UNIT-IV

SPATIAL DATA ANALYSIS

6.1 CLASSIFICATION OF ANALYTICAL GIS CAPABILITIES

There are many ways to classify the analytical functions of a GIS. The classifications are the one put forward by Aronoff making the following distinctions

1. Classification, retrieval, and measurement functions

All functions in this category are performed on a single (vector or raster) data layer, often using the associated attribute data.

- Classification allows the assignment of features to a class on the basisof attribute values or attributes ranges (i.e. definition of data patterns). On the basis of reflectance characteristics found in a raster, pixels may be classified as representing different crops, such as cotton and jute.
- Retrieval functions allow the selective search of data. Example: retrieve all agricultural fields where cotton is grown.
- Generalization is a function that joins different classes of objects with common characteristics to a higher level (generalized) class. For ex ample, we might generalize fields where potato or maize, and possibly other crops, are grown as 'kharif crop fields'.
- Measurement functions allow the calculation of distances, lengths, or areas.

2. Overlay functions

These belong to the most frequently used functions in a GIS application. They allow the combination of two (or more) spatial data layers comparing them position by position, and treating areas of overlap—and of non-overlap—in distinct ways. Many GISs support overlays through an algebraic language, expressing an overlay function as a formula in which the data layers are the arguments. In this way, we can find

- The cotton fields on black soils (select the 'cotton' cover in the crop data layer and the 'black' cover in the soil data layer and perform an intersection),
- The fields where cotton or jowar is the crop (select both areas of 'cotton' and 'jowar' cover in the crop data layer and take their union),

- The cotton fields not on red soils (perform a difference operator of areas with 'cotton' cover with the areas having red soil),
- The fields that do not have wheat as crop (take the complement of the wheat areas).

3. Neighborhood functions

Whereas overlays combine features at the same location, neighborhood functions evaluate the characteristics of an area surrounding a feature's location. A neighborhood function 'scans' the neighborhood of the given feature(s), and performs a computation on it.

- Search functions allow the retrieval of features that fall within a given search window. This window may be a rectangle, circle, or polygon.
- Buffer zone generation (or buffering) is one of the best known neighborhood functions. It determines a spatial envelope (buffer) around (a) given feature(s). The created buffer may have a fixed width, or a variable width that depends on characteristics of the area.
- Interpolation functions predict unknown values using the known values at nearby locations. This typically occurs for continuous fields, like elevation, when the data actually stored does not provide the direct answer for the location(s) of interest.
- Topographic functions determine characteristics of an area by looking at the immediate neighborhood as well. Typical examples are slope computations on digital terrain models (i.e. continuous spatial fields). The slope in a location is defined as the plane tangent to the topography in that location. Various computations can be performed, such as determination of slope angle, slope aspect, slope length, contour lines.

These are lines that connect points with the same value (for elevation, depth, temperature, barometric pressure, water salinity etc).

4. Connectivity functions

These functions work on the basis of networks, including road networks, water courses in coastal zones, and communication lines in mobile telephony. These networks represent spatial linkages between features. Main functions of this type include:

- Contiguity functions evaluate a characteristic of a set of connected spatial units. One can think of the search for a contiguous area of forest of certain size and shape in a satellite image.
- Network analytic functions are used to compute over connected line features that make up a network. The network may consist of roads, public transport routes, high voltage lines or other forms of transportation infrastructure. Analysis of such networks may entail shortest path computations (in terms of distance or travel time) between two points in a network for routing purposes. Other forms are to find all points reachable within a given distance or duration from a start point for allocation purposes, or determination of the capacity of the network for transportation between an indicated source location and sink location.
- Visibility functions also fit in this list as they are used to compute the points visible from a given location (viewshed modelling or viewshed mapping) using a digital terrain model.

5.2 RETRIEVAL, CLASSIFICATION AND MEASUREMENT

- **Measurement**

Geometric measurement on spatial features includes counting, distance and area size computations. For the sake of simplicity, this section discusses such measurements in a planar spatial reference system. We limit ourselves to geometric measurements, and do not include attribute data measurement. Measurements on vector data are more advanced, thus, also more complex, than those on raster data.

Measurements on vector data

The primitives of vector data sets are point, (poly)line and polygon. Related geometric measurements are location, length, distance and area size. Some of these are geometric properties of a feature in isolation (location, length, area size); others (distance) require two features to be identified. The location property of a vector feature is always stored by the GIS: a single coordinate pair for a point, or a list of pairs for a polyline or polygon boundary. Occasionally, there is a need to obtain the location of the centroid of a polygon; some GISs store these also, others compute them 'on-the-fly'.

Length is a geometric property associated with polylines, by themselves, or in their function as polygon boundary. It can obviously be computed by the GIS as the sum of lengths of the constituent line segments—but it quite often is also stored with the polyline.

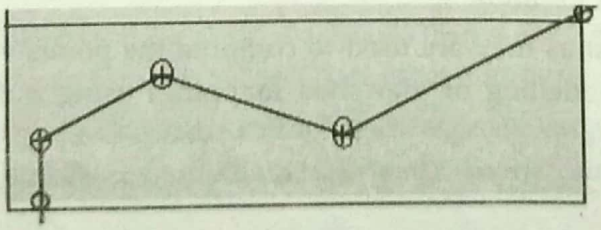
Area size is associated with polygon features. Again, it can be computed, but usually is stored with the polygon as an extra attribute value. This speeds up the computation of other functions that require area size values. The attentive reader will have noted that all of the above 'measurements' do not actually require computation, but only retrieval of stored data. Measuring distance between two features is another important function. If both features are points, say p and q, the computation in a Cartesian spatial reference system are given by the well-known Pythagorean distance function :

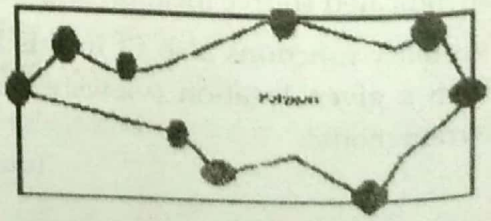
$$d = \sqrt{(x_p - x_q)2 + (y_p - y_q)2}$$

If one of the features is not a point, or both are not, we must be precise in defining what we mean by their distance. All these cases can be summarized as computation of the minimal distance between a location occupied by the first and a location occupied by the second feature. Features that intersect or meet, or when one contains the other have a distance of 0.

It is not possible to store all distance values for all possible combinations of two features in any reasonably sized spatial database. As a result, the system must compute 'on the fly' whenever a distance computation request is made.

Another geometric measurement used by the GIS is the minimal bounding box computation. It applies to polylines and polygons, and determines the minimal rectangle—with sides parallel to the axes of the spatial reference system—that covers the feature.





Bounding box computation is an important function for the GIS: for instance, if the bounding boxes of two polygons do not overlap, we know the polygons cannot possibly intersect each other. Since polygon intersection is a complicated function, but bounding box computation is not, the GIS will always first apply as a test to see whether it must do the first.

A common use of area size measurements is when one wants to sum up the area sizes of all polygons belonging to some class. This class could be crop type:

What is the size of the area covered by cotton? If our crop classification is in a stored data layer, the computation would include (a) selecting the cotton areas, and (b) summing up their area sizes.

Measurements on raster data

Measurements on raster data layers are simpler because of the regularity of the cells. The area size of a cell is constant, and is determined by the cell resolution. Horizontal and vertical resolution may differ, but typically do not. Together with the location of a so-called anchor point, this is the only geometric information stored with the raster data, so all other measurements by the GIS are computed. The anchor point is fixed by convention to be the lower left (or sometimes upper left) location of the raster.

Location of an individual cell derives from the raster's anchor point, the cell resolution, and the position of the cell in the raster.

Again, there are two conventions: the cell's location can be its lower left corner, or the cell's midpoint. These conventions are set by the software in use, and in case of low resolution data they become more important to be aware of.

The area size of a selected part of the raster (a group of cells) is calculated as the number of cells multiplied by the cell area size.

