Analysis of Discrete Entities in Space

The aim of GIS is not just to create a database of digital representations of geographical phenomena, but to provide means of selecting, retrieving, and analysing them. The previous chapter outlined some approaches to summarizing geographical data and assessing patterns. This chapter explains some of the methods available for dealing with crisp entities ('things')—how they can be selected from the database in terms of their attributes, and how new attributes can be computed ('modelled') using the rules of Boolean logic and mathematics to yield useful groups or classes, or to generalize complex map images. Many of these procedures are not really spatial because they only affect the attributes and not the size, shape, or form of the spatial entities, which can be any geographical primitive—point, line, area, or pixel. Spatial analysis often begins with the determination of spatial inclusion or exclusion, and with the intersection of lines and areas of different kinds to yield new entities. Spatial interactions are not just limited to the boundaries of existing entities, but may be extended to include neighbourhood functions such as crow's flight distances, topological proximity, and distance over networks such as roads or rivers. These procedures are illustrated by examples from several disciplines, including land evaluation and planning.

Learning objectives

By the end of this chapter, you will:

- **understand some key concepts and classes of approaches which can be used to work with discrete entities**
- **have developed a knowledge of some key ways of analysing data on discrete entities**
- **be able to apply your knowledge to reclassify data, measure distances from objects, and assess if and how features in different layers overlap.**

7.1 Spatial analysis is more than asking questions

The kinds of analytical techniques that can be used on spatial data depend greatly on the data model and the representation that have been used. It is important to realize that different data models and different kinds of representation can require different approaches to the

way spatial queries can be formulated. The fundamental question is whether the basic data model refers to *entities in space* or to the *continuous variation of an attribute over space*. In the case of entities, data retrieval and analysis concern the attributes, location, and connectivity of the entities and measures of the way they are distributed in space. In the case of continuous fields, data analysis concerns the spatial properties of the fields. The matter is made more complicated by the fact that continuous fields are usually **discretized** to a set of triangles or a regular grid, and the individual triangles or grid cells (or particular sets of contiguous triangles or grid cells) can also be treated as individual entities. In Chapter 2, the vector data model was introduced and this model was shown to be appropriate for representing discrete features. Chapter 3 included a discussion of databases in GIS and made links to vector topology. Most of this chapter assumes a vector data model, although not exclusively since the raster model is often used to represent objects, with remotely sensed images (for example) containing discrete features (such as roads and buildings) as well the terrain surface.

This chapter concentrates on the methods of data analysis that are most useful for dealing with *entities in space*, either in the relational or object-oriented model; the analysis of continuous fields is covered in Chapter 10. The fundamental axioms for data modelling and analysis were presented in Chapter 2. This chapter demonstrates the applications of these axioms and how they can be translated into computer commands to solve practical problems using discrete entities. Chapter 10 extends this discussion to continuous fields.

7.2 The basic classes of operations for spatial analysis

In the entity model of objects in space, three kinds of information are important: *what is it?*, *where is it?*, and *what is its relation to other entities*? The nature of an entity is given by its attributes, its whereabouts by its geographical location or coordinates, and the spatial relations between different entities in terms of **proximity** and connectivity (topology). The aspects of location, proximity, and topology distinguish geographical data from many other kinds of data that are routinely handled in information systems.

We distinguish the following basic classes of data analysis options for entities:

Attribute operations

- Operations on one or more attributes of an entity
- Operations on one or more attributes of multiple entities that overlap in space
- Operations on one or more attributes that are linked by directed pointers (object orientation)
- Operations on the attributes of entities that are contained by other entities (point in polygon)

Distance/location operations

- Operations to locate entities with respect to simple Euclidean distance or location criteria
- Operations to create buffer zones around an entity

Operations using built-in spatial topology

- Operations to model spatial interactions over a connected net.

All of these operations can result in new attributes, which can be attached to the original entities, thereby increasing the size and value of the database. Certain operations also create new spatial entities, requiring the database to be expanded to include these new items. Data that have been retrieved from the database can be displayed on the screen, plotted as a paper map, or written as a digital file for future processing.

Computing areas (see Box 7.1) and perimeters are common operations on discrete entities.

7.3 Operations on the attributes of geographic entities

As explained in Chapter 2, attributes are properties of entities that define what they are. They can be divided into three types—those that refer to location (the **geographical attributes** of latitude, longitude (or easting, northing), and elevation); those that are simply attached as qualitative or quantitative descriptors of some non-spatial property; and those that are derived from the spatial properties of the entity itself. For example, the attributes of parcel number, name of the owner, and land cover describe non-spatial properties of a piece of land. The length of the fence bordering the road, the area, shape, and contiguity are attributes that are derived from the form of the piece of land.

As in conventional information systems, new attributes can be attached to entities as the result of a database operation. For example, a new attribute (or a new value of

Box 7.1 Computing polygon areas using the trapezoidal rule

A polygon may be described in terms of a series of trapeziums, as shown in the figure below.

