

18MAG34E -GIS AND ITS APPLICATIONS-UNIT 3

DATA EDITING

Geographic Information System simply represents real world conditions with the aid of computer. It is a tool for analyzing the problems. For that we need some data. It may be spatial or non-spatial. These data may include errors. We could expect errors from the original source as well as derived during encoding. Before the processing of data it is essential to identify and eliminate the error, otherwise it will contaminate the GIS data base.

The pre-processing of GIS data i.e. data editing can be grouped in to the following:

- Detecting and correcting errors
- Reprojection, transformation and generalization.
- Edge matching and rubber sheeting.

DETECTION AND CORRECTION OF ERRORS

Errors in input data may derive from three main sources. They are:

Errors in the sources of data: It may the errors in maps used by surveyors or printing errors.

Errors while encoding: It may be scanning errors, digitizing errors, typing errors, etc.

Errors at the time of transfer and conversion: While transferring and converting data different formats makes errors and data loss.

ATTRIBUTE DATA EDITING

Attribute data may also have some errors and it could be identify easily by manually and could compare with original data. There are many methods for checking and correcting attribute data. Some of them are:

- * **Impossible values:** We could check the error value, when we know the range of data.
- * **Extreme values:** We could identify the errors in the data by extreme values.
- * **Internal consistency:** By tallying we could check the error in total and averages.

- * **Scatter diagram:** The error in correlation of two attribute data could be identify using scatter diagrams.
- * **Trend surface:** It will highlight the points which have a long range from other figures.

SPATIAL DATA EDITING

Spatial data error creates more problems and it is difficult to identify. We have two types of spatial data one is raster and the other is vector. Both types have different types or errors and correcting measures.

1. **RASTER DATA EDITING**

Raster data editing refers to correcting specific contents of raster images than their general characteristics. Commonly used raster data editing functions are:

Filling holes and gaps: it used to fill holes and gaps appear in raster images.

Edge smoothing and boundary simplification: remove or fill single pixel irregularities along line edges.

Deskewing: it is used to rotate the image.

Speckle removal or filtering: to remove speckles or random high or low valued pixels in the image.

Erase and delete: remove unwanted pixels.

Thinning: to reduce the representation of linear features to single cell width. It is done to preserve sharp corners and round corners.

Clipping: to cut and remove specific portions of raster image.

Drawing and rasterisation: to add vector graphics or text to raster form in a new image.

Raster Editing Functions	Raster Before Editing	Raster After Editing
Filling holes and gaps		
Edge smoothing or boundary simplification		
Deskewing		
Speckle removal or filtering		
Erasing and deleting		
Thinning		
Clipping		
Drawing and rasterizing		

VECTOR DATA EDITING

Errors may occur in vector data also. These errors are mainly because of digitizing process. Most GIS packages are providing editing tools for identification and removal of errors in vector data. Some of the errors and correcting measures are:

PSEUDO NODES

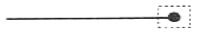

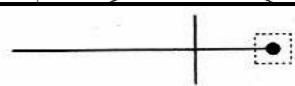

These are false nodes occur where a line connects itself, or where two lines intersect along a parallel path rather than crossing. These incorrect nodes can be corrected by either selecting or deleting when necessary or by adding nodes where needed to convert a polygon.

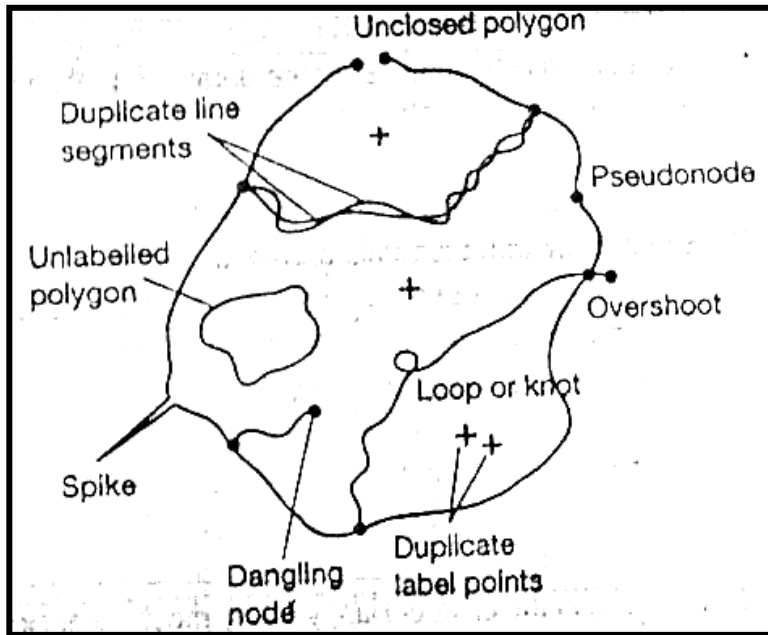
DANGLING NODES

It can be defined as a single node connected to a single line entity and it can be result from three possible mistakes.

- * Failure to close the polygon [unclosed polygon]: failure to close the polygon.
- * Undershoot: failure to connect the node to the object it was supposed to be connected.
- * Overshoot: if a node going beyond the entity where it is supposed to be connected is called overshoot.

For undershoot, the node is moved or snapped to the object to which it should be connected. Overshoot errors can be corrected by identifying intended line intersection points and clipping the line so that it connects where it is supposed to. Open polygon merely move one of the nodes to connect with other.