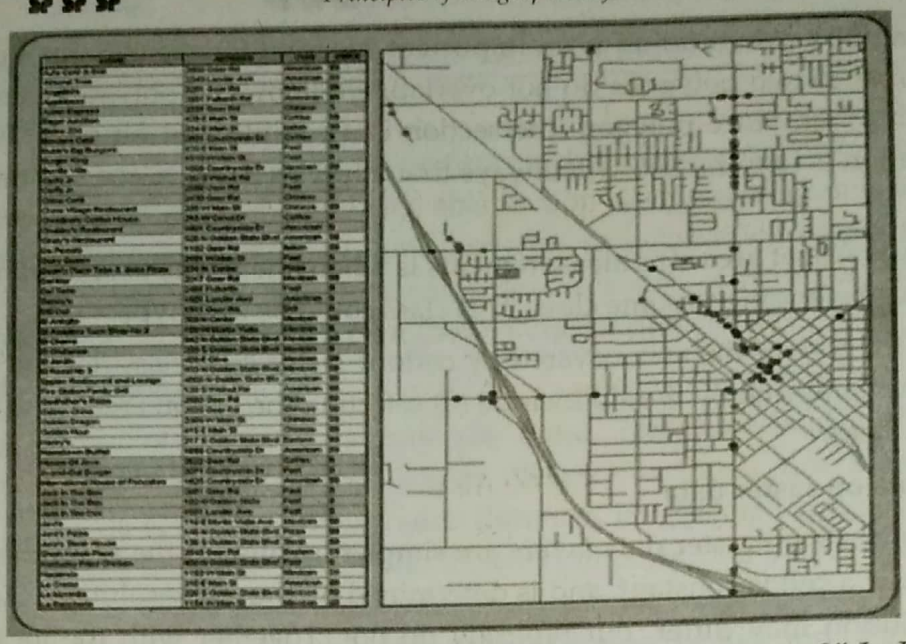
The distance between two raster cells is the standard distance function applied to the locations of their respective mid-points, obviously taking into account the cell resolution. Where a raster is used to represent line features as strings of cells through the raster, the length of a line feature is computed as the sum of distances between consecutive cells. This computation is prone to error

- **Spatial Selection Queries**

When exploring a spatial data set, the first thing one usually wants is to select certain features, to (temporarily) restrict the exploration. Such selections can be made on geometric/spatial grounds, or on the basis of attribute data associated with the spatial features.

Interactive spatial selection

In interactive spatial selection, one defines the selection condition by pointing at or drawing spatial objects on the screen display, after having indicated the spatial data layer(s) from which to select features. The interactively defined objects are called the selection objects; they can be points, lines, or polygons. The GIS then selects the features in the indicated data layer(s) that overlap i.e. intersect, meet, contain, or are contained in; as show in Figure below with the selection objects.



Interactive spatial selection answers questions like "What is at . . . ?" In Figure above, the selection object is a circle and the selected objects are the black and orange dots; they overlap with the selection object. All city regions that overlap with the selection object indicated using dots, and their corresponding attribute records are highlighted in the data table.

Spatial selection by attribute conditions

It is also possible to select features by using selection conditions on feature attributes. These conditions are formulated in SQL if the attribute data reside in a geodatabase. This type of selection answers questions like "where are the features with . . . ?".

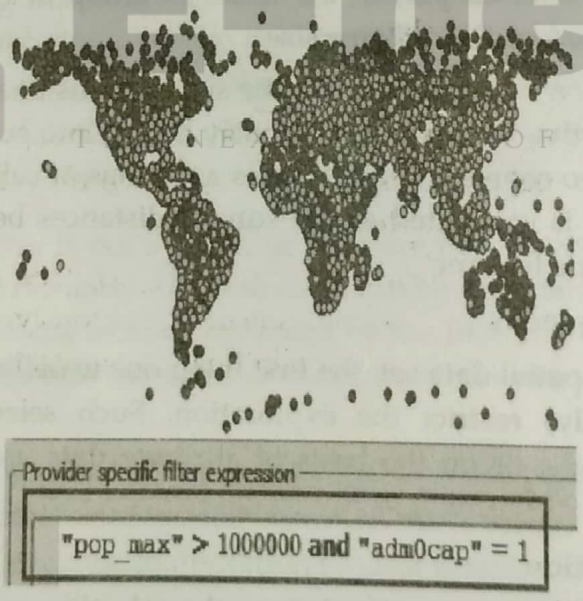


Figure above shows an example of selection by attribute condition. The query expression is pop_max > 1000000, which can be interpreted as "select all the cities of which the population is more than 10, 00, 000." The circles in yellow are the selected cities;

Combining attribute conditions

When multiple criteria have to be used for selection, we need to carefully express all of these in a single composite condition. The tools for this come from a field of mathematical logic, known as propositional calculus.

Atomic conditions such as Area < 400000, and LandUse = 80. Atomic conditions use a

predicate symbol, such as < (less than) or = (equals). Other possibilities are <= (less than or equal), > (greater than), >= (greater than or equal) and <> (does not equal). Any of these symbols is combined with an expression on the left and one on the right. For instance, LandUse <> 80 can be used to select all areas with a land use class different from 80. Expressions are either constants like 400000 and 80, attribute names like Area and LandUse, or possibly composite arithmetic expressions like 0.15 × Area, which would compute 15% of the area size.

Atomic conditions can be combined into composite conditions using logical connectives. The most important ones are AND, OR, NOT and the bracket pair (· · ·). If we write a composite condition like

Area < 400000 AND LandUse = 80,

We can use it to select areas for which both atomic conditions hold true. This is the meaning of the AND connective. If we had written

Area < 400000 OR LandUse = 80 instead, the condition would have selected areas for which either condition holds, so effectively those with an area size less than 400, 000, but also those with land use class 80.

The NOT connective can be used to negate a condition. For instance, the Condition NOT (LandUse = 80) would select all areas with a different land use class than 80. Brackets can be applied to force grouping amongst atomic parts of a composite condition. For instance, the condition (Area < 30000 AND LandUse = 70) OR (Area < 400000 AND LandUse = 80) will select areas of class 70 less than 30, 000 in size, as well as class 80 areas less than 400, 000 in size.

Spatial selection using topological relationships

Various forms of topological relationship can be useful to select features as well. The steps carried out are:

1. To select one or more features as the selection objects, and
2. To apply a chosen spatial relationship function to determine the selected features that have that relationship with the selection objects.

Selecting features that are inside selection objects This type of query uses the containment relationship between spatial objects. Obviously, polygons can contain polygons, lines or points, and lines can contain lines or points, but no other containment relationships are possible.

Here, we are interested in finding the location of medical clinics in the area of Thane District. We first selected all areas of Thane District, using the technique of selection by attribute condition District = "Thane". Then, these selected areas were used as selection objects to determine which medical clinics (as point objects) were within them.

Selecting features that intersect The intersect operator identifies features that are not disjoint to include points and lines.

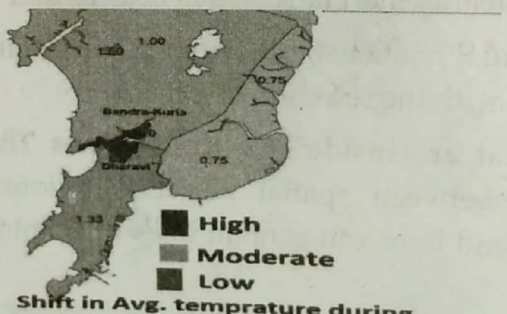
Selecting features adjacent to selection objects Adjacency is the meet relationship. It expresses that features share boundaries, and therefore it applies only to line and polygon features. If we want to select all parcels adjacent to an industrial area. The first step is to select that area (in dark green) and then apply the adjacency function to select all land use areas (in red) that are adjacent to it.

Selecting features based on their distance One may also want to use the distance function of the GIS as a tool in selecting features. Such selections can be searches within a given distance from the selection objects, at a given distance, or even beyond a given distance.

Afterthought on selecting features The selection conditions on attribute values can be combined using logical connectives like AND,OR and NOT. A fact is that the other techniques of selecting features can usually also be combined. Any set of selected features can be used as the input for a subsequent selection procedure. This means, for instance, that we can select all medical clinics first, then identify the roads within 200 meters, then select from them only the major roads, then select the nearest clinics to these remaining roads, as the ones that should receive our financial support. In this way, we are combining various techniques of selection.

- **Classification**

Classification is a technique of purposefully removing detail from an input data set, in the hope of revealing important patterns (of spatial distribution). In the process, we produce an output data set, so that the input set can be left intact. We do so by assigning a characteristic value to each element in the input set, which is usually a collection of spatial features that can be raster cells or points, lines or polygons. If the number of characteristic values is small in comparison to the size of the input set, we have classified the input set. The pattern that we look for may be the distribution of household income in a city. Temperature Shift is called the classification parameter. If we know for each ward in the city the associated average recorded temperature, will have many different values. It can be defined in three different categories (or: classes): 'low', 'Moderate' and 'high', and provide value ranges for each category. If these three categories are mapped in a sensible color scheme, this may reveal interesting information. This has been done for Shift in Average temperature recorded during December between 2008 and 2018 in Figure below.