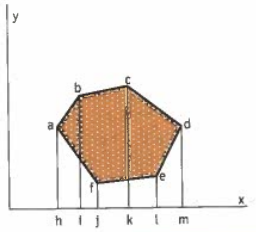
The area under the polygon is calculated by summing the areas of the various trapeziums that make up the total shape.

The area of a trapezium = (half the sum of its sides) × (horizontal distance)

The way to derive the total area, accounting for the varying levels of the sides, is to sum all the trapeziums that make each side of the shape in one direction and then subtract the total in the opposite direction.

The area of the polygon in the figure is computed by:

Add the areas of upper trapeziums A, B, C, and subtract the areas of lower trapeziums D, E, F. Upper trapeziums: A (h, a, b, i); B (i, b, c, k); C (k, c, d, m). Lower trapeziums: D (h, a, f, j); E (j, f, e, l); F (l, e, d, m). Lloyd (2010b) illustrates another approach to area calculation.



an attribute) can be computed for land parcels larger than a given size, or for those having owners that live abroad. For displaying information, the new attribute could be the colour or the symbol chosen to represent this kind of entity on the map. The new attribute can be derived by any legitimate method of logical and mathematical analysis, including operations on the proximity and topological properties of entities. Simple data retrieval can be seen as the creation of a new temporary attribute 'selected' when the set of attributes attached to an entity match the search criteria.

The process of selection or creation of new attributes can be formalized as follows. For any given location x, the value of a derived attribute U_i is given by:

$$U_i = f(A, B, C, D, \ldots) \qquad 7.1$$

where A, B, C, ... are the values of the attributes used to estimate U_i. The function $f()$ can be any of the following, singly or in combination:

a) logical (Boolean) operations
b) simple and complex arithmetical operations and numerical models
c) univariate statistical analysis
d) multivariate statistical methods or Bayesian statistics for classification and discrimination
e) multicriteria methods, artificial intelligence (AI)-based methods: neural networks.

Vector data in multiple separate layers can be easily combined and queries conducted using these combinations.

Data retrieval using the attributes attached to individual entities

Entities can be selectively retrieved or reclassified on their attributes by using the standard rules of Boolean algebra, which are incorporated in database languages such as SQL (Worboys 2005 provides an introduction to relational database management systems in GIS, as well as other database frameworks). **Boolean algebra** uses the logical operators AND, OR, XOR, NOT to determine whether a particular condition is true or false (Box 7.2). Each attribute is thought of as defining a *set*. The operator *AND* (symbol $\wedge$) is the *intersection* of two sets—those entities that belong to both A and B; *OR* (symbol $\vee$) is the *union* of two sets—those entities that belong either to set A or to set B; *NOT* (symbol $\neg$) is the *difference* operator, identifying those entities that belong to A but not to B, and *XOR* (symbol $\veebar$) is the *exclusive OR*, or the set of objects that belong to one set or another, but not to both. These simple set relations are often portrayed visually in the form of Venn diagrams (Figure 7.1). Note that logical operations can be applied to all data types, whether they are Boolean, nominal, ordinal, scalar, or directional.

Two simple examples illustrate the principles. Consider first a spatial database used by an estate agent. A typical retrieval query from a prospective buyer might be the following: 'Please show me the locations of all houses costing between $200 000 and $300 000 with four bedrooms and plots measuring at least 300 m²'. If the data set contains the attributes 'cost', 'number of bedrooms', 'area of plot', and location, a map of the desired premises

Box 7.2 Mathematical operations for transforming attribute data

a) Logical operations

Truth or falsehood (0 or 1) resulting from union (∨ logical OR), *intersection* (∧ logical AND), *negation* (¬ logical NOT), and *exclusion* (⊻ logical exclusive or XOR) of two or more sets.

b) Arithmetical operations

New attribute is the result of addition (+), subtraction (–), multiplication (*), division (/), raising to power (**), exponentiation (exp), logarithms (ln—natural, log—base 10), truncation, or square root.

c) Trigonometric operations

New attribute is the sine (sin), cosine (cos), tangent (tan) or their inverse (arcsin, arccos, arctan), or is converted from degrees to radians or grad representation.

d) Data type operations

New attribute is the original attribute expressed as a different data type (Boolean, nominal, ordinal, directional, integer, real, or topological data type).

e) Statistical operations

New attribute is the mean, mode, median, standard deviation, variance, minimum, maximum, range, skewness, kurtosis, etc. of a given attribute represented by *n* entities.

f) Multivariate operations

New attribute is computed by a multivariate regression model.

New attribute is computed by a numerical model of a physical process.

New attribute is computed by a *principal component analysis*, *factor analysis*, or *correspondence analysis* transformation of multivariate data.