Feature	Errors	Examples
Node	Dangling node	
	Pseudo-node	
Overshoot		
Undershoot		



LABEL ERRORS

There are two types of errors that can occur related with polygons. One is missing labels and the other is too many labels. We can rectify it by adding or deleting labels wherever necessary.

Label	Missing label	
	Duplicate/multiple labels	

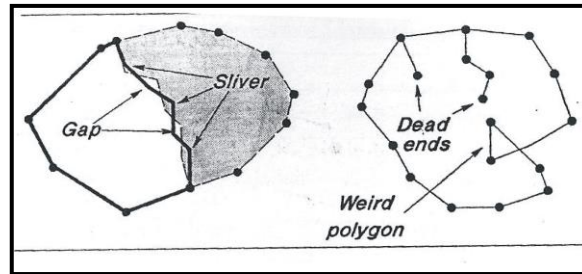
SLIVER POLYGONS

Vector data creates each polygon as a separate entity. In such cases, if required to digitize the adjacent lines between polygons more than once, failure to digitizing exactly will open result

of overlay operation. The easiest way to avoid this does not require digitizing the same line twice and requirement is become very common.

WEIRD POLYGON

These are inadmissible loops occurs while digitizing.

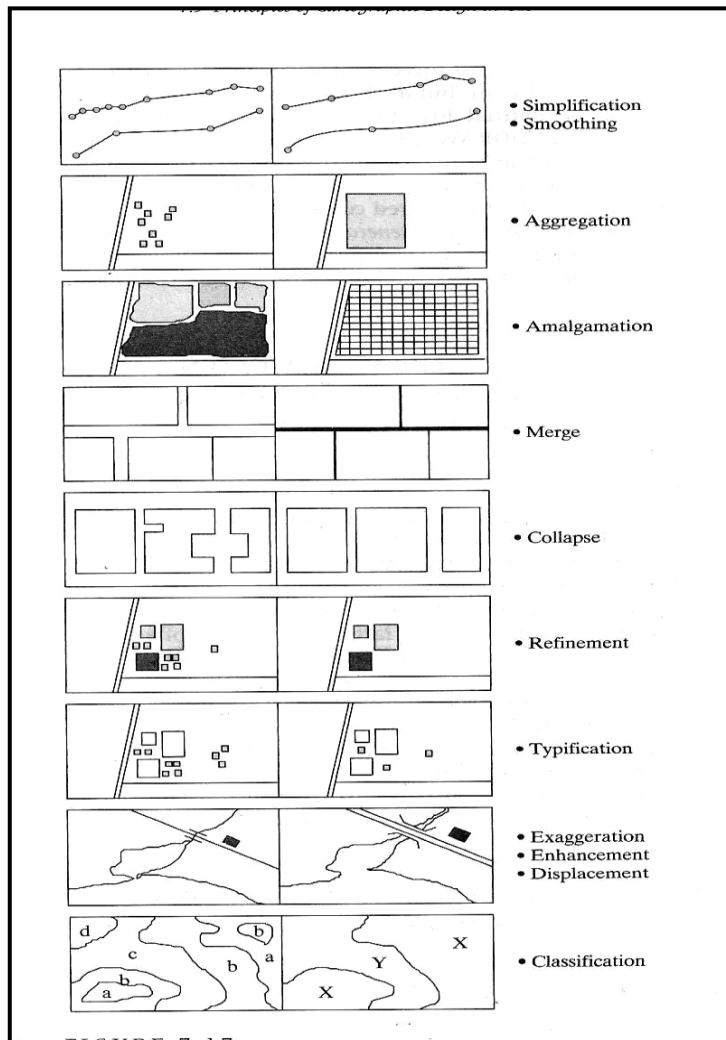


REPROJECTION, TRANSFORMATION AND GENERALIZATION

Once spatial and attribute data have been encoded and edited, it is necessary to process data geometrically in order to provide a common reference. The data derived from various sources should be converted in to a common projection before they combined and analyzed. If it not reprojected, data derived from a source map using one projection will not plot the same location data derived from another source using another projection.

Data derived from different sources may also have different co-ordinate systems. They may have different origins, units of measurements and orientations. So it is necessary to transform it in to a common grid system. It involves some mathematical calculations.

Data may be derived from different maps with different scales. The generalization should be done while comparing data of large and small scales. This will also helps to save time and reduce the space of storage. The simplest method for generalization is to delete points between two points with in a specified interval. But it will not preserve the space of the object. When we generalize a map data loss is a min problem. But it is necessary with comparison of different scale maps. Instead of this compaction technique could be used it will help to reduce the space with out any data loss.



Methods of generalization

EDGE MATCHING AND RUBBER SHEETING

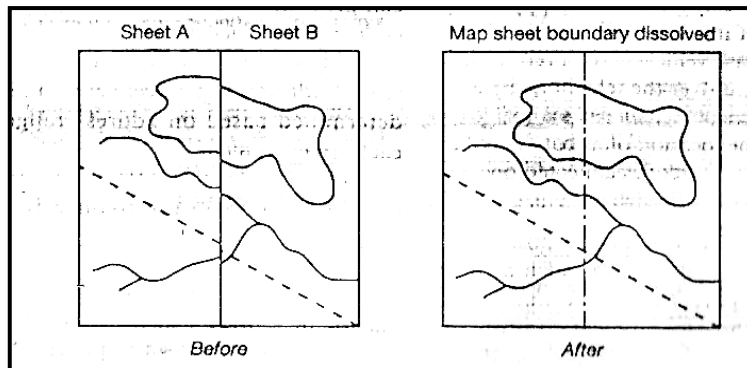
When our study area extends across two or more map sheets, small difference and miss matches may occur. For that normally each map sheets would be digitize separately and then adjacent sheets joined after editing, projection, transformation and generalization. This joining process is known as edge matching. This involves three basic steps:

Mismatches at sheet boundaries must be resolved. When the maps are joining, the adjacent lines and polygons may not join. It should be corrected to complete features and ensure that the data are correct topologically.

