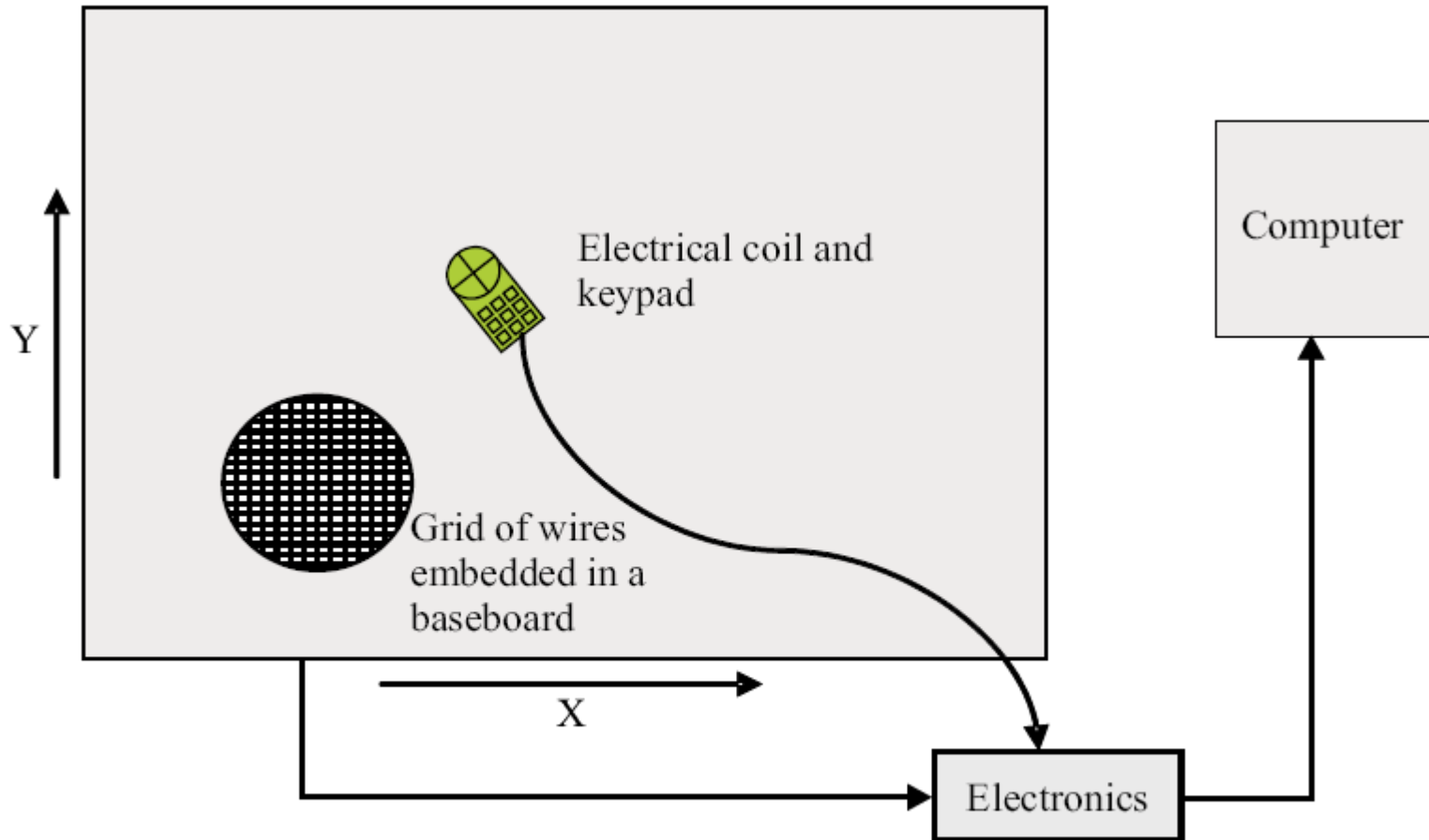


Figure 5.19: The basic components of a digitising tablet.








Terms	Example	Description
Arc		Line feature: a node at each end; vertices at each change of direction.
Node		Endpoint of an arc (also found at intersections between lines).
Vertex		A point on an arc that signals a change of direction.
Pseudo Node		On an (island) arc that connects to itself or where an attribute changes or on a long arc
Dangling Node		Arc endpoint that is not connected.
Label Point		Identifies a point feature or polygon.
Tic		Geographic control point; features can be registered to the same coordinate system.

Figure 5.20: Digitizing terms.