Entity is assigned to a given *class* (new attribute = class name) by methods of multivariate numerical taxonomy.

Entity is assigned a *probability* (based on statistical chance) by discriminant analysis, maximum likelihood, or Bayesian techniques, of belonging to a given set.

Entity is assigned a *fuzzy membership value* for a given set.

Entity is assigned to a class using neural network methods.

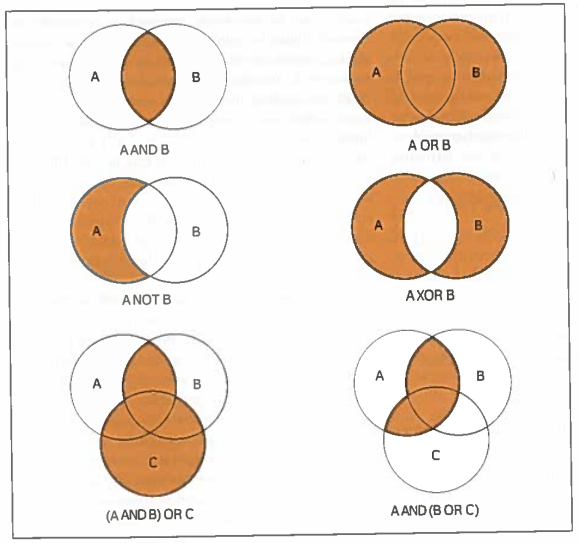


Figure 7.1 Venn diagrams showing the results of applying Boolean logic to the union and intersection of two or more sets; in each case the shaded zones are 'true'

is easily produced by a multiple AND query on the specified attributes to highlight the matching plots:

IF COST GE $200 000 AND COST LT $300 000 AND N BEDROOM = 4 AND PLOT AREA GE 300 THEN ITEM = 1 ELSE ITEM = 0

The selected entities are given a Boolean value of 1 (true) if they match the specifications, and a 0 (false) if not. Display of the results follows by assigning a new colour to entities with ITEM = 1.

Now consider a query in land suitability classification. In a database of soil mapping units, each mapping unit may have attributes describing the texture and pH of the topsoil. If set *A* is the set of mapping units called *Oregon loam* (nominal data type), and if set *B* is the set of mapping units for which the topsoil pH equals or exceeds 7.0 (scalar data type), then the data retrieval statements work as follows:

X = *A AND B* finds all occurrences of Oregon loam with pH ≥ 7.0

X = *A OR B* finds all occurrences of Oregon loam, and all mapping units with pH ≥ 7.0.

X = *A XOR B* finds all mapping units that are either Oregon loam, or have a pH ≥ 7.0, but not in combination.

X = *A NOT B* finds all mapping units that are Oregon loam where the pH is less than 7.0.

Selected entities can also be renamed and/or given a new display symbol (Figure 7.2) by statements such as: 'Give the designation "Suitable" to all mapping units with soil texture = "loam" and pH ≥ 5.5.' This is a particular instance of the logical statement 'IF condition (*C*) THEN carry out specified task.'

Note that, unlike arithmetic operations, Boolean operations are not commutative. The result of *A AND B OR C* depends on the priority of *AND* with respect to *OR*. Parentheses are usually used to indicate clearly the order of evaluation when there are more than two sets (Figure 7.1). For example, if set *C* contains mapping units of poorly drained soil, then *X* = *(A AND B) OR C* returns all mapping units that are either (a) Oregon loam with a pH ≥ 7.0 or (b) units of poorly drained soil. The relation *X* = *A OR (B AND C)* returns (a) all Oregon loam mapping units and (b) those mapping units with a combination of pH ≥ 7.0 and poor drainage.

Note also that Boolean operations may require an exact match in attributes to return data, and they take no account of errors or uncertainty unless that is specifically incorporated into the definitions of the sets.

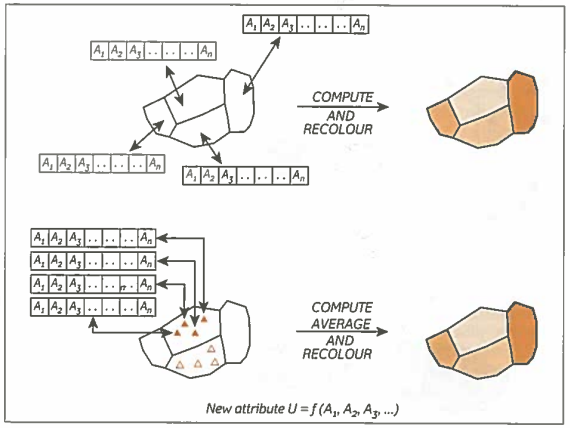


Figure 7.2 When retrieving entities on attributes alone, or computing new attributes from old ones, the spatial units of the map do not change shape, only colour or shading (top). The same occurs when a point-in-polygon search is used to find enclosed point objects that are used to compute area averages (bottom).