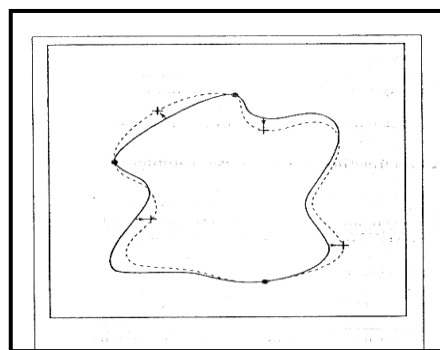
For use as a vector data layers topology must be rebuilt as new lines and polygons have been created from the segments that lie across the sheet. It could be automated, but the problem may arise due to tolerance. If the tolerance is too large, some small lines and polygons may miss. If the tolerance is too small some of the lines may remain unjoined.

The redundant map sheet boundary lines should be deleted or dissolved.

Some data sources may have some internal distortions with in individual map sheets. E.g.: some aerial photographs may have internal inaccuracies even after Reprojection due to movement of air craft or distortion due to camera. This could be rectified by ‘rubber sheeting.’ This process involves stretching of maps in various directions as it drawn in a rubber sheet. Some control points are fixing and map is stretched. Lack of control points and the processing time makes problems.



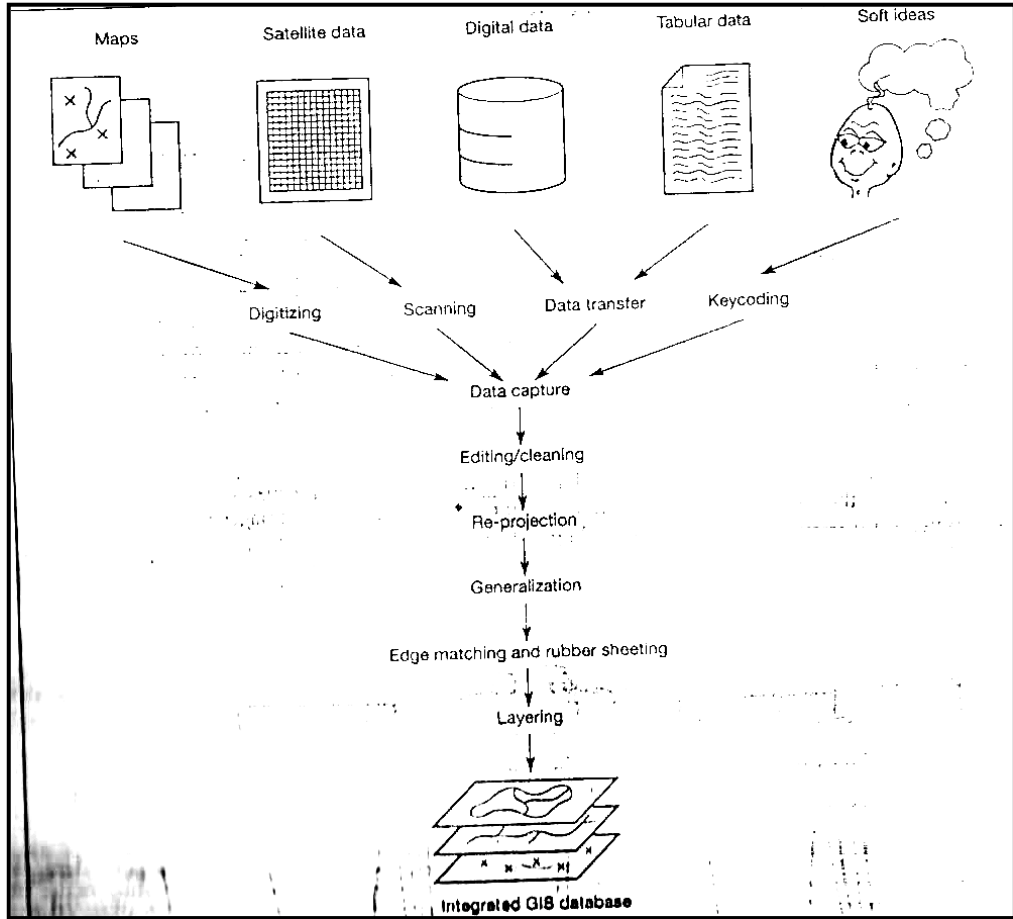
Edge matching



Rubber sheeting

TO AN INTEGRATED GIS DATA BASE

We are preparing an integrated GIS data base using the edited and reprojected data from various sources.



Data stream which leads to an integrated data base

DATA ANALYSIS TERMINOLOGY, MEASUREMENTS IN GIS, QUERIES

Data analysis

- The nature of data used as been reviewed, and encoding and structuring of these data to produce a computer representation of the real world.
It takes the user from data to information and to decision making.
It covers some of the options in GIS for data analysis
- There is a wide range of function for data analysis in GIS package, they are measurement techniques attribute queries, proximity analysis, over lay operations and models of surface networks.

Entity:

An individual point, line, area in G I S.

Attribute:

Data about an entity. In a vector GIS attributes are stored in a database.example, the street name for a line entity that represents a road. In raster the value of a cell in raster grid is numerical code used to represent the attribute present.