## ❖ Heads-up digitizing and vectorization

- Vectorization is the process of converting raster data into vector data.
  - One of the main reasons for scanning maps is as a prelude to vectorization.
- Heads-up digitizing
  - The simplest way to create vectors from raster layers is to digitize vector objects manually straight off a computer screen using a mouse or digitizing cursor.
  - It is widely used for selective capture of, for example, land parcels, buildings, and utility assets.

- A faster and more consistent approach is to use software to perform automated vectorization in either batch or interactive mode.
- Batch vectorization
  - Batch vectorization takes an entire raster file and converts it to vector objects in a single operation. Vector objects are created using software algorithms that build simple (spaghetti) line strings from the original pixel values.
  - Batch vectorization software is far from perfect and post-vectorization editing is required to clean up errors.
  - Batch vectorization is best suited to simple bi-level maps of, for example, contours, streams, and highways.

- Interactive vectorization
  - For more complicated maps and where selective vectorization is required (for example, digitizing fittings off topographic maps), interactive vectorization is preferred.
  - It is also called semiautomatic vectorization, line following, or tracing.
  - In interactive vectorization, software is used to automate digitizing. The operator snaps the cursor to a pixel, indicates a direction for line following, and the software then automatically digitizes lines.
  - Although quite labour intensive, interactive vectorization generally results in much greater productivity than manual or heads-up digitizing. It also produces high-quality data.

## ❖ Photogrammetry

- Photogrammetry is the science and technology of making measurements from pictures, aerial photographs, and images.
  - In the strict sense it includes 2D measurements taken from single aerial photographs.
  - Today in GIS It is almost exclusively concerned with capturing 2.5D and 3D measurements from models derived from stereo-pairs of photographs and images.
- To obtain true georeferenced coordinates from a model it is necessary to georeference photographs using control points.
- Measurements are captured from overlapping pairs of photographs using stereoplotters.
  - Stereo-pairs usual have 60% overlap along each flight line and 30% overlap between flight lines.
  - Stereoplotter build a model and allow 3D measurements to be captured, edited, stored, and plotted.



- Orthophotographs
  - Orthophotographs result from using a DEM to correct distortions in an aerial photograph derived from varying land elevation.
  - They have become popular because of their relatively low cost of creation (when compared with topographic maps) and ease of interpretation as base maps.
  - They can also be used as accurate data sources for heads-up digitizing.
- Photogrammetry is a very cost effective data capture technique that is sometimes the only practical method of obtaining detailed topographic data about an area of interest.
- Unfortunately, the complexity and high cost of equipment have restricted its use to large scale primary data capture projects and specialist data capture organizations.

## ❖ COGO data entry

- COGO, a contraction of the term coordinate geometry, is a methodology for capturing and representing geographic data.
- COGO uses survey style bearings and distances to define each part of an object.
- The COGO system is widely used in North America to represent land records and property parcels.
- Coordinates can be obtained from COGO measurements by geometric transformation (*i.e.*, bearings and distances are converted into X, Y coordinates).
- COGO data are very precise measurements and are often regarded as the only legally acceptable definition of land parcels.

## OBTAINING DATA FROM EXTERNAL SOURCES (DATA TRANSFER)

- Some of data captured by others are freely available, but many of them are sold as a commodity from a variety of outlets including, increasingly, Internet sites.
- The characteristics and availability of datasets are constantly changing so those seeking an up-to-date list should consult one of the good online sources.
  - The best way to find geographic data is to search the Internet using one of the specialist geographic search engines such as the US NSDI Clearinghouse or the Geography Network.
  - An interesting new trend initiated by the Geography Network Project is the idea of providing data online in ready-to-use GIS formats.
    - The Geography Network is global collection of data users and providers connected by the Internet.
    - Information about available data sources can be found by consulting the Geography Network Web site ([www. GeographyNetwork.com](http://www.GeographyNetwork.com)).

# GEOGRAPHIC DATA FORMATS

- There are so many different geographic data formats because no single format is appropriate for all tasks and applications.
- Many people have asked for tools to move data between systems and to re-use data through open application programming interfaces (APIs).
  - In the former case, the approach has been to develop software that is able to translate data, either by a direct read into memory, or via an intermediate file format.
  - In the latter case, software developers have created open interfaces to allow access to data.
- More than 25 organizations are involved in the standardization of various aspects of geographic data and geoprocessing.
  - Such as ISO (the International Standards Organization) through technical committees TC 211 and 287, CEN(Commission European Normalization).

- Geographic data translation software must address both syntactic and semantic translation issues.
  - Syntactic translation involves converting specific digital symbols (letters and numbers) between systems.
  - Semantic translation is concerned with converting the meaning inherent in geographic information.
  - While the former is relatively simple to encode and decode, the latter is much more difficult and has seldom met with much success to date.