Shift in Avg. temprature during month of December from 2008 - 2018

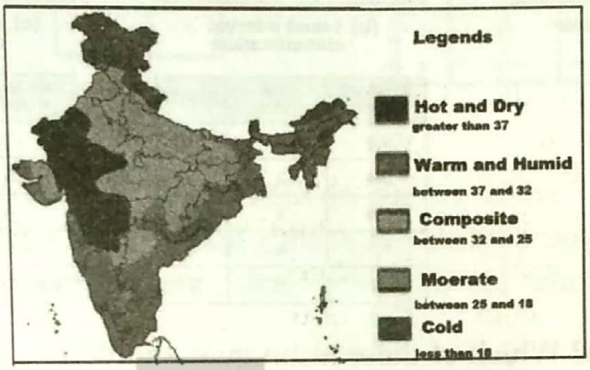
The input data set may have itself been the result of a classification, and in such a case we call it a reclassification. For example, we may have a soil map that shows different soil type units and we would like to show the suitability of units for a specific crop. In this case, it is better to assign to the soil units an attribute of suitability for the crop. Since different soil types may have the same crop suitability, a classification may merge soil units of different type into the same category of crop suitability.

In classification of vector data, there are two possible results. In the first, the input features may become the output features in a new data layer, with an additional category assigned. In other words, nothing changes with respect to the spatial extents of the original features. A second type of output is obtained when adjacent features with the same category are merged into one bigger feature. Such post-processing functions are

called spatial merging, aggregation or dissolving. An illustration of this second type is found in Figure for Dharavi region and Bandra Kurla Complex. This type of merging is only an option in vector data, as merging cells in an output raster on the basis of a classification makes little sense. Vector data classification can be performed on point sets, line sets or polygon sets; the optional merge phase is sensible only for lines and polygons.

User-controlled classification

In user-controlled classification, a user selects the attribute(s) that will be used as the classification parameter(s) and defines the classification method. The latter involves declaring the number of classes as well as the correspondence between the old attribute values and the new classes. This is usually done via a classification table.



It is rather typical for cases in which the used parameter domain is continuous (as in temperature, humidity etc.), then, the table indicates value ranges to be mapped to the same category. Another case exists when the classification parameter is nominal or at least discrete. We must also define the data format of the output, as a spatial data layer, which will contain the new classification attribute. The data type of this attribute is always categorical, i.e. integer or string, no matter what is the data type of the attribute(s) from which the classification was obtained.

Sometimes, one may want to perform classification only on a selection of features. In such cases, there are two options for the features that are not selected. One option is to keep their original values, while the other is to assign a null value to them in the output data set. A null value is a special value that means that no applicable value is present. Care must be taken to deal with these values correctly, both in computation and in visualization.

Automatic classification

User-controlled classifications require a classification table or user interaction. GIS software can also perform automatic classification, in which a user only specifies the number of classes in the output data set. The system automatically determines the class break points. Two main techniques of determining break points are in use.

1. **Equal interval technique**

The minimum and maximum values vmin and vmax of the classification parameter are determined and the (constant) interval size for each category is calculated as (vmax - vmin)/n, where n is the number of classes chosen by the user. This classification is useful in revealing the distribution patterns as it determines the number of features in each category.

2. Equal frequency technique

This technique is also known as quantile classification. The objective is to create categories with roughly equal numbers of features per category. The total number of features is determined first and by the required number of categories, the number of features per category is calculated. The class break points are then determined by counting off the features in order of classification parameter value.

1	1	1	2	8
4	4	5	4	9
4	3	3	2	10
4	5	6	8	8
4	2	1	1	1

(a) original raster

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

(b) equal interval classification

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

(c) equal frequency classification

original value	new value	# cells
1,2	1	9
3,4	2	8
5,6	3	3
7,8	4	3
9,10	5	2

original value	new value	# cells
1	1	6
2,3	2	5
4	3	6
5,6	4	3
8,9,10	5	5

When to use which? Which of these techniques should be applied to a given dataset depends upon the purpose of the analysis (what the user is trying to achieve) as well as the characteristics of the data itself.

6.3 OVERLAY FUNCTIONS

Overlay is a technique of combining two spatial data layers and producing a third from them. The binary operators that we discuss are known as spatial overlay operators. We will firstly discuss vector overlay operators, and then focus on the raster case. Standard overlay operators take two input data layers, and assume they are georeferenced in the same system, and overlap in study area. If either of these requirements is not met, the use of an overlay operator is senseless. The principle of spatial overlay is to compare the characteristics of the same location in both data layers, and to produce a result for each location in the output data layer. The specific result to produce is determined by the user. It might involve a calculation, or some other logical function to be applied to every area or location.

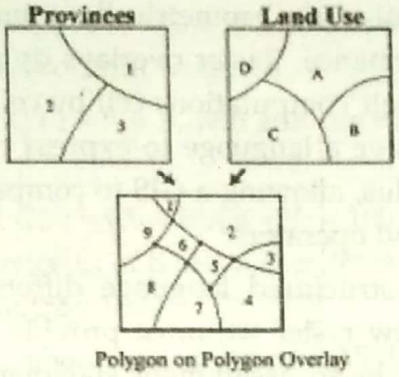
In raster data, as we shall see, these comparisons are carried out between pairs of cells, one from each input raster. In vector data, the same principle of comparing locations applies, but the underlying computations rely on determining the spatial intersections of features from each input layer.

- **Vector Overlay Operators**

In the vector domain, overlay is computationally more demanding than in the raster domain. Here we will only discuss overlays from polygon data layers, but we note that most of the ideas also apply to overlay operations with point or line data layers.

The standard overlay operator for two layers of polygons is the polygon intersection operator. It is fundamental, as many other overlay operators proposed in the literature or implemented in systems can be defined in terms of it. The result of this operator is the

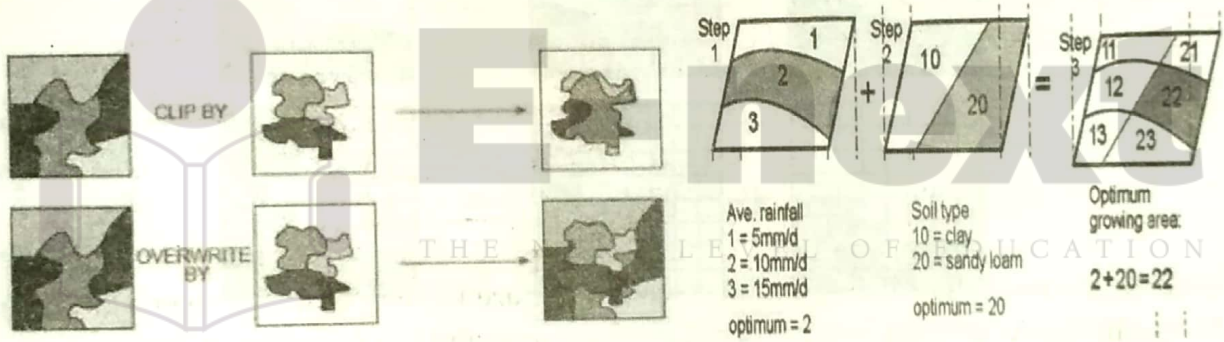
collection of all possible polygon intersections; the attribute table result is a join—in the relational database of the two input attribute tables. This output attribute table only contains one table for each intersection polygon found, and this explains why we call this operator a spatial join.



Polygon Topology

ID	Provinces	Land Use
1	1	D
2	1	A
3	1	B
4	3	B
5	3	A
6	2	A
7	3	C
8	2	C
9	2	D

A more practical example is provided in Figure above, which was produced by polygon intersection of the rainfall polygons with soil type polygons classified. This has allowed select the areas with optimum soil with optimum rain fall for growing a specific type of crop.



Two more polygon overlay operators are illustrated in the above figure. The first is known as the polygon clipping operator. It takes a polygon data layer and restricts its spatial extent to the generalized outer boundary obtained from all (selected) polygons in a second input layer. Besides this generalized outer boundary, no other polygon boundaries from the second layer play a role in the result.

A second overlay operator is polygon overwrite. The result of this binary operator is defined is a polygon layer with the polygons of the first layer, except where polygons existed in the second layer, as these take priority.

Most GISs do not force the user to apply overlay operators to the full polygon data set. One is allowed to first select relevant polygons in the data layer, and then use the selected set of polygons as an operator argument.

The fundamental operator of all these is polygon intersection. The others can be defined in terms of it, usually in combination with polygon selection and/or classification. For instance, the polygon overwrite of A by B can be defined as polygon intersection between A and B, followed by a (well-chosen) classification that prioritizes polygons in B, followed by a merge.