Features:

An object in a real world to be encoded in a GIS data base.

Data layer:

A data set for the area of interest in a GIS.Some of the data layers are land use, soils, hydrology, geology,roads,village,hotels,etc.

Image:

A data layer in a raster GIS.It should be rembered that each cell in a raster image will contain single value that is a key to the attribute present there.

Cell:

An individual pixel in a raster image.

Function or operation:

A data analysis procedure performed by a GIS.

Algorithm:

The computer implemented of a sequence of actions designed to solve a problem.

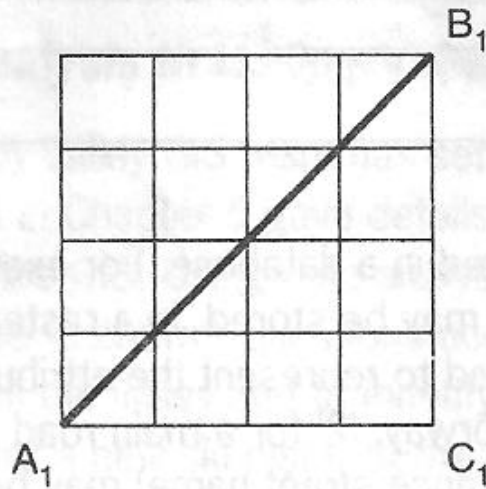
Measurements in GIS- Lengths, perimeters, and areas.

All measurements from a GIS will be an approximation, since vector data are made up of straight line segments (even curve lines are stored as collection of short straight line segments), and all raster entities are approximated using a grid cell representation.

Raster GIS measurement

Length:

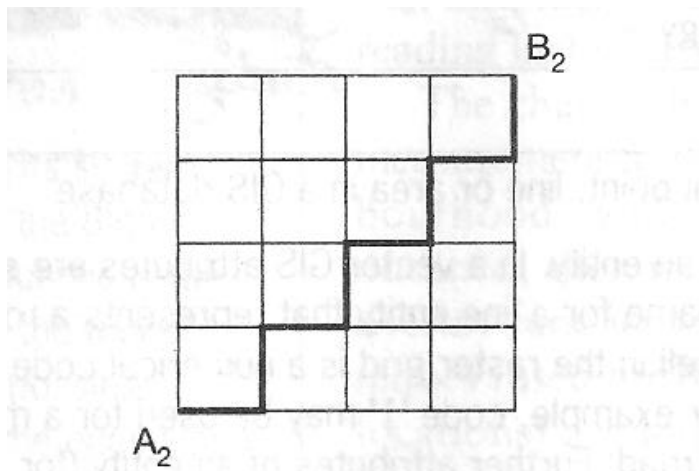
(Pythagorean distance)



$$(a) A_1B_1 = \sqrt{A_1C_1^2 + C_1B_1^2}$$
$$= 5.7 \text{ units}$$

It is calculated by drawing a straight line between the end points of a line and creating a right angle triangle so that Pythagorean geometry can be used to calculate the distance AB.

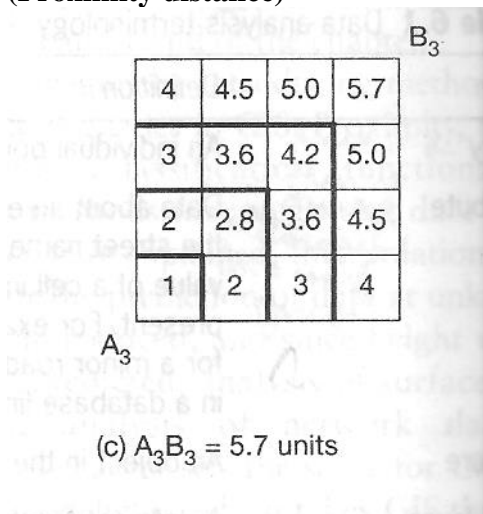
Length:
(Manhattan distance)



(b) $A_2B_2 = 8$ units

This distance can be calculated along the raster cell sides from one point to the other. The name for this method comes from the way in which you would get across, a city, like manhattan, consisting of dense blocks of building ,as it is Impossible to pass diagonal through a block have traverse the sides)

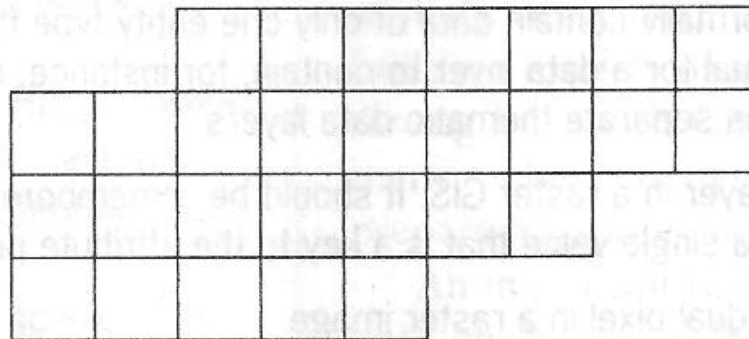
Length:
(Proximity distance)



(c) $A_3B_3 = 5.7$ units

In this method concentric equidistant zones are established around the start point A. The resulting shows the shortest straight line distance from every point on the map including end point B to location of interest A. Thus distance A to B calculated.