If the value of the attribute 'elevation' is set at 2000 m above sea level to define the class 'mountain,' then hills with elevations up to 1999.999 999 999 ... m will be rejected. This is not a problem with ordinal and nominal data types but it can present problems when working with scalar data types that represent quantities like elevation, pH, clay content, soil depth, atmospheric pressure, salinity, population, and so on, that are subject to various sources of measurement error and uncertainty. If the error bands on these data span the boundary values of sets then strict application of Boolean rules may yield results that are counter-intuitive.

Spatial aspects of Boolean retrieval on multiple attributes of single entities

Carrying out logical retrieval and reclassification on the non-spatial attributes of spatial entities has little effect on the map image, except in terms of symbolism and boundary removal. Computing a new attribute or condition requires the preparation of a legend and a recolouring or reshading of the selected entities (see Figure 7.2). When selection leads to adjacent polygons receiving the same code it may be sensible to dissolve the boundaries between them, achieving a form of map generalization (Figure 7.3). Figure 7.4 illustrates the use of this option to simplify a complex soil map.

Boolean operations are not only applicable to the non-spatial attributes of the geographical elements—they can also be applied to the geographic location and attributes derived from the spatial or topological properties of the geographic entities. For example, one might wish to find all mapping units exceeding 5 ha in areas having soil with clay loam texture in combination with a pH > 7.0. More complicated searches may involve the shapes of areas, the properties of the boundaries of areas, or the properties of neighbouring areas such as the areas of woodland bordering urban areas. In these cases, the results of the search would have an effect on the spatial patterns.

Simple and complex arithmetical operations on attributes of single entities

New attributes can be computed using all normal arithmetical rules (+, −, /, *, logarithms, trigonometric functions, exponents, and all combinations of them, including complex mathematical models). Arithmetical and trigonometrical operations can obviously only be used on scalar data and certain kinds of ordinal data. Arithmetical operations on Boolean data types and nominal data types are nonsensical and therefore not allowed (an expression such as $X = sqrt$ (London) has no meaning). Arithmetical operations can be very simple, or very complicated, but in all cases the operation is the same—a new attribute (a result) is computed from existing data.

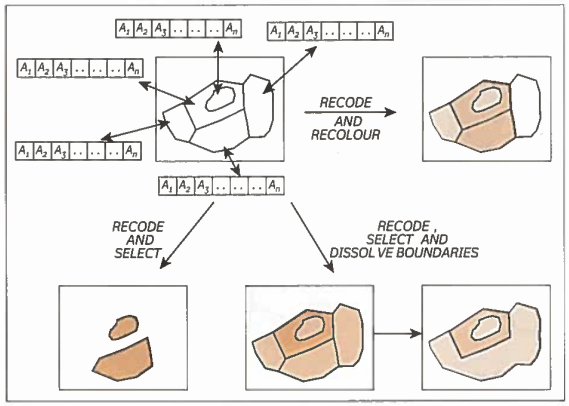


Figure 7.3 If, during a retrieval or recoding operation, two adjacent polygons receive the same new code, boundaries between them can be dissolved, leading to map generalization

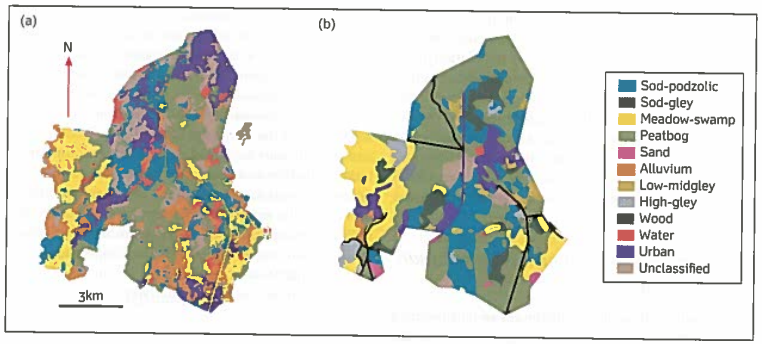


Figure 7.4 Using reclassification as a means of map generalization. Left: original soil map with 95 different units; right: reclassified map with 12 units. Note that reclassification preserves the original geometry and that the legend and shading codes only apply to the simplified map.

Some hypothetical examples of computing new attributes for a given administrative area (polygon) or point location are:

- Population increase = population 2010 − population 2000
- Total spending power = average income × number of persons
- Average wheat yield per farm = (total yield) / (number of farms)
- Predicted wheat yield = *f*(crop), where *f*(crop) is a complex mathematical model that computes wheat yield as a function of the soil, moisture, nutrients, and properties of a site (point entity)
- Class allocation = *Result* (multivariate classification) where (multivariate classification) might be any statistical or multicriteria analysis of the numerical attributes on the entity.