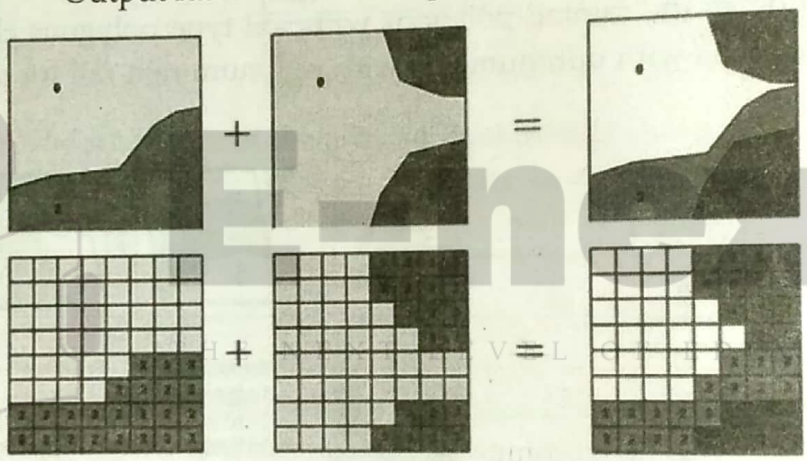
Vector overlays are usually also defined for point or line data layers. Their definition parallels the definitions of operators discussed above. Different GISs use different names for these operators, and one is advised to carefully check the documentation before applying any of these operators.

- **Raster Overlay Operators**

Vector overlay operators are useful, but geometrically complicated, and this sometimes results in poor operator performance. Raster overlays do not suffer from this disadvantage, as most of them perform their computations cell by cell, and thus they are fast. GIS that support raster processing have a language to express operations on raster referred to as map algebra, or raster calculus, allowing a GIS to compute new raster from existing ones, using a range of functions and operators.

The key operations using a logical structured language differs for different GIS software packages. When producing a new raster we must provide a name for it, and define how it is computed. This is done in an assignment statement of the following format:

Output raster name := Map algebra expression.

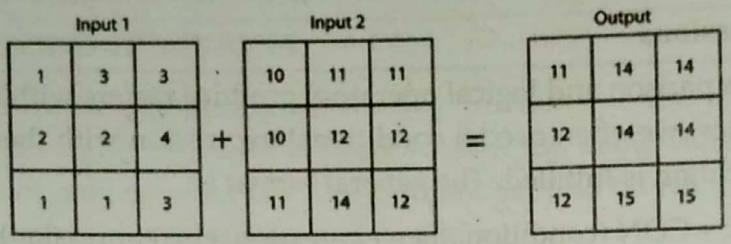


The expression on the right is evaluated by the GIS, and the raster in which it results is then stored under the name on the left. The expression may contain references to existing rasters, operators and functions; the format is made clear below. The raster names and constants that are used in the expression are called its operands. When the expression is evaluated, the GIS will perform the calculation on a pixel by pixel basis, starting from the first pixel in the first row, and continuing until the last pixel in the last row.

Arithmetic operators

Various arithmetic operators are supported. The standard ones are multiplication (×), division (/), subtraction (-) and addition (+). Other arithmetic operators may include modulo division (MOD) and integer division (DIV). Modulo division returns the remainder of division: for example, 11 MOD 5 will return 1 as $10 - 5 \times 2 = 1$. Similarly, 10 DIV 2 will return 5.

More operators are goniometric: sine (sin), cosine (cos), tangent (tan), and their inverse functions asin, acos, and atan, which return radian angles as real values. A simple map algebra assignment is illustrated in Figure below.

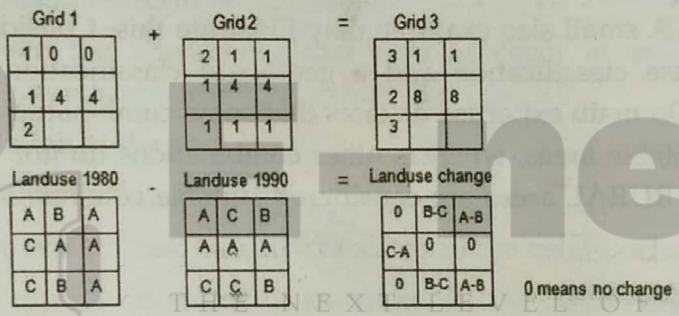


The assignment : C1 := A + 5, will add a constant factor of 5 to all cell values of raster A and store the result as output raster C1.

The assignment C2 : = A + B, will add the values of A and B cell by cell, and store the result as raster C2.

The assignment C3 : = (A + B)/(A - B) × 100 will create output raster C3, as the result of the addition (cell by cell,) of B cell values from A cell values, divided by their difference. The result is multi plied by 100.

This expression, when carried out on AVHRR channel 1 (red) and AVHRR channel 2 (near infrared) of NOAA satellite imagery, is known as the NDVI (Normalized Difference Vegetation Index). It has proven to be a good indicator of the presence of green vegetation.

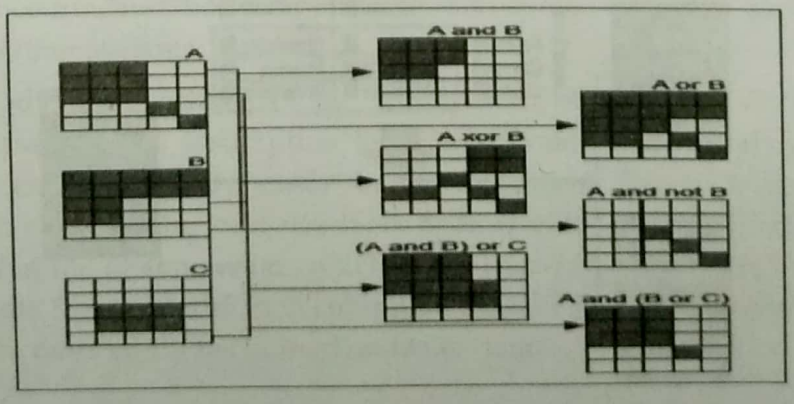


Comparison and logical operators

Map algebra also allows the comparison of rasters cell by cell. To this end, we may use the standard comparison operators (<, <=, =, >=, > and <>) that we introduced before. A simple raster comparison assignment is: C: = A <> B, will store truth value either true orfalse in the output raster C.

Logical connectives like AND, OR, XOR, NOT are also supported in map algebra.

The figure below provides various raster computations in search of black soil on different rain fall area. Raster A indicates black soil with rainfall below 5 mm, B indicates red soil with rain fall between 5 mm and 10 mm and raster C areas that are having rainfall above 10 mm and having clay soil.



Conditional expressions

The above comparison and logical operators produce rasters with the truth value true and false. In practice, we often need a conditional expression with them that allows us to test whether a condition is fulfilled. The general format is:

Output raster : = CON (condition, then expression, else expression).

Here, condition is the tested condition, then expression is evaluated if condition holds, and else expression is evaluated if it does not hold.

For example an expression like CON (GridIn>3, 1, 0) will evaluate to 1 for each cell in the output raster where the same cell in GridIn is classified as greater than 3. In each cell where this is not true, the else expression is evaluated, resulting in 0.

5	3	6
2	4	4
2	1	6

GridIn

1	0	1
0	1	1
0	0	1

GridOut

- **Overlays Using a Decision Table**

Conditional expressions are powerful tools in cases where multiple criteria must be taken into account. A small size example may illustrate this. Consider a suitability study in which a land use classification and a geological classification must be used. The respective rasters. Do main expertise dictates that some combinations of land use and are a Type result in suitable areas, whereas other combinations do not. In our example, NA Land on CITY and RURAL areas are considered suitable combinations, while the others are not.

A map algebra expression

Suitability : = CON ((Landuse = "Non-Agriculture" AND areaType = "CITY") OR

(Landuse = "Non-Agriculture" AND areaType = "RURAL"), "Suitable", "Unsuitable")

The above type of computation becomes simpler by setting up a separate decision table that will guide the raster overlay process. This extra table carries domain expertise, and dictates which combinations of input raster cell values will produce which output raster cell value. This gives us a raster overlay operator using a decision table, as illustrated in Figure below. The GIS will have supporting functions to generate the additional table from the input rasters, and to enter appropriate values in the table.