Box 13: Some examples of geographic data formats

Vector	Raster (Image)
Automated Mapping System (AMS)	Arc Digitized Raster Graphics (ADRG)
ESRI Coverage	Band Interleaved by line (BIL)
Computer Graphics Metafile (CGM)	Band Interleaved by Pixel (BIP)
Digital Feature Analysis Data (DFAD)	Band Sequential (BSQ)
Encapsulated Postscript (EPS)	Windows Bitmap (BMP)
Microstation drawing file format (DGN)	Device-Independent Bitmap (DIB)
Dual Independent Map Encoding (DIME)	Compressed Arc Digitized
Digital line Graph (DLG)	Raster Graphics (CADRG)
AutoCAD Drawing Exchange Format (DXF)	Controlled Image Base (CIB)
AutoCAD Drawing (DWG)	Digital Terrain Elevation Data (DTED)
MapBase file (ETAK)	ERMapper
ESRI Geodatabase	Graphics Interchange Format (GIF)
Land Use and Land Cover Data (GIRAS)	ERDAS IMAGINE (IMG)
Interactive Graphic Design Software (IGDS)	ERDAS 7.5 (GIS)
Initial Graphics Exchange Standard (IGES)	ESRI GRID file (GRID)
Map Information Assembly Display System	JPEG File Interchange Format (JFIF)
(MIADS)	Multi-resolution Seamless Image
MOSS Export File (MOSS)	Database (MrSID)
TIGER/line file: Topologically Integrated	Tag Image File Format (TIFF; GeoTIFF)
Geographic Encoding and Referencing (TIGER)	Portable Network Graphics (PNG)
Spatial Data Transfer Standard/Topological	
Vector Profile (SDTS/TVP)	

# CAPTURING ATTRIBUTE DATA

- Attribute data capture is a relatively simple task that can be undertaken by lower-cost clerical staff.
- Attributes can be entered by direct data loggers, manual keyboard entry, optical character recognition (OCR) or, increasingly; voice recognition, which do not require expensive hardware and software systems.
- Metadata are a special type of non-geometric data that are increasingly being collected.
  - Some metadata are derived automatically by the GIS software system (for example, length and area, extent of data layer, and count of features).
  - Some must be explicitly collected (for example, owner name, quality estimate, and original source).
    - Explicitly collected metadata can be entered in the same way as other attributes as described above.

# MANAGING A DATA CAPTURE PROJECT

- In any data capture project there is a fundamental trade-off between quality, speed, and price.
  - Capturing high quality data quickly is possible, but it is very expensive. If price is a key consideration then lower quality data can be captured over a longer period.
- A key decision facing managers of data capture projects is whether to pursue a strategy of incremental capture or 'Blitzkrieg' – that is, to capture all data as rapidly as possible.
  - Incremental data capture involves breaking the data capture project into small manageable sub-projects.
- A further important decision is whether data capture is to use in-house or external resources.
  - Three factors influencing this decision are: cost – schedule, quality, and long-term ramifications.



# DATA EDITING

- Data editing or “cleaning” includes – detection and correction of errors; re-projection, transformation and generalization; and edge matching and rubber sheeting.

## ❖ Detecting and correcting errors

- Errors in input data may derive from three main sources:
  - errors in the source data
  - errors introduced during encoding
  - errors propagated during data transfer and conversion
- Errors in attribute data are relatively easy to spot and may be identified using manual comparison with the original data.
- Errors in spatial data are often more difficult to identify and correct than errors in attribute data.
- Chrisman (1997) suggests that certain types of error can help to identify other problems with encoded data.
  - For example, in an area data layer ‘dead-end nodes’ might indicate missing lines, overshoots or undershoots.