For a set of river catchments, the proportion of precipitation discharging through the outlet can be computed by dividing the annual precipitation for each catchment by the cumulative annual river discharge measured at the outlet.

Arithmetical operations can be easily combined with logical operators:

```
IF (A + B)/(C) ≥ TEST VALUE THEN CLASS = GOOD
```

The statistical analysis of attributes

Chapter 6 provides a brief introduction to the principles of univariate and multivariate statistical analysis. Readers requiring further details of these methods should consult a standard text such as Davis (2002) for geology and earth sciences, Rogerson (2006) for geography, Jongman et al. (1995) for ecology, and Haining (2003) for the environmental and social sciences. In this chapter we assume that statistical methods are just one set of procedures out of many for computing new attributes.

Simple statistical analysis can be used to compute means and standard deviations, and to conduct correlation and regression analysis (see Chapter 6). The operations can be applied to a set of attributes attached to single entities or to any set of entities that are retrieved by a logical search. For example, compute the mean and variance of nitrate levels for all groundwater monitoring sites included in a province P (e.g. Figure 7.2, lower example), or compute the average takings for all fast-food outlets in the postcode districts that include railway stations.

Numerical models

The range of arithmetical operations that can be applied to numerical attributes is unlimited. They are often used to compute the values of attributes that are difficult or impossible to measure, or which can be derived from cheap, readily available base data, such as data from censuses or natural resources surveys. Standard sets of mathematical operations that have been derived from empirical (regression) modelling are sometimes called **transfer functions** (see Ghermandi and Nunes 2013 for an application concerned with the values of coastal recreation) in disciplines like soil science and land evaluation. More complex sets of mathematical functions

that represent a physical process such as crop growth, air quality, groundwater movement, pesticide leaching, epidemiological hazards, increase of population pressure, etc. are often referred to as **models** (described in more detail in Chapter 12). Most GIS do not provide the functionality to program these complex models; instead they are used to assemble the data and export them to the model, which might reside on another computer on the network. Development of new tools using the R programming language[1] is, however, providing greater flexibility.

Neural networks, multicriteria evaluation, and fuzzy logic

All methods of deriving new attributes so far presented are **parametric**, which is to say that they assume that the definition of a new attribute can be expressed by a logical or numerical equation in which weights or parameters can be objectively assigned. Regression analysis and the calibration of numerical models are just two examples of ways in which the 'best' parameter values are chosen for classification or calculation. It is worth adding that in most cases the numerical models are also linear—i.e. there is a direct relation between a parameter value and its effect on the output.

These assumptions derive from classical, mechanical science. But parametric methods are difficult to use in complex, non-linear situations where attribute values are not normally distributed and where causal or even statistical relations are tenuous. These difficult conditions surround many spatial data, whose interrelations may violate many of the basic tenets of parametric methods, and problems as simple as how best to classify complex spatial objects may be intractable. This problem has received particular attention in the classification of remotely sensed data into land cover classes (Lees 1996).

Neural networks are powerful ways of classifying geographic entities (entities or pixels) into sensible groups (Atkinson and Tatnall 1997; Kia et al. 2012 detail a recent application concerned with flood simulation using GIS). In many cases the analyses are not really spatial, but require an entity to be assigned to a class on the basis of a non-statistical method of computation. A neural network is a processing device, implemented as an algorithm or computer program, that is said to mimic the human brain. It comprises many simple processing elements that are connected by unidirectional communication channels carrying numeric data. The elements operate only on local data but the network as a whole organizes itself according to perceived patterns in the input data. These patterns can be created by 'self-learning'—the system determines the 'best' set of classes for the data—or 'supervised classification', in which the system is supplied with a template of the required classification.

Multicriteria evaluation and fuzzy logic. Neural networks are not the only ways of dealing with complexity and non-linearity in spatial data. Methods of **multicriteria evaluation** and multicriteria decision analysis (as defined in Section 7.7) have been developed to provide users with the means to determine new attributes that indicate alternative responses to problems involving multiple and conflicting criteria. In Chapter 13 we explore the particular use of fuzzy logic and continuous classification for spatial data analysis in GIS.

[1] http://www.r-project.org/

7.4 Examples of deriving new attributes for spatial entities

The results of computing new attributes or reclassification are usually displayed by reshading or recolouring the entities (Figure 7.2). As with Boolean selection, the spatial properties of the entities (location, shape, form, topology) do not change, except in the case that neighbouring entities are reclassified as being the same, when generalization can take place. Note that if the data are in raster form, these operations are carried out on each pixel separately, unless the data structure uses a map unit-based approach to raster data coding (see Chapter 3). The following sections give examples of spatial analysis based entirely on the derivation of new attributes. Section 12.1 gives an example of a regression model of surface elevation and temperature in the Swiss Alps—the regression equation can be used to derive estimates of temperature at all cell locations for an altitude matrix covering the study area.