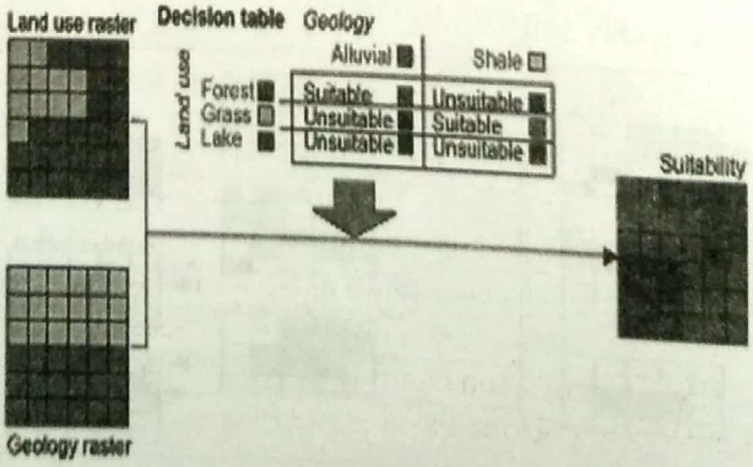


Figure : Classification using Decision Table

6.4 NEIGHBORHOOD FUNCTIONS

The principle in Neighborhood function is to find out the characteristics of the vicinity, here called neighborhood, of a location. After all, many suitability questions, for instance, depend not only on what is at the location, but also on what is near the location. Thus, the GIS must allow us 'to look around locally'.

To perform neighborhood analysis, we must :

1. State which target locations are of interest to us, and define their spatial extent,
2. Define how to determine the neighborhood for each target,
3. Define which characteristic(s) must be computed for each neighborhood.

For instance, our target might be a nearby ATM. Its neighborhood could be defined as:

- An area within 100m walking distance of an State Bank ATM, or
- An area within 2 km travel distance, or
- All roads within 500 m travel distance, or
- All other Bank ATM within 5 minutes travel time, or
- All Banks, for which the ATM is the closest.

To discover about the phenomena that exist or occur in the neighborhood. E. g. spatial extent, also require statistical information like:

- The total population of the area,
- Average household income, or
- The distribution of high-risk industries located in the neighborhood.

The above are typical questions in an urban setting. When our interest is more in natural phenomena, different examples of locations, neighborhoods and neighborhood characteristics arise. Since raster data are the more commonly used in this case, neighborhood characteristics often are obtained via statistical summary functions that compute values such as average, minimum, maximum, and standard deviation of the cells in the identified neighborhood.

Determining neighborhood extent

To select target locations, one can use the selection techniques. To obtain characteristics from an eventually identified neighborhood, the same techniques apply. So what remains to be discussed here is the proper determination of a neighborhood. One way of determining a neighborhood around a target location is by making use of the geometric distance function. For example, pollution spread by rivers, ground water flow, or prevailing weather systems.

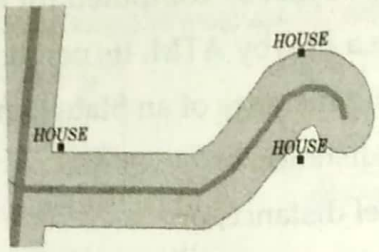
The more advanced techniques for computation of flow and diffusion. Diffusion functions are based on the assumption that the phenomenon spreads in all directions, though not necessarily equally easily in all directions. Hence, it uses local terrain characteristics to compute the local resistance against diffusion. In flow computations, the assumption is that the phenomenon will choose a least-resistance path, and not spread in all directions. Both flow and diffusion computations take local characteristics into account, and are therefore more easily performed on raster data.

- **Proximity Computations**

In proximity computations, we use geometric distance to define the neighborhood of one or more target locations. The most common and useful technique is buffer zone generation.

Buffer zone generation

The principle of buffer zone generation is simple : we select one or more target locations, and then determine the area around them, within a certain distance. In Figure below, the main roads were selected as targets, and a 75 meter buffer was computed from them.

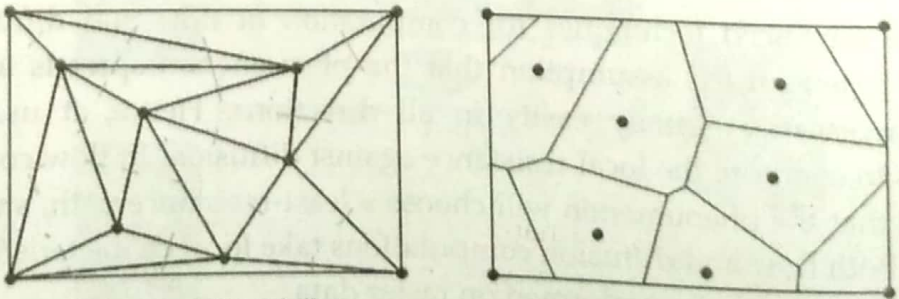


In vector-based buffer generation, the buffers themselves become polygon features, usually in a separate data layer, that can be used in further spatial analysis. Buffer generation on rasters is a fairly simple function. The target location or locations are always represented by a selection of the raster's cells, and geometric distance is defined, using cell resolution as the unit. The distance function applied is the Pythagorean distance between the cell centers. The distance from a non-target cell to the target is the minimal distance one can find between that non-target cell and any target cell.

Thiessen polygon generation

Thiessen polygon partitions make use of geometric distance for determining neighbourhoods. This is useful if we have a spatially distributed set of points as target locations, and we want to know for each location in the study to which target it is closest. This technique will generate a polygon around each target location that identifies all those locations that 'belong to' that target. We have already seen the use of Thiessen polygons in the context of interpolation of point data. Given an input point set that will be the polygon's midpoints, it is not difficult to construct such a partition. It is even much easier to construct if we already have a Delaunay triangulation for the same input point set.

Figure below repeats the Delaunay triangulation of the Thiesse polygon partition constructed from it is on the right. The construction first creates the perpendiculars of all the triangle sides; observe that a perpendicular of a triangle side that connects two points is the divide between the areas closer to both points. The perpendiculars become part of the boundary of each Thiessen polygon.



Delaunay triangulation

Thiessen polygons

Computation of Diffusion

The determination of neighborhood of one or more target locations may depend not only on distance—cases which we discussed above—but also on direction and differences in the terrain in different directions. This typically is the case when the target location contains a 'source material' that spreads over time, referred to as diffusion. This 'source material' may be air, water or soil pollution, commuters exiting a train station, people from an opened-up refugee camp, a water spring uphill, or the radio waves emitted from a radio relay station. In all these cases, one will not expect the spread to occur evenly in all directions. There will be local terrain factors that influence the spread, making it easier or more difficult.

Diffusion computation involves one or more target locations, which are better called source locations in this context. They are the locations of the source of whatever spreads. The computation also involves a local resistance raster, which for each cell provides a value that indicates how difficult it is for the 'source material' to pass by that cell. The value in the cell must be normalized: i.e. valid for a standardized length (usually the cell's width) of spread path. From the source location(s) and the local resistance raster, the GIS will be able to compute a new raster that indicates how much minimal total resistance the spread has witnessed for reaching a raster cell. This process is illustrated in Figure below.

1	1	1	2	8
4	4	5	4	9
4	3	3	2	10
4	5	6	8	8
4	2	1	1	1

(a)

14.50	14.95	15.95	17.45	22.45
12.00	12.45	14.61	16.66	21.44
8.00	8.95	11.95	13.66	19.66
4.00	6.36	8.00	10.00	11.00
0.00	3.00	4.50	5.50	6.50

(b)

While computing total resistances, the GIS take proper care of the path lengths. Obviously, the diffusion from a cell csrc to its neighbor cell to the east ce is shorter than to the cell that is its northeast neighbor cne.

The distance ratio between these two cases is 1: $\sqrt{2}$, If val(c) indicates the local resistance value for cell c, the GIS computes the total incurred resistance for diffusion from csrc to ce as

$$\frac{1}{2}(\text{val}(C_{src}) + \text{val}(C_e))$$

while the same for csrc to cne is

$$\frac{1}{2}(\text{val}(C_{src}) + \text{val}(C_{ne})) \times \sqrt{2}$$

The accumulated resistance along a path of cells is simply the sum of these incurred resistances from pair-wise neighbor cells.

Since 'source material' has the habit of taking the easiest route to spread, we must determine at what minimal cost (i.e. at what minimal resistance) it may have arrived in a cell. Therefore, we are interested in the minimal cost path. To determine the minimal total resistance along a path from the source location csrc to an arbitrary cell cx, the GIS

determines all possible paths from csrc to cx, and then determines which one has the lowest total resistance. This value is found, for each cell, in the raster of Figure (b) above.

For instance, there are three paths from the green source location to its northeast neighbor cell (with local resistance 5). We can define them as path 1 (N–E), path 2 (E–N) and path 3 (NE), using compass directions to define the path from the green cell. For path 1, the total resistance is computed as:

$$\frac{1}{2}(4+4)+\frac{1}{2}(4+5)=8.5$$

Path 2, in similar style, gives us a total value of 6.5. For path 3, we find

$\frac{1}{2}(4+5)\times\sqrt{2}=6.36$ is the minimal cost path.