Perimeter and area:



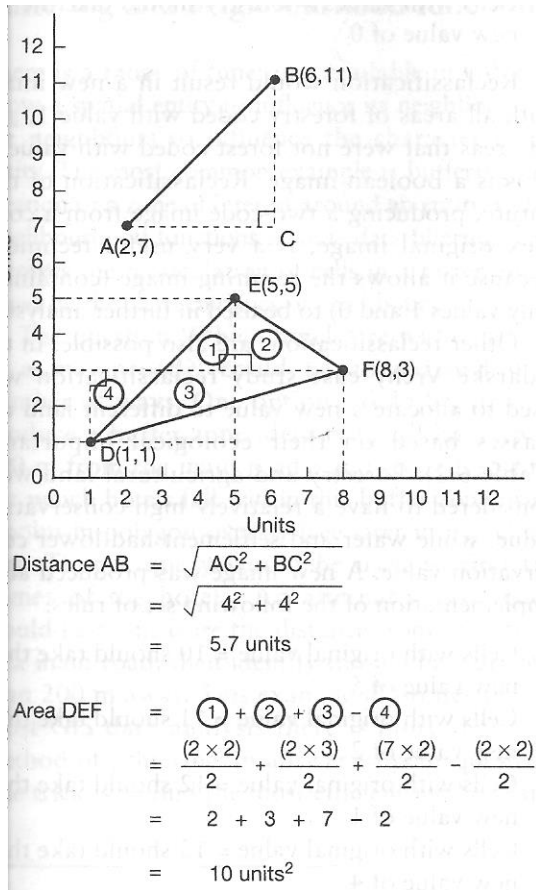
(d) Perimeter = 26 units

Area = 28 units²

To find perimeter in raster GIS, the number of cell sides that make up the boundary of a feature is multiplied by the known resolution of raster grid.

For area calculation the number of cells a feature occupies is multiplied by the known area of the individual grid cell.

1. Vector GIS measurement



In vector GIS distance is measured using Pythagoras theorem to obtain the distance. Perimeter is calculated by built up of the sum of straight line and lengths. Areas are calculated by totaling the areas of simple geometric shapes formed by subdividing the feature of interest.

Queries

Queries on GIS data base to retrieve data, is an essential part of most GIS projects

Queries offer a method of data retrieval, and can be performed on data that are part of GIS data base.

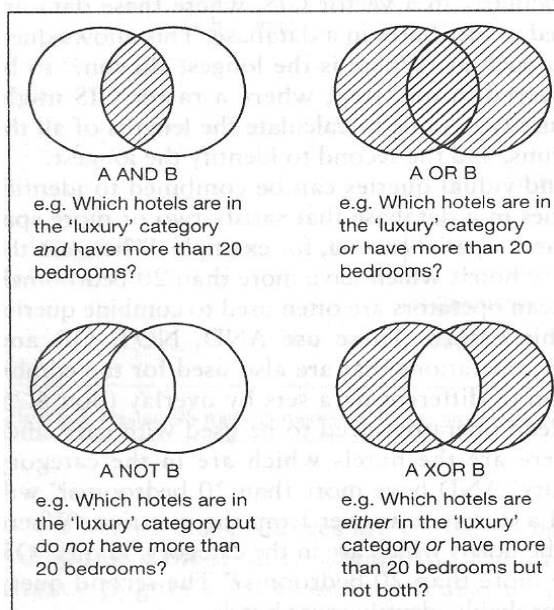
Queries are useful at all stages of GIS analysis for checking the quality of raster GIS measurement. Example , if a data point representing a hotel is found erroneously in the sea after data encoding .A query may establish that the address of the hotel had been entered in a data base, resulting in the allocation of an incorrect spatial reference.

There are two types of query can be performed in GISspatial and aspatial.

Queries can be made more complex by combination of questions about distance , areas, and perimeters.Particlrly in vector GIS,where these data are stored as attributes in a data base. This allows questions such as “where is the longest ski run?” ,to be answered in one step, where a raster GIS might require two, one to calculate the length of all the ski runs, and the second to identify the longest.

Boolean Operators Queries:

Individual queries can be combined to identify entities in a data base that satisfy two or spatial and aspatial criteria ,example with the help of Boolean operators are used to combine the queries, the use of AND,NOT,OR,XOR,operations that are also used for the combination of different data set.



Here in the circle A is the set of luxury hotels and circle B is the set of hotels having more than 20 bed rooms.

RECLASSIFICATION BUFFERING AND NEIGHBOURHOOD FUNCTIONS.

Reclassification based on locational attributes is very useful, but it limits us to the attributes within each object or feature. It would be nice if we could classify features based on a bird's-eye view.

Reclassification is an important variation on the query idea in GIS. And can be used in place of a query in raster GIS, Creating a new raster model by classification is often referred to as Reclassification.

METHODS OF RECLASSIFICATION

Two methods of data reclassification may be used.

The first method is a one-to-one change meaning a cell value in the input grid is assigned a new value in the output grid. Ex, irrigated cropland in a land use grid is assigned a value of 1 in the output grid.

The second method assigns a new value to a range of cell values in the input grid.

Ex, Cells with population density between 0 and 25 persons per square mile in a population density grid are assigned a value of 1 in the output grid.

Both methods of reclassification can be applied to integer grids. Floating-point grids. One the other hand , can only be reclassification by assigning a new value to a range of cell values in the input grid. For Ex, a value of 2 is assigned to cells with slope values between 10.0 and 20.00% Reclassification of a floating-point grid results in an integer grid.

Where are all the areas of forestry, could be obtained using a query or by reclassifying the image. Ex,If cells representing forestry in the original image had a value of 10, a set of rules for the reclassification could be.