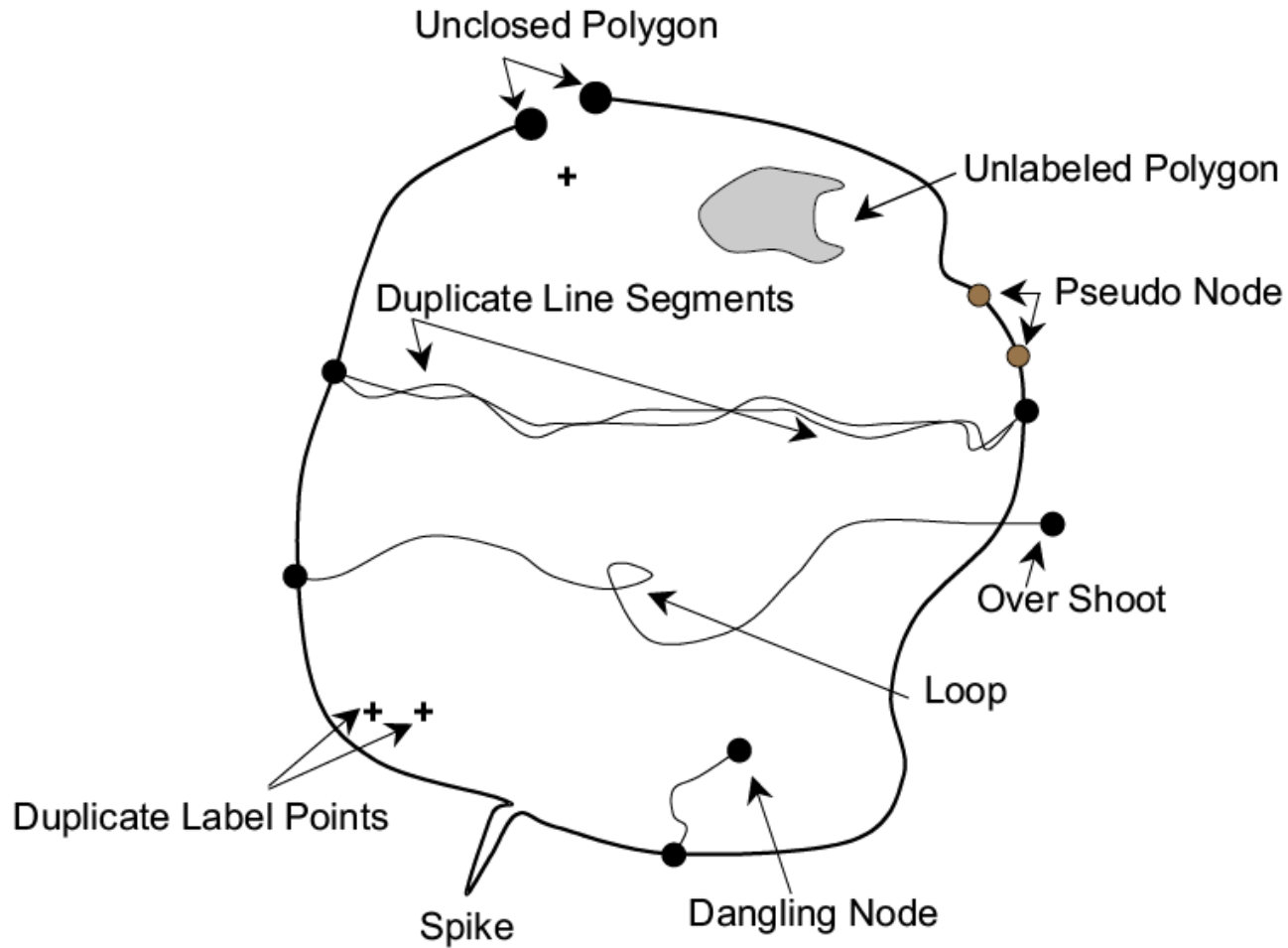


Figure 5.21: Examples of spatial errors.

## Box 14: Common spatial errors

Error	Description
Missing entities	Missing points, lines or boundary segments.
Duplicate entities	Points, lines or boundary segments that have been digitized twice.
Mislocated entities	Points, lines or boundary segments that have been digitized at wrong place.
Missing labels	Unidentified polygons.
Duplicate labels	Two or more identification labels for same polygon.
Artifacts of digitizing	Undershoot, overshoot, loops, spikes etc.
Noise	Irrelevant data entry during digitizing or scanning.

- Most GIS packages will provide a suite of editing tools for the identification and removal of errors in vector data.
  - Corrections can be done interactively by the operator 'on-screen', or automatically by the GIS software.
  - Automatic corrections can save many hours of work but need to be used with care as incorrectly specified tolerances may miss some errors or correct 'errors' that never existed in the first place.
- Noise in raster data may be removed by filtering.
  - Filtering involves passing a filter (a small grid of pixels specified by the user-often a  $3 \times 3$  pixel square is used) over the noisy data set and recalculating the value of the central (target) pixel as a function of all the pixel values within the filter.
  - Genuine features in the data can be lost if too large a filter is used.

## ❖ **Re-projection, transformation and generalization**

- Data derived from maps drawn on different projections will need to be converted to a common projection system before they can be combined or analyzed.
- If the grid systems used may have different origins, different units of measurement or different orientation, it will be necessary to transform the co-ordinates of each of the input data sets onto a common grid system.