- **Flow Computation**

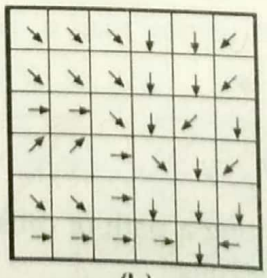
Flow computations determine how a phenomenon spreads over the area, in principle in all directions, though with varying difficulty or resistance. There are also cases where a phenomenon does not spread in all directions, but moves or 'flows' along a given, least-cost path, determined again by local terrain characteristics. The typical case arises when we want to determine the drainage patterns in a catchment: the rainfall water 'chooses' a way to leave the area.

This principle is illustrated with a simple elevation raster, in below Figure (a). For each cell in that raster, the steepest downward slope to a neighbor cell is computed, and its direction is stored in a new raster Figure (b). This computation determines the elevation difference between the cell and a neighbor cell, and takes into account cell distance—1 for neighbor cells in N–S or W–E direction, √2 for cells in NE–SW or NW-SE direction. Among its eight neighbor cells, it picks the one with the steepest path to it. The directions in raster (b), thus obtained, are encoded in integer values, and we have 'decoded' them for the sake of illustration. Raster (b) can be called the flow direction raster. From raster (b), the GIS can compute the accumulated flow count raster, a raster that for each cell indicates how many cells have their water flow into the cell.

Cells with a high accumulated flow count represent areas of concentrated flow, and thus may belong to a stream. By using some appropriately chosen threshold value in a map algebra expression, we may decide whether they do. Cells with an accumulated flow count of zero are local topographic highs, and can be used to identify ridges.

156	144	138	142	116	98
148	134	112	98	92	100
138	106	88	74	76	96
128	116	110	44	62	48
136	122	94	42	32	38
148	106	68	24	22	24

(a)



(b)

0	0	0	0	0	0
0	1	1	2	2	0
0	3	7	5	4	0
0	0	0	20	0	1
0	0	0	1	24	0
0	2	4	7	35	1

(c)

- **Raster Based Surface Analysis**

Continuous fields have a number of characteristics not shared by discrete fields. Since the field changes continuously, we can talk about slope angle, slope aspect and concavity/convexity of the slope. These notions are not applicable to discrete fields.

The discussions here use terrain elevation as the prototypical example of a continuous field, but all issues discussed are equally applicable to other types of continuous fields. Nonetheless, we regularly refer to the continuous field representation as a DEM, to conform with the most common situation.

Applications

There are numerous examples where more advanced computations on continuous field representations are needed. A short list is provided below.

- **Slope angle calculation**

The calculation of the slope steepness, expressed as an angle in degrees or percentages, for any or all locations.

- **Slope aspect calculation**

The calculation of the aspect (or orientation) of the slope in degrees (between 0 and 360 degrees), for any or all locations.

Slope convexity/concavity calculation : Slope convexity—defined as the change of the slope (negative when the slope is concave and positive when the slope is convex)—can be derived as the second derivative of the field.

- **Slope length calculation**

With the use of neighborhood operations, it is possible to calculate for each cell the nearest distance to a watershed boundary (the upslope length) and to the nearest stream (the downslope length). This information is useful for hydrological modelling. Hillshading is used to portray relief difference and terrain morphology in hilly and mountainous areas. The application of a special filter to a DEM produces hillshading. The colour tones in a hillshading raster represent the amount of reflected light in each location, depending on its orientation relative to the illumination source. This illumination source is usually chosen at an angle of 45° above the horizon in the north-west.

- **Three-dimensional map display**

With GIS software, three-dimensional views of a DEM can be constructed, in which the location of the viewer, the angle under which s/he is looking, the zoom angle, and the amplification factor of relief exaggeration can be specified. Three-dimensional views can be constructed using only a predefined mesh, covering the surface, or using other rasters (e.g. a hillshading raster) or images (e.g. satellite images) which are draped over the DEM.

- **Determination of change in elevation through time**

The cut-and-fill volume of soil to be removed or to be brought in to make a site ready for construction can be computed by overlaying the DEM of the site before the work begins with the DEM of the expected modified topography. It is also possible to determine landslide effects by comparing DEMs of before and after the landslide event.

- **Automatic catchment delineation**

Catchment boundaries or drainage lines can be automatically generated from a good quality DEM with the use of neighborhood functions. The system will determine the lowest point in the DEM, which is considered the outlet of the catchment. From there, it will repeatedly search the neighboring pixels with the highest altitude.

This process is continued until the highest location (i.e. cell with highest value) is found, and the path followed determines the catchment boundary. For delineating the

drainage network, the process is reversed. Now, the system will work from the watershed downwards, each time looking for the lowest neighboring cells, which determines the direction of water flow.

- **Dynamic modeling**

DEMs are increasingly used in GIS-based dynamic modeling, such as the computation of surface run-off and erosion, groundwater flow, the delineation of areas affected by pollution, the computation of areas that will be covered by processes such as debris flows and lava flows.

- **Visibility analysis**

A viewshed is the area that can be 'seen', i.e. in the direct line-of-sight from a specified target location. Visibility analysis determines the area visible from a scenic lookout, the area that can be reached by a radar antenna, or assesses how effectively a road or quarry will be hidden from view.

Filtering

The principle of filtering is quite similar to that of moving window averaging. We define a window and let the GIS move it over the raster cell-by-cell. For each cell, the system performs some computation, and assigns the result of this computation to the cell in the output raster. The difference with moving window averaging is that the moving window in filtering is itself a little raster, which contains cell values that are used in the computation for the output cell value. This little raster is a filter, also known as a kernel which may be square (such as a 3x3 kernel), but it does not have to be. The values in the filter are used as weight factors.

As an example, let us consider a 3 × 3 cell filter, in which all values are equal to 1, as illustrated in below Figure. The use of this filter means that the nine cells considered are given equal weight in the computation of the filtering step. Let the input raster cell values, for the current filtering step, be denoted by rij and the corresponding filter values by wij. The output value for the cell under consideration will be computed as the sum of the weighted input values divided by the sum of weights:

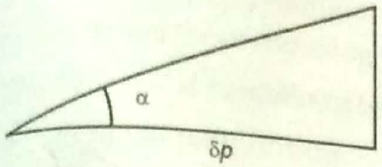
$$\sum_{i,j} (w_{ij} \cdot r_{ij}) / \sum_{i,j} |w_{ij}|$$

1	1	1
1	1	1
1	1	1

Computation of slope angle and slope aspect

A different choice of weight factors may provide other information. Special filters exist to perform computations on the slope of the terrain.

Slope angle, which is also known as slope gradient, is the angle α, between a path p in the horizontal plane and the sloping terrain. The path p must be chosen such that the angle α is maximal. A slope angle can be expressed as elevation gain in a percentage or as a geometric angle, in degrees or radians. The two respective formulas are:



$$\text{Slope perc} = 100.\frac{\delta f}{\delta p} \text{ and slope\_angle} = \arctan\left(\frac{df}{dp}\right)$$

The path p must be chosen to provide the highest slope angle value, and thus it can lie in any direction. The compass direction, converted to an angle with the North, of this maximal down-slope path p is what we call the slope aspect. From an elevation raster, we cannot 'read' the slope angle or slope aspect directly. Yet, that information can be extracted. After all, for an arbitrary cell, we have its elevation value, plus those of its eight neighbor cells. A simple approach to slope angle computation is to make use of x-gradient and y-gradient filters. Figure (a) and (b) illustrate an x-gradient filter, and y-gradient filter, respectively. The x-gradient filter determines the slope increase ratio from west to east: if the elevation to the west of the centre cell is 1540 m and that to the east of the centre cell is 1552 m, then apparently along this transect the elevation increases 12 m per two cell widths, i.e. the x-gradient is 6 m per cell width. The y-gradient filter operates entirely analogously, though in south-north direction.

Both filters express elevation gain per cell width. This means that we must divide by the cell width—given in meters, for example, to obtain the true derivatives $\delta f/\delta x$ and $\delta f/\delta y$. Here, f stands for the elevation field as a function of x and y, and $\delta f/\delta x$, for instance, is the elevation gain per unit of length in the x-direction.

To obtain the real slope angle α along path p, observe that both the x and y gradient contribute to it. As shown in figure C,

A, geometric derivation show that

$$\text{Tan}(\alpha) = \sqrt{(\delta f/\delta x)^2 + (\delta f/\delta y)^2}$$

Now what does this mean in the practice of computing local slope angles from an elevation raster? It means that we must perform the following steps:

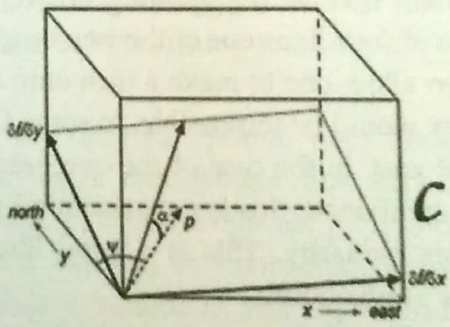
1. Compute from (input) elevation raster R the non-normalized x- and y-gradients, using the filters of Figure (a) and (b), respectively.