Cells with values = Forestry (value 10) should take the new value of 1

Cells with values = Forestry should take the new value of 0

Reclassification would result in a new image with all areas of forestry coded with value 1, and all areas of forestry coded with value 0. Reclassification of this nature, producing a two coded image from a complex original image is a very useful technique because it allows the resulting image to be used in further analysis.

The Zdarsk Vrchy case study reclassification was used to allocate a new value to different land use classes based on their ecological importance. Forestry and agricultural land were considered to have a relatively high conservation value. While water and settlement had lower conservation value. A new images was produced after implementation of the following set of rules.

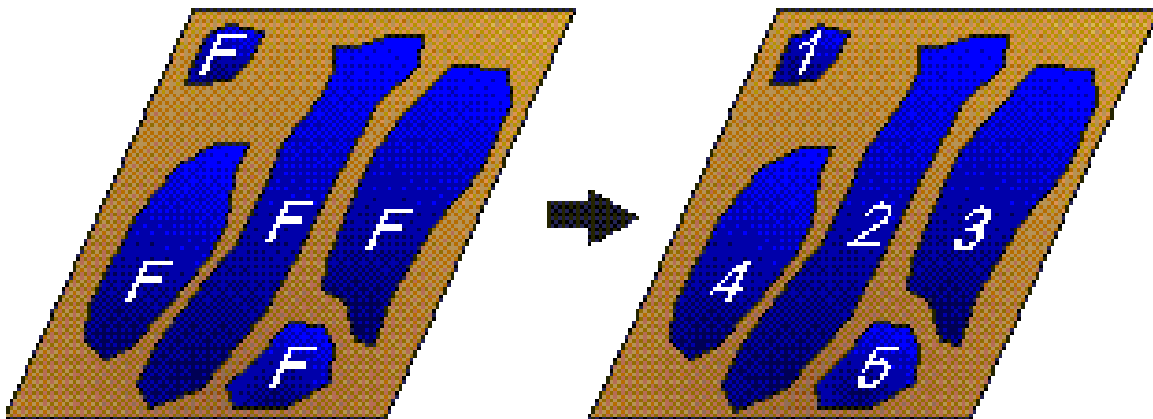
Reclassification values for the land use data layer

Land use	Old value	New value after reclassification: Boolean example	New value after reclassification: Weighting example
Forestry	10	1	5
Water	11	0	2
Settlement	12	0	1
Agricultural land	13	0	4

Cells With original value= 10 should take the new value of 5
Cells With original value= 11 should take the new value of 2
Cells With original value= 12 should take the new value of 1
Cells With original value= 13 should take the new value of 4

The resulting map helped the ecologist to identify areas of high conservation value. However, the new classes (1,2,4 and 5) are still simply labels and care needs to be taken to ensure appropriate further analysis of the image.

Reclassify by Contiguity



Work with individual forest stands, rather than the class forest as a whole.

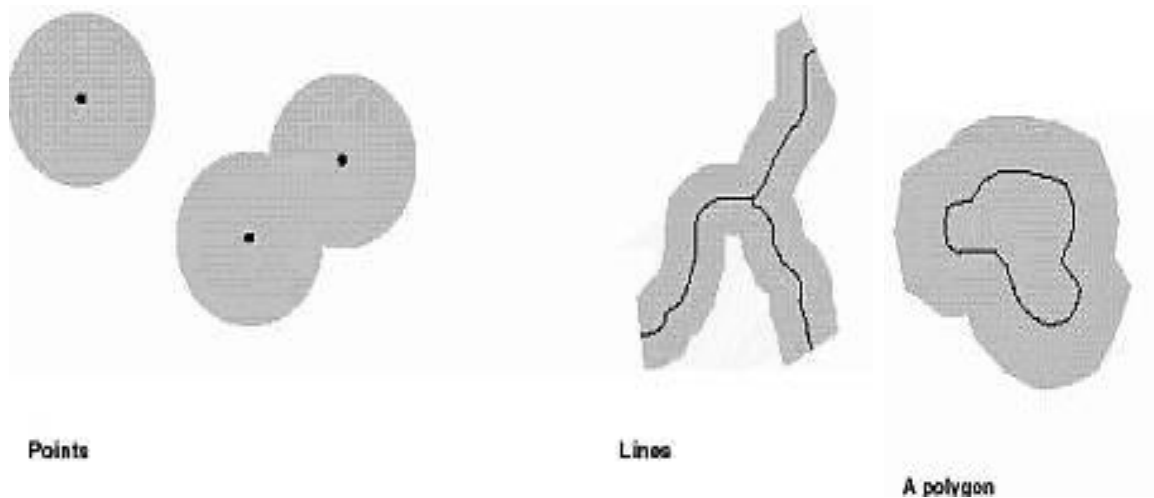
BUFFERING

The creation a zone of interest around an entity. It is an important function used to determine spatial proximity or nearness of various features by defining a distance zone around map features. Buffer can be generated for points, lines and polygons.

Based on the concept of proximity, buffering separates a map into two areas.

- One area that is within a specified distance of selected map features and the other area that is beyond.
- The area that is with in the specified distance is called the buffer zone.
- Selected map features for buffering may be points, lines, areas.
- Buffering around points creates circular buffer zones extending outward or from the polygon boundaries.
- It is used to identify a zone of interest around an entity, or set of entities.

Example: Connectivity (Vector)



Proximity Operation: Buffer Types

A GIS operation in which areas that are within a specified distance of selected map features are separated from areas that are beyond. The area that is within the specified distance is called the buffer zone. Selected map features for buffering may be points, lines, areas. Buffering around points creates circular buffer zones. Lines creates a series of elongated buffer zones. Buffering around polygons creates buffer zones extending outward or from the polygon boundaries.