Using multivariate clustering

Geodemographic segmentation (Harris et al. 2005) is a method used by multinational marketing companies to classify residential areas of Western countries into distinct neighbourhood types based on statistical information about the consumers who live in them. The spatial entities are provided by census districts, postal code districts, or mail order address units linked to the universe of national house addresses. We consider a case where each spatial unit is characterized by four key criteria: age (young, middle, and old), income (high, middle, and low), urbanization (metropolitan, urban, suburban, rural), and family type (married couples with children, singles and childless couples, and pensioners). These attributes yield 108 possible combinations of classes, which can be recoded to ten core classes for the identification of characteristic socio-economic types. Multivariate methods of class allocation may then be used to assign a basic spatial unit to one of these ten classes, and they are also linked to other information including empirical sales data and consumer preference attributes obtained by questionnaire. Simple logical retrieval of spatial units in terms of their class and attributes provides maps at local or national level that show the spatial distribution of market opportunities and brand preferences.

Using simple Boolean logic

In many parts of the world there are insufficient data to compute crop yields with numerical models of crop growth as a function of available moisture, energy, and water, so qualitative predictions based on simple rules may be the only useful way to assess the suitability of land for agricultural development. This is the philosophy behind the now classic FAO land evaluation procedure (FAO 1976; Beek 1978).

Prescriptive land evaluation (Rossiter 1996; Castella et al. 2007) is based on the simple idea that a landscape can be divided into basic entities called *mapping units*, separated by crisp boundaries. Soil survey is often the basis for this kind of landscape division, but alternative methods use landform, ecological zones, or vegetation communities. The idea is that once basic spatial entities have been mapped and defined in terms of representative attribute values, their suitability for any given purpose may be determined by reclassification or by computing new attributes on the basis of existing information. The general procedure is called 'top-down' logic, because it starts with the presumption that global rules or physical models exist for translating primary units of information into units that can be used for a specific purpose. The procedure is exactly the same whether the data are stored and displayed as vector polygons or as grid cells; the only difference is the kind of spatial representation for the conceptual entity carrying the information.

The following is an example of this rule-based reasoning. In order to grow well, a crop needs a moist, well-drained, oxygenated, fertile soil. Growing a monocrop leads to the soil being bare, or nearly bare, for part of the year, and in this time the soil should be able to resist the effects of erosion by rain. The four attributes—available moisture, available oxygen, nutrients, and erosion susceptibility—are known as *land qualities* (LQ_i), and they may be derived from primary soil and land data using simple logical transfer functions derived by agronomists and soil experts. The overall suitability of a site (its *land quality* for the use intended) is determined by the most limiting of the land characteristics—a case of worst takes all.

Figure 7.5 illustrates the complete procedure using data from a soil series map and report and a digital elevation model of a small part of the Kisii District in Kenya (Wielemaker and Boxem 1982). The study area covers some 1406 ha (3750 × 3750 m²) of the area mapped by the 1:12 500 detailed soil survey of the Marongo area (Boerma et al. 1974) and was chosen for its wide variety of parent material (seven major geological types), relief (altitude ranges from 4700 to 5300 feet above sea level—1420 to 1600 m), and soil (12 mapping units). Detailed soil survey information describes parent material, soil series, soil depth to weathered bedrock, stoniness and rockiness, and tabular information relating soil series to land qualities. Each attribute was digitized as a separate polygon overlay and converted to a 60 × 60 array of 62.5 m square cells. The digital elevation model was obtained by interpolation from digitized contours and spot heights and converted to local relief (minimum 40 m, maximum 560 m). Information about the climate, the chemical status of the soil, the land use, and cultural practices is also available. Slope lengths (needed for estimating erosion) were interpreted from stereo aerial photographs and digitized as a separate overlay.

In this example we consider suitability for smallholder maize, which is determined by the land qualities *nutrient supply*, *oxygen supply*, *water supply*, and *erosion susceptibility*. These land qualities can be ranked by assigning values of 1, 2, 3, respectively, to the following classes:

No limitation	assign 1
Moderate limitation	assign 2
Severe limitation	assign 3

The rules for deriving the land qualities are: water availability is a Boolean union (AND) of soil depth and soil series (B); oxygen availability and nutrient availability can be derived directly from the soil series information by recoding using a lookup table (L). Erosion susceptibility or hazard can be determined as a Boolean union (AND) of slope classes and soil series. Examples of the rules are:

If soil series is S_2 then assign nutrient quality W_3 from lookup table L_w.

If soil series is S_2 and slope class is 'flat' assign erosion susceptibility 1.