0	0	0
-1	0	1
0	0	0

(a)

(b)

0	1	0
0	0	0
0	-1	0



2. Normalize the resulting rasters by dividing by the cell width, expressed in units of length like meters.
3. Use both rasters for generating a third raster, applying the formula, possibly even applying an arctan function to the result to obtain the slope angle α for each cell.

It can also be shown that for the slope aspect ψ we have

$$\text{Tan}(\psi) = \frac{\delta f / \delta x}{\delta f / \delta y}$$

So slope aspect can also be computed from the normalized gradients.

$$\psi = \arctan\left(\frac{\delta f / \delta x}{\delta f / \delta y}\right)$$

The reason being that the latter formula does not account for southeast and southwest quadrants, nor for cases where $\delta f/\delta y = 0$.

6.5 NETWORK ANALYSIS

A completely different set of analytical functions in GIS consists of computations on networks. A network is a connected set of lines, representing some geographic phenomenon, typically of the transportation type. The 'goods' transported can be almost anything: people, cars and other vehicles along a road network, commercial goods along a logistic network, phone calls along a telephone network, or water pollution along a stream/river network.

Network analysis can be performed on either raster or vector data layers, but they are more commonly done in the latter, as line features can be associated with a network, and hence can be assigned typical transportation characteristics such as capacity and cost per unit. A fundamental characteristic of any network is whether the network lines are considered directed or not.

Associate with each line a direction of transportation; undirected networks do not.

The 'goods' can be transported along a line in both directions.

For many applications of network analysis, a planar network, i.e. one that can be embedded in a two-dimensional plane, will do the job. Many networks are naturally planar, like stream/river networks. A large-scale traffic network, on the other end, is not planar: motorways have multi-level crossings and are constructed with underpasses and overpasses. Planar networks are easier to deal with computationally, as they have simpler topological rules.

Not all GISs accommodate non-planar networks, or can do so only using 'tricks'. These may involve the splitting of overpassing lines at the intersection vertex and the creation of four lines out of the two original lines. Without further attention, the network will then allow one to make a turn onto another line at this new intersection node, which in reality would be impossible. In some GISs we can allocate a cost with turning at a node and that cost, in the case of the overpass, can be made infinite to ensure it is prohibited. But, as mentioned, this is a workaround to fit a non-planar situation into a data layer that presumes planarity. This is a good illustration of geometry not fully determining the network's behaviour.

Additional application-specific rules are usually required to define what can and cannot happen in the network. Most GISs provide rule-based tools that allow the definition of these extra application rules. Various classical spatial analysis functions on networks are supported by GIS software packages. The most important ones are:

1. Optimal path finding which generates a least cost-path on a network between a pair of predefined locations using both geometric and attribute data.

2. Network partitioning which assigns network elements (nodes or line segments) to different locations using predefined criteria.

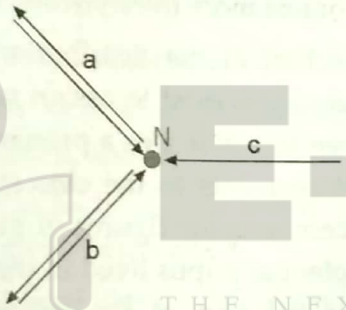
The two typical functions discussed here are.

- **Optimal path finding**
- **Network partitioning**

Optimal path finding

Optimal path finding techniques are used when a least-cost path between two nodes in a network must be found. The two nodes are called origin and destnation, respectively. The aim is to find a sequence of connected lines to traverse from the origin to the destination at the lowest possible cost.

The cost function can be simple: for instance, it can be defined as the total length of all lines on the path. The cost function can also be more elaborate and take into account not only length of the lines, but also their capacity, maximum transmission (travel) rate and other line characteristics, for instance to obtain a reasonable approximation of travel time. There can even be cases in which the nodes visited add to the cost of the path as well. These may be called turning costs, which are defined in a separate turning cost table for each node, indicating the cost of turning at the node when entering from one line and continuing on another. This is illustrated in Figure below.



from	onto	cost
a	a	×
a	b	5
b	a	8
b	b	15
c	a	3
c	b	6

Notice that it is possible to travel on line b in Figure above, then take a U-turn at node N, and return along a to where one came from. The question is whether doing this makes sense in optimal path finding. After all, to go back to where one comes from will only increase the total cost. In fact, there are situations where it is optimal to do so. Suppose it is node M that is connected by line b with node N, and that we actually wanted to travel to another node L from M. The turn at M towards node L coming via another line may be prohibitively expensive, whereas turning towards L at M returning to M along b may not be so expensive.

Problems related to optimal path finding are ordered optimal path finding and unordered optimal path finding. Both have an extra requirement that a number of additional nodes needs to be visited along the path. In ordered optimal path finding, the sequence in which these extra nodes are visited matters; in unordered optimal path finding it does not. Here, a path is found from node A to node D, visiting nodes B and C. Obviously, the length of the path found under non-ordered requirements is at most as long as the one found under ordered requirements. Some GIS provide support for these more complicated path finding problems.

Network partitioning

In network partitioning, the purpose is to assign lines and/or nodes of the network, in a mutually exclusive way, to a number of target locations. Typically, the target locations

play the role of service centre for the network. This may be any type of service: medical treatment, education, water supply. This type of network partitioning is known as a network allocation problem.

Another problem is trace analysis. Here, one wants to determine that part of the network that is upstream (or downstream) from a given target location. Such problems exist in pollution tracing along river/stream systems, but also in network failure chasing in energy distribution networks.

Network allocation

In network allocation, we have a number of target locations that function as resource centres, and the problem is which part of the network to exclusively assign to which service centre. This may sound like a simple allocation problem, in which a service centre is assigned those lines (segments) to which it is nearest, but usually the problem statement is more complicated. These further complications stem from the requirements to take into account

- The capacity with which a centre can produce the resources (whether they are medical operations, school pupil positions, kilowatts, or bottles of milk), and
- The consumption of the resources, which may vary amongst lines or line segments. After all, some streets have more accidents, more children who live there, more industry in high demand of electricity or just more thirsty workers.

The service area of any centre is a subset of the distribution network, in fact, a connected part of the network. Various techniques exist to assign network lines, or their segments, to a centre. In Figure (a), the green star indicates a primary school and the GIS has been used to assign streets and street segments to the closest school within 2 km distance, along the network. Then, using demographic figures of pupils living along the streets, it was determined that too many potential pupils lived in the area for the school's capacity. So in part (b), the already selected part of the network was reduced to accommodate precisely the school's pupil capacity for the next academic year.

Trace analysis

Trace analysis is performed when we want to understand which part of a network is 'conditionally connected' to a chosen node on the network, known as the trace origin. For a node or line to be conditionally connected, it means that a path exists from the node/line to the trace origin, and that the connecting path fulfills the conditions set. What these conditions are depends on the application, and they may involve direction of the path, capacity, length, or resource consumption along it. The condition typically is a logical expression, as we have seen before, for example:

- The path must be directed from the node/line to the trace origin,
- Its capacity (defined as the minimum capacity of the lines that constitute the path) must be above a given threshold, and
- The path's length must not exceed a given maximum length.

Tracing is the computation that the GIS perform to find the paths from the trace origin that obey the tracing conditions. It is a rather useful function for many network-related problems.

6.6 GIS AND APPLICATION MODELS

Models are simplified abstractions of reality representing or describing its most important elements and their interactions. Modelling and GIS are more or less inseparable, as GIS is itself a tool for modelling 'the real world'.

The solution to a (spatial) problem usually depends on a large number of parameters. Since these parameters are often interrelated, their interaction is made more precise in an application model.

Here we define application models to include any kind of GIS based model including analytical and process models for a specific real-world application. Such a model, in one way or other, describes as faithfully as possible how the relevant geographic phenomena behave, and it does so in terms of the parameters.

The nature of application models varies enormously. GIS applications for famine relief programs, for instance, are very different from earthquake risk assessment applications, though both can make use of GIS to derive a solution. Many kinds of application models exist, and they can be classified in many different ways.

Here we identify five characteristics of GIS-based application models :

1. The purpose of the model,
2. The methodology underlying the model,
3. The scale at which the model works,
4. Its dimensionality - i.e. whether the model includes spatial, temporal or spatial and temporal dimensions, and
5. Its implementation logic - i.e. the extent to which the model uses existing knowledge about the implementation context.