- Some of the other methods commonly used are:
  - *Translation and scaling:*
    - If a common grid system of 1-metre coordinates is required, then this is simply a case of multiplying the coordinates in the 10metre data set by a factor of 10.
  - *Creating a common origin:*
    - The origin of one of the data sets may be shifted in line with the other simply by adding the difference between the two origins.
  - *Rotation:*
    - Map co-ordinates may be rotated using simple trigonometry to fit one or more data sets onto a grid of common orientation.
- If source maps of widely differing scales are to be used together, data derived from larger-scale mapping should be generalized to be comparable with the data derived from smaller-scale maps.

- Generalization of vector data
  - The simplest techniques for generalization delete points along a line at a fixed interval (for example, every third point).
    - These techniques have the disadvantage that the shape of features may not be preserved.
  - Most other methods are based on the Douglas-Peucker algorithm (Douglas and Peucker, 1973). This involves the following stages:
    - i. Joining the start and end nodes of a line with a straight line.
    - ii. Examining the perpendicular distance from this straight line to individual vertices along the digitized line.
    - iii. Discarding points within a certain threshold distance of the straight line.
    - iv. Moving the straight line to join the start node with the point on the digitized line that was the greatest distance away from the straight line.
    - v. Repeating the process until there are no points left which are closer than the threshold distance.

- Generalization of raster data
  - The most common method employed is to aggregate or amalgamate cells with the same attribute values.
    - This approach results in a loss of detail which is often very severe.
  - A more sympathetic approach is to use a filtering algorithm.
  - If the main motivation for generalization is to save storage space, then, it may be better to use an appropriate data compaction technique.
    - This will result in a volume reduction without any loss in detail.