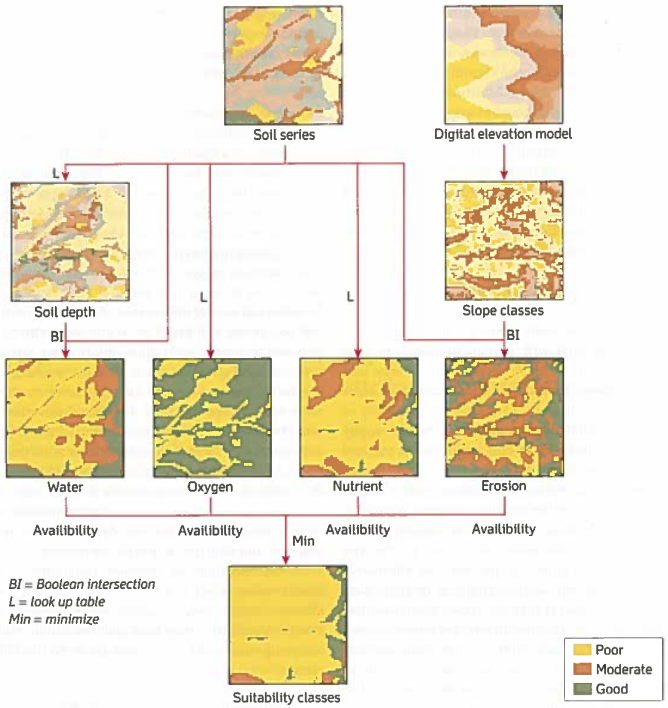


Figure 7.5 The flow chart of operations for 'top-down' land evaluation for determining the suitability of land to grow maize

If soil series is S_2 and slope class is 'steep' assign erosion hazard value 3.

Once the individual land qualities have been assigned, the overall suitability per polygon or pixel is determined by the land quality with the most limiting (largest) value:

$$Suitability = maximum\ (LQ_{water}, LQ_{oxygen}, LQ_{nutrients}, LQ_{erosion})$$

This gives suitabilities of *poor*—at least one serious limitation; *moderate*—not severe but at least one moderate limitation; and *good*—no limitations.

It is simple to repeat the analysis for the same area using other values of the conversion factors relating soil to the land qualities, to see how the areas of suitable land may increase as limitations are dealt with by irrigation, mulching, or terracing. This *scenario exploration* can be achieved by replacing the real data with values of the land characteristics that represent a more degraded or an improved situation. In this way one can use the logical model to explore how the suitability of an area depends on the different factors. One must not forget, however, that the results are no better than the data and the insight in the land evaluation procedure allow.

7.5 Operations that depend on a simple distance between A and B: buffering

Operations of the type '*A* is within/beyond distance *D* from *B*' where *D* is a simple crow's flight distance are carried out with the help of a **buffering** command. This is used to draw a zone around the initial entity where the boundaries of the zone or buffer are all distance *D* from the coordinates of the original entity. If it is a point entity then the zone is a circle, if a straight line, the zone is a rectangle with rounded ends, or if it is an irregular line or polygon, an enlarged version of the same shape (Figure 7.6). The buffer is in effect a new polygon that may be used as a temporary aid to spatial query, or may be itself added to the database. The determination of whether an entity is inside/outside or overlaps the buffer zone is then carried out using the overlay operations, as described later in this chapter (see Section 7.7), and logical or mathematical operations on those entities proceed as before.

Typical examples of using the zoning/buffering command with other analysis options include:

- 'Determine the number of fast-food restaurants within 5 km of the White House.'
- 'Investigate the potential for water pollution in terms of the proximity of filling stations to natural waterways.'
- 'Compute the total value of the houses lying within 200 m of the proposed route for a new road.'
- 'Compute the proportion of the world population that lives within 100 km of the sea.'
- 'Compute the number of cattle grazing within 5 km of a waterhole.'
- 'Determine the potential amount of arable land within 1 hour's walk from a Neolithic village.'

Figure 7.7 shows railways in Wales, while Figure 7.8 shows a 5 km buffer computed for the railways. Buffers are very widely used in site-selection exercises where simple distance from a feature is of interest. Other approaches to dealing with distances are considered in the next section.

In such cases, distances from multiple features (e.g. roads or rivers) may be essential inputs, and so buffer polygons may be created and used to identify areas which fall within or outside of these distance bands.

7.6 Operations that depend on connectivity

These are operations in which the entities are directly linked in the database. The linkage can be spatial, as in the contiguity case where *A* is a direct neighbour of *B* or the case where *A* is connected to *B* by a topological network that models roads or other lines of communication. Entities can also be linked by an internal topology, so that

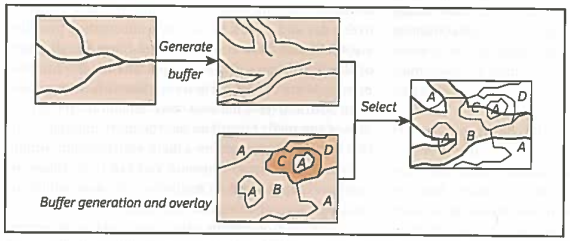


Figure 7.6 Generating buffer zones around exact entities such as points, lines, or polygons yields new polygons which can be used in polygon overlays to select defined areas of the map