Buffering as already stated, is used to identify a zone of interest around an entity, or set of entities. If a point is buffered a circular zone is created,. Buffering lines and areas creates new areas. Creating buffer zones around point features is the easiest operation , a circle of the required radius simply drawn around each point creating buffer zones around line and area features is more complicated.

GIS do this by placing a circle of the required radius at one end of the line or area boundry to be buffered.

This circle tangential to the line makes is used to define the boundry of the buffer zone. Only the most basic set of buffer operations as there are many variations on this theme. For Ex, buffer zone may be of fixed or varying width according to feature attributes. When analyzing a road network, wide buffer zones could be attached to motorways and narrower buffer zone to minor roads to reflect traffic densities.

*Example: Connectivity (Vector)
Proximity Operation - Buffers & Setbacks*

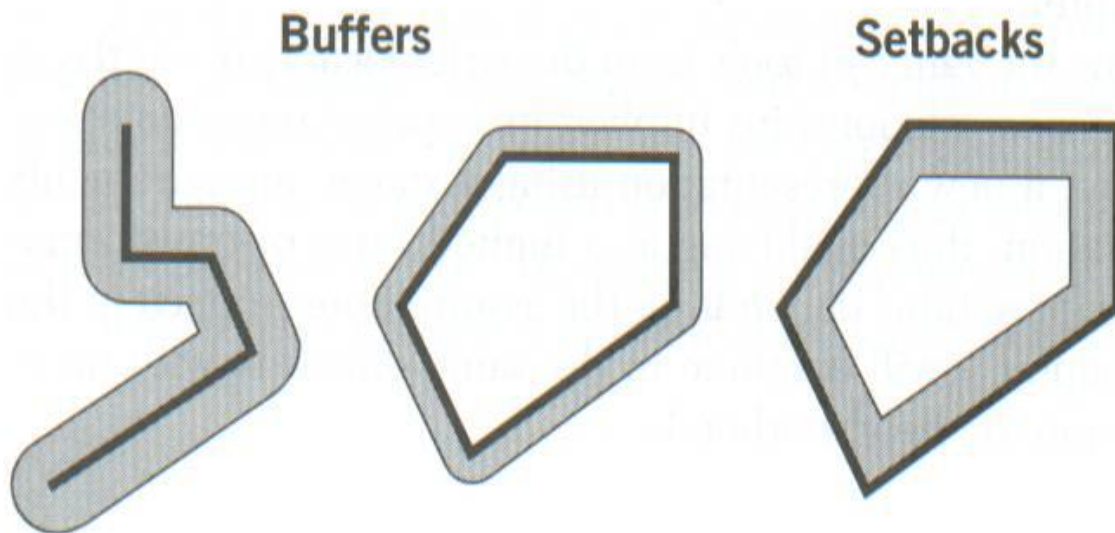


Diagram of simple buffers and a setback.

NOTE: Buffers go outward from lines or areas; setbacks run inside of areas (not lines).

BUFFERING TECHNIQUES

- Buffer zones of varying widths were drawn around roads and railway lines.
- Motorways were buffered at 3 km and primary routes at 1.5 km. Railways were also buffered at 3 km sites lying within any of these buffer zones were potentially feasible waste disposal sites as they were sufficiently accessible.
- Buffering has a whole series of uses when combined with other data layers.
- In the radioactive waste example the buffer zone were used as part of a process to identify the land use.
- Population totals and conservation status of accessible land
- Population data to identify the best locations for bus stops.
- The population density within this buffer zone was then calculated.

Example: Connectivity (Vector)



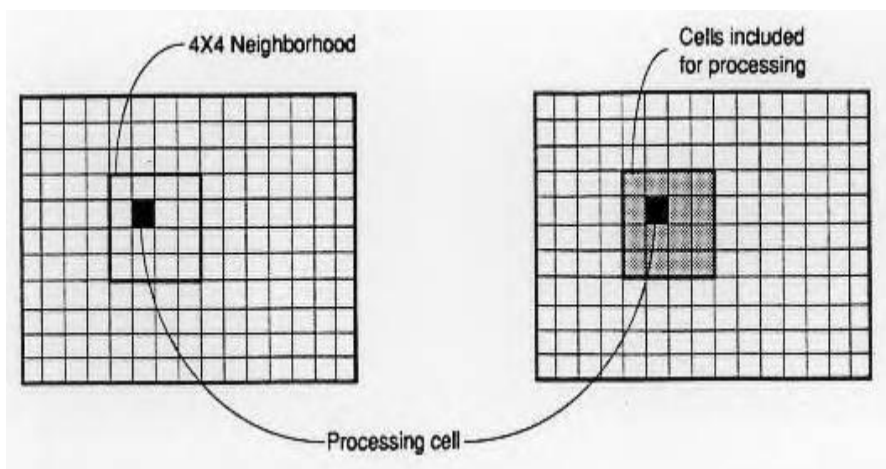
Proximity Operation: Road Buffer

NEIGHBORHOOD FUNCTIONS

- Such reclassification procedures are called “Neighborhood functions” because the idea is to characterize each object as part of a larger neighborhood of objects based on some shared attributes.
- It is easy to recall growing up in a neighborhood. you might have classified that neighborhood based on size or because you happened to interact with the people in it.
- Neighborhoods may also be defined in terms of a unifying attribute for an entire area or the focus may be on smaller portions of the total area (A targeted analysis)
- A targeted analysis also called immediate neighborhoods, includes only locations that are in immediate proximity (adjacent to) the target area or location.
- The total neighborhoods analysis also called extended neighborhoods, includes locations that are immediately in contact with the target area or location plus areas some distance beyond.
- Although this separation of neighborhood functions is both intellectually stimulating and quite useful to the advanced GIS analysis.
- Neighborhood functions operating on 2 and 3-dimensional objects.
- Be able to separate both this neighborhood functions types as either static neighborhood functions.
- The analysis takes place all at once for the selected target area or rowing window neighborhood function.