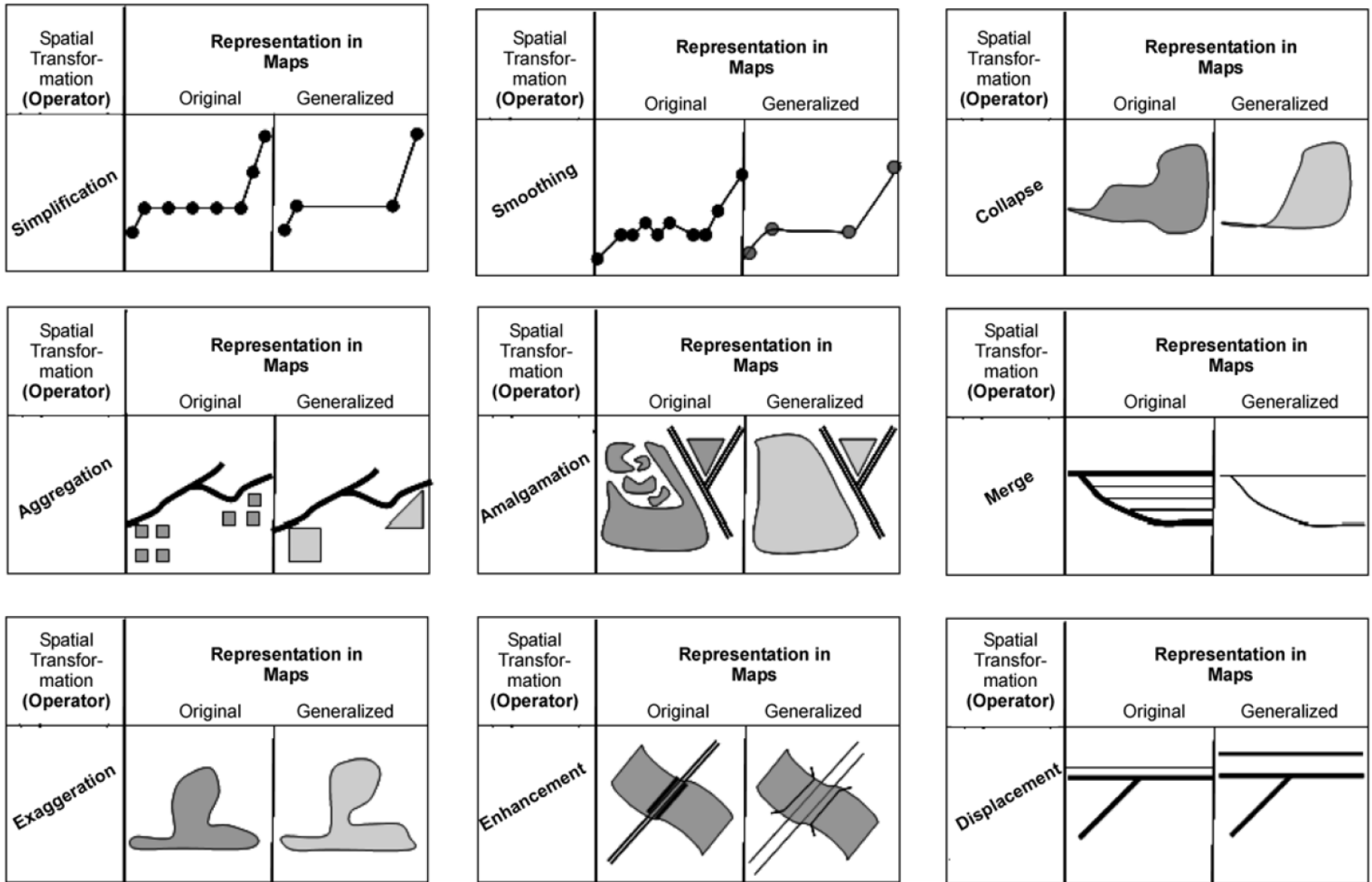


Figure 5.22: Different forms of generalization (Based on McMaster and Shea, 1992).

## ❖ Edge Matching and Rubber Sheeting

- Edge matching
  - Differences or mismatches between adjacent map sheets may need to be resolved.
  - Normally, each map sheet would be digitized separately and then the adjacent sheets joined after editing, re-projection, transformation and generalization. The joining process is known as edge matching
  - Three basic steps of edge matching
    - First, mismatches at sheet boundaries must be resolved.
    - Second, for use as a vector data layer, topology must be rebuilt as new lines and polygons have been created from the segments that lie across map sheets.
    - Finally, redundant map sheet boundary lines are deleted or dissolved (Jackson and Woodsford, 1991)

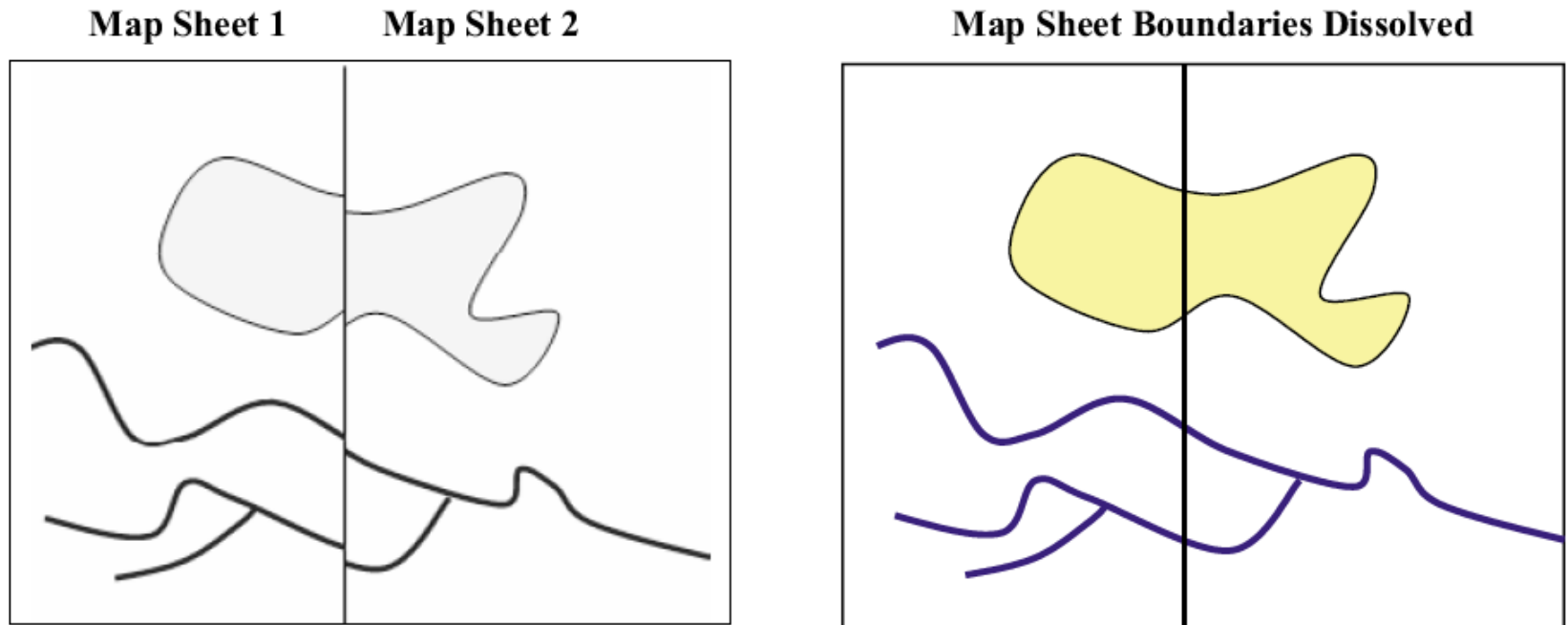


Figure 5.23: Example of edge matching.