It is important to note that the categories above are merely different characteristics of any given application model. Any model can be described according to these characteristics. Each is briefly discussed below.

Purpose of the model refers to whether the model is descriptive, prescriptive or predictive in nature. Descriptive models attempt to answer the "what is" question. Prescriptive models usually answer the "what should be" question by determining the best solution from a given set of conditions.

Models for planning and site selection are usually prescriptive, in that they quantify environmental, economic and social factors to determine 'best' or optimal locations. So-called Predictive models focus upon the "what is likely to be" questions, and predict outcomes based upon a set of input conditions. Examples of predictive models include forecasting models, such as those attempting to predict landslides or sea–level rise.

Methodology refers to the operational components of the model. Stochastic models use statistical or probability functions to represent random or semi-random behaviour of phenomena. In contrast, deterministic models are based upon a well-defined cause and effect relationship. Examples of deterministic models include hydrological flow and pollution models, where the 'effect' can often be described by numerical methods and differential equations.

Rule-based models attempt to model processes by using local (spatial) rules. Cellular Automata (CA) are examples of models in this category. These are often used to

understand systems which are generally not well understood, but for which their local processes are well known. For example, the characteristics of neighborhood cells (such as wind direction and vegetation type) in a raster based CA model might be used to model the direction of spread of a fire over several time steps.

Rule -based models (ABM) attempt to model movement and development of multiple interacting agents (which might represent individuals), often using sets of decision-rules about what the agent can and cannot do. Complex agent-based models have been developed to understand aspects of travel behavior and crowd interactions which also incorporate stochastic components.

Scale refers to whether the components of the model are individual or aggregate in nature. Essentially this refers to the 'level' at which the model operates. Individual-based models are based on individual entities, such as the agent-based models described above, whereas aggregate models deal with 'grouped' data, such as population census data. Aggregate models may operate on data at the level of a city block (for example, using population census data for particular social groups), at the regional, or even at a global scale.

Dimensionality is the term chosen to refer to whether a model is static or dynamic, and spatial or aspatial. Some models are explicitly spatial, meaning they operate in some geographically defined space. Some models are aspatial, meaning they have no direct spatial reference.

Models can also be static, meaning they do not incorporate a notion of time or change. In dynamic models, time is an essential parameter. Dynamic models include various types of models referred to as process models or simulations. These types of models aim to generate future scenarios from existing scenarios, and might include deterministic or stochastic components, or some kind of local rule (for example, to drive a simulation of urban growth and spread). The fire spread example given above is a good example of an explicitly spatial, dynamic model which might incorporate both local rules and stochastic components.

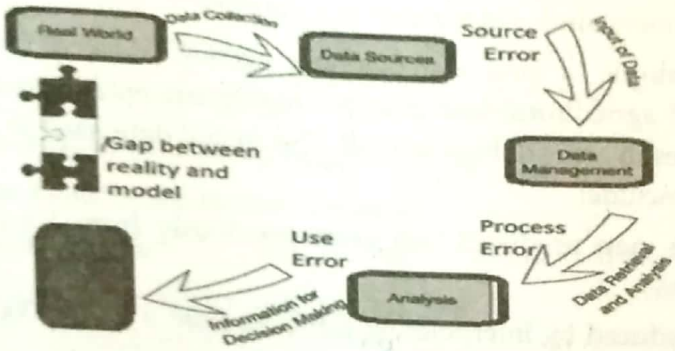
Implementation logic refers to how the model uses existing theory or knowledge to create new knowledge. Deductive approaches use knowledge of the overall situation in order to predict outcome conditions. This includes models that have some kind of formalized set of criteria, often with known weightings for the inputs, and existing algorithms are used to derive outcomes. Inductive approaches, on the other hand, are less straightforward, in that they try to generalize (often based upon samples of a specific data set) in order to derive more general models. While an inductive approach is useful if we do not know the general conditions or rules which apply in a given domain, it is typically a trial and-error approach which requires empirical testing to determine the parameters of each input variable.

Most GIS only come equipped with a limited range of tools for modeling. For complex models, or functions which are not natively supported in our GIS, external software environments are frequently used. In some cases, GIS and models can be fully integrated known as embedded coupling or linked through data and interface known as tight coupling. If neither of these is possible, the external model might be run independently of our GIS, and the output exported from our model into the GIS for further analysis and visualization. This is known as loose coupling.

6.7 ERROR PROPAGATION IN SPATIAL DATA PROCESSING

• How Errors Propagate

Error may be present in source data. It is important to note that the acquisition of base data to a high standard of quality still does not guarantee that the results of further, complex processing can be treated with certainty. As the number of processing steps increases, it becomes difficult to predict the behavior of error propagation. These various errors may affect the outcome of spatial data manipulations. In addition, further errors may be introduced during the various processing steps as illustrated in Figure below.



One of the most commonly applied operations in geographic information systems is analysis by overlaying two or more spatial data layers. Each such layer will contain errors, due to both inherent inaccuracies in the source data and errors arising from some form of computer processing, for example, rasterization. During the process of spatial overlay, all the errors in the individual data layers contribute to the final error of the output. The amount of error in the output depends on the type of overlay operation applied. For example, errors in the results of overlay using the logical operator AND are not the same as those created using the OR operator.

Coordinate Adjustment Transformation Projection changes Rescaling	**Feature Editing** Line Snapping Extension and reshaping Copying and Moving	**Generalization** Linear alignment Line Simplification Vertex editing	**Conversions** Raster cell to polygon Polygon to raster cell Cell alignment Line Thinning
Surface Modeling Contour Generation TIN Formation Cross section Generation Slope/Aspect determination	**Attribute Editing** Numeric calculation Value change/substitution Attribute redefinition Attribute update	**Boolean Operation** Polygon on Polygon Polygon on line Polygon on point Line on line, Overlay	**Data Input and Mgt.** Digitizing Scanning Topology Construction Dissolving Polygon
Display and Analysis cluster analysis calculation of surface lengths shortest route/path	**Display and Analysis** class intervals choice areal interpolation perimeter/area size/		

computation buffer creation display and query adjacency/contiguity	volume computation distance computation spatial statistics label/text placement		

Table : Lists common sources of error in GIS.

Note that these are from a wide range of sources, and include various common tasks relating to both data preparation and data analysis. It is the combination of different errors that are generated at each stage of preparation and analysis which may bring about various errors and uncertainties in the eventual outputs.

In another example, a land use planning agency is faced with the problem of identifying areas of agricultural land that are highly susceptible to erosion. Such areas occur on steep slopes in areas of high rainfall. The spatial data used in a GIS to obtain this information might include:

- A land use map produced five years previously from 1 : 25, 000 scale aerial photographs,
- A DEM produced by interpolating contours from a 1 : 50, 000 scale topographic map, and
- Annual rainfall statistics collected at two rainfall gauges.

- **Quantifying error propagation**

Chrisman noted that "the ultimate arbiter of cartographic error is the real world, not a mathematical formulation". It is an unavoidable fact that we will never be able to capture and represent everything that happens in the real world perfectly in a GIS. Hence there is much to recommend the use of testing procedures for accuracy assessment.

Various perspectives, motives and approaches to dealing with uncertainty have given rise to a wide range of conceptual models and indices for the description and measurement of error in spatial data. All these approaches have their origins in academic research and have strong theoretical bases in mathematics and statistics. Here we identify two main approaches for assessing the nature and amount of error propagation:

1. Testing the accuracy of each state by measurement against the real world, and
2. Modelling error propagation, either analytically or by means of simulation techniques.

Modeling of error propagation has been defined by Veregin as: "the application of formal mathematical models that describe the mechanisms whereby errors in source data layers are modified by particular data transformation operations." In other words, we would like to know how errors in the source data behave under manipulations that we subject them to in a GIS. If we are able to quantify the error in the source data as well as their behaviour under GIS manipulations, we have a means of judging the uncertainty of the results. Error propagation models are very complex and valid only for certain data types (e.g. numerical attributes). Initially, they described only the propagation of attribute error. More recent research has addressed the spatial aspects of error propagation and the development of models incorporating both attribute and locational components.

Rather than explicitly modeling error propagation, is often more practical to test the results of each step in the process against some independently measured reference data.

QUESTIONS

1. Explain Aronoff's classification of analytical function of GIS.
2. List and explain the measurements on vector data.
3. List and explain the measurements on raster data.
4. Write a note on spatial selection queries.
5. Write a short note on classification. Also explain user controlled and automatic classification.
6. Explain vector overlay operation.
7. Explain raster overlay operation.
8. Explain using suitable diagram overlays using decision table.
9. Write a note on neighborhood function.
10. Write a note on raster based surface analysis.
11. Explain network analysis.
12. How error propagates in spatial data processing?