Neighborhood Functions:

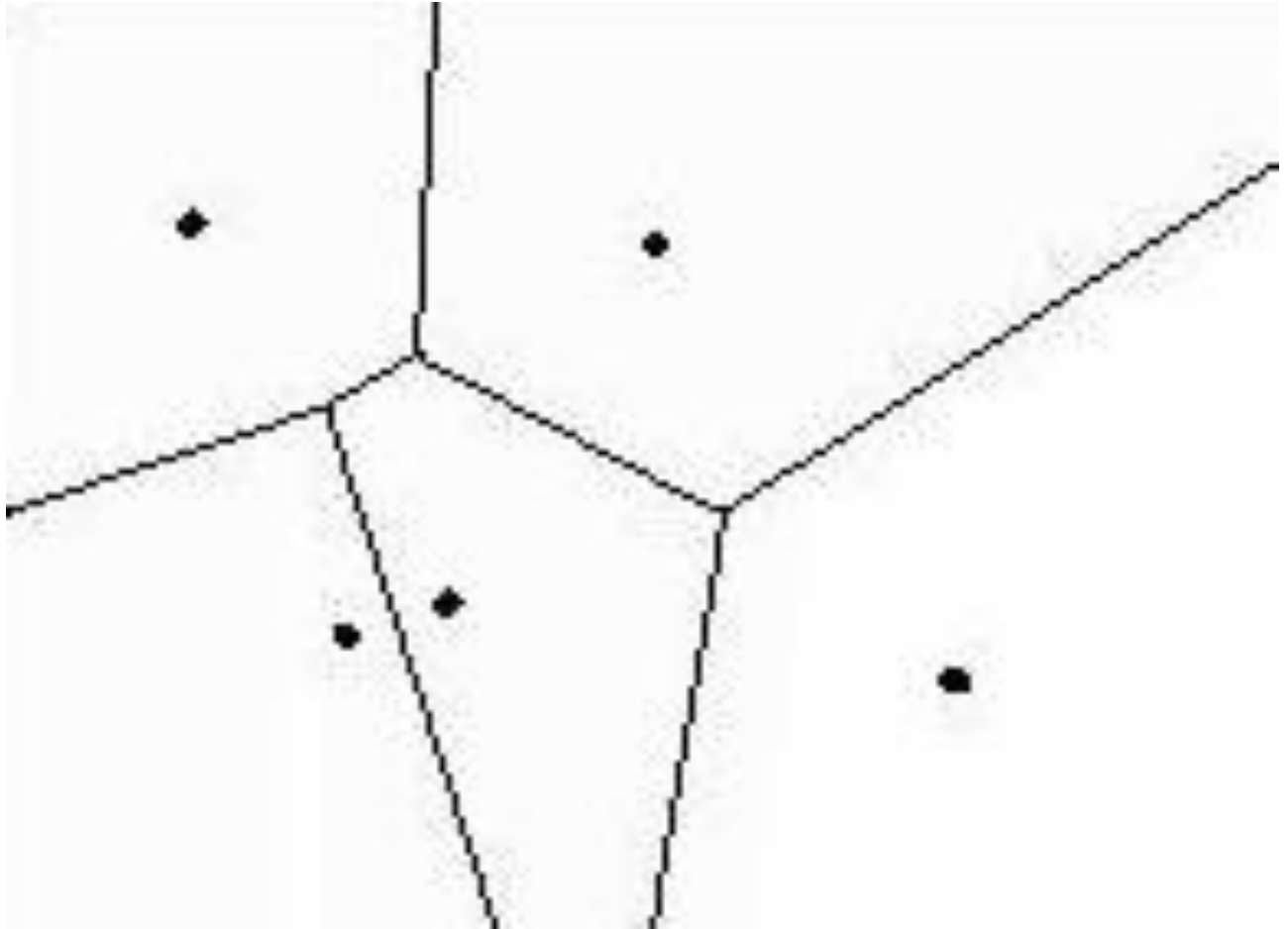
4 x 4 Window Processing



Thiessen Polygons Operation

- Defines the individual area of influence around a point
- Used to predict values at surrounding points from a single point observation
- Can produce polygons with shapes unrelated to phenomenon being mapped

Example: Neighborhood Function



Thiessen Polygons

Filtering

- The processing of remotely sensed imagery. Filtering will change the value of a cell based on the attributes of neighboring cells.
- The filter is defined as a group of cells around a target cell. The size and shape of the filter are determined by the operator.
- Filter shapes are squares and circles and the dimensions of the filter determine the number of neighboring cells used in the filtering process.
- The filter is based across the Raster data set and used to recalculate the value of the target cell that lies at its center. The new value assigned to the target cell is calculated using one of a number of algorithms. Examples include the maximum cell value within the filter and the most frequent value.

The filter and other filtering techniques are shown in filtering operations might be used in the preparation of the forestry data layer in the Happy Valley GIS. The raster data obtained from a classified map may require filtering to 'smooth' noisy data caused by high spatial variability in vegetation cover or problems with the data collection device.