- Rubber sheeting
  - Internal inaccuracies in the location of features within the image can be rectified through a process known as rubber sheeting (or conflation).
    - Rubber sheeting involves stretching the map in various directions as if it were drawn on a rubber sheet.
  - Objects on the map that are accurately placed are ‘tacked down’ and kept still; while others that are in the wrong location or have the wrong shape are stretched to fit with the control points.
    - Control points are fixed features that may be easily identified on the ground and on the image.
  - Figure 5.24 illustrates the process of rubber sheeting.
  - This technique may also be used for re-projection where details of the base projection used in the source data are lacking.
  - Difficulties associated with this technique include the lack of suitable control points and the processing time required for large and complex data sets.

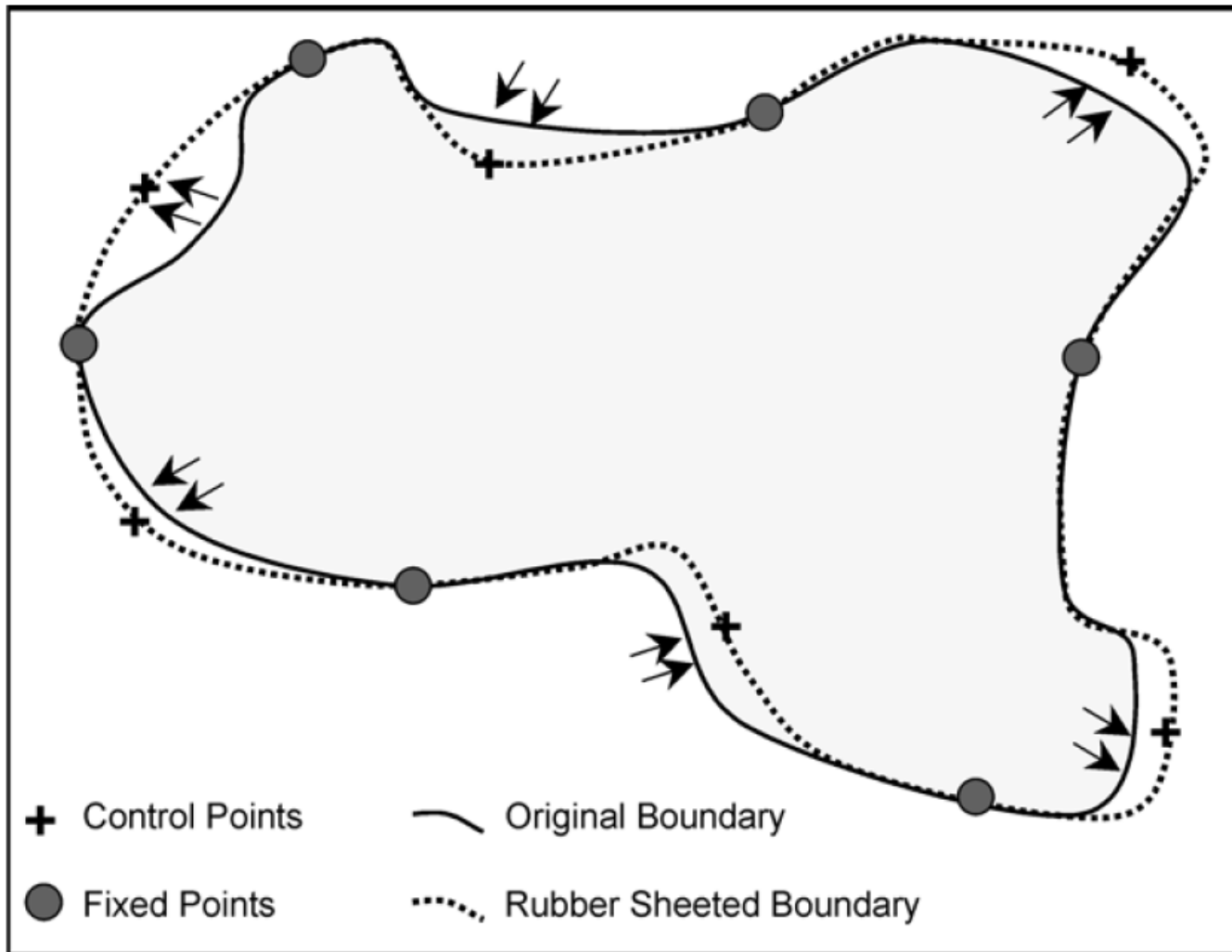


Figure 5.24: Example of rubber sheeting.

## ❖ Geocoding address data

- Geocoding is the process of converting an address into a point location (McDonnell and Kemp, 1998).
  - During geocoding the address itself, a postcode or another non-geographic descriptor (such as place name, land owner or land parcel reference number) is used to determine the geographical co-ordinates of a location.
- Address matching is the process of geocoding street addresses to a street network.
  - Locations are determined based on address ranges stored for each street segment.

- Address data are frequently inconsistent:
  - place names may be spelt incorrectly,
  - addresses may be written in different formats
  - different abbreviations exist for words that appear frequently in addresses,
- The use of standards for address data is particularly relevant to geocoding.

# DATA CONVERSION

- Vector to raster
  - Usually the conversion is from vector to raster, because the biggest part of the analysis is done in the raster domain.
  - Vector data are transformed to raster data by overlaying a grid with a user-defined cell size.
- Raster to vector
  - This is the case especially if one wants to achieve data reduction because the data storage needed for raster data is much larger than for vector data.
- Remote sensing images usually have to be converted into the format of the spatial (raster) database before they can be downloaded.



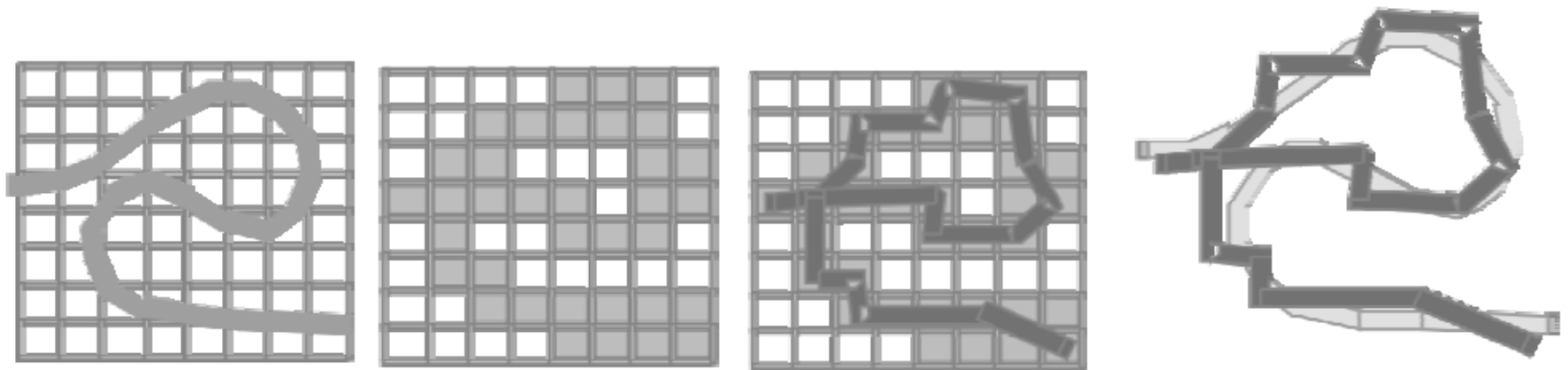


Figure 5.25: Vector to raster exchange errors.

# GEOGRAPHIC DATA – LINKAGES AND MATCHING

- A GIS can carry out linkages and matching because it uses geography, as a common key between the data sets.
- Information is linked only if it relates to the same geographical area.

## ❖ Linkages

- A GIS typically links different sets.
- For example, we have one file that contains the number of children in this age group, and another that contains the mortality rate from malnutrition.
- Suppose we want to know the mortality rate to malnutrition among children under 10 years of age in any state.
- We must first combine or link the two data files. Once this is done, we can divide one figure by the other to obtain the desired answer.

## ❖ Exact Matching

- Exact matching means when we have information in one computer file about many geographic features (e.g., towns) and additional information in another file about the same set of features.
- The operation to bring them together is easily achieved by using a key common to both files -- in this case, the town name.
- Thus, the record in each file with the same name is extracted, and the two are joined and stored in another file.

## ❖ Hierarchical Matching

- Some types of information are collected in more detail and less frequently than other types of information.
  - For example, land use data covering a large area are collected quite frequently; On the other hand, land transformation data are collected in small areas but at less frequent intervals.
- If the smaller areas nest (*i.e.*, fit exactly) within the larger ones, then the way to make the data match of the same area is to use hierarchical matching -- add the data for the small areas together until the grouped areas match the bigger ones and then match them exactly.

## ❖ Fuzzy Matching

- On many occasions, the boundaries of the smaller areas do not match those of the larger ones.
  - For example, crop boundaries rarely match the boundaries between the soil types.
- If we want to determine the most productive soil for a particular crop, we need to overlay the two sets and compute crop productivity for each and every soil type.
- This is like laying one map over another and noting the combinations of soil and productivity.