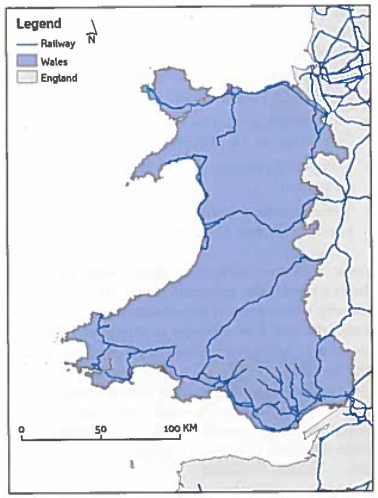


Figure 7.7 Railways in Wales (Contains Ordnance Survey data © Crown copyright and database right 2014)

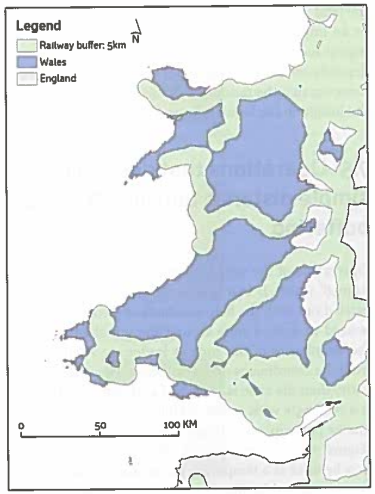


Figure 7.8 Buffer for railways in Wales: 5 km (Contains Ordnance Survey data © Crown copyright and database right 2014)

complex spatial entities are made up of sets of subentities, as is the case with object orientation.

The operations '*A* is a direct neighbour of *B*' and '*A* is connected to *B* by a topological network' are two versions of the same class of operations which, for topologically connected lines and polygons, use explicit information from the spatial database (see Chapter 3) to determine how two entities or locations are connected. Inter-entity distances over the network or other measures of connectivity such as travel times, attractiveness of a route, etc. can be used to determine indices of interaction. These operations are much used for determining the location of emergency services or for optimizing delivery routes.

For example, when a boundary between two land cover polygons is also defined as a road, it is a simple matter to select those roads/boundary lines that have particular kinds of land cover on both sides. Such an analysis would easily distinguish rural roads (agriculture on both sides) from urban roads (built-up areas on both sides) from coastal roads (sea on at least one side—to take account of sea dykes and breakwaters).

The analysis of connectivity over a topologically directed net is much used in automated route-finding (vehicle navigation systems) and for the optimum location of emergency services (see Cromley and McLafferty 2011 for a summary). Attributes attached to the line elements representing roads, rivers, or rail links can identify the character of the connector. For example a road can be identified not only by its width, surface, class, and number of lanes, but also by its visual attractiveness or otherwise (potential tourist route) or traffic densities. Linking time series data on traffic densities over a day and a week to the route information provides a sound basis for estimating travel times for all hours of the day, factors that are important for the location of emergency services. Figure 7.9 shows how the route from *A* to *B* over a network may depend on the attributes of the roads taken and may be quite different from that computed from a crow's flight path based on simple buffering. Figure 7.10 (Ritsema Van Eck 1993) shows an example output from an analysis of the accessibility of built-up areas for emergency services.

A common operation which uses data on transport networks is the **shortest path algorithm**, which is used to find the shortest (or perhaps cheapest) route between two or more points on a network (see Lloyd 2010b for more details).

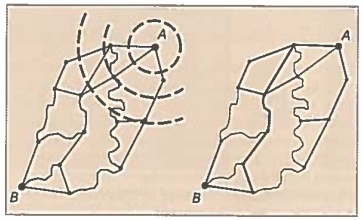


Figure 7.9 The analysis of transport times from A to B in terms of (a) crow's flight distance and (b) times along different routes in a network to determine expected travel times for different road conditions

7.7 Operations on attributes of multiple entities that overlap in space

Here we extend the discussion of operations on attributes to include attributes from two or more entities that completely or partially occupy or cover the same space. In other words, we consider the **inclusion problem**:

A contains *B*, or

A is contained by *B*,

and the overlap and intersection problem:

A crosses *B*

A overlaps with *B*

where *A* and *B* are two different spatial entities.

Inclusion

The cases '*A* contains *B*' and '*A* is contained by *B*' are solved by extending the rules of Boolean algebra from attributes of entities to measures of how entities occupy space. The problem is the well-known 'point-in-polygon' issue, which is outlined in Box 7.3 . The first step in the analysis is to determine which entities are included or excluded in the location sense—e.g. 'Which restaurants are located in Soho?', 'Which groundwater observation wells have been drilled in formation X?' Once the entities have been selected and tagged, the procedures for attribute analysis can be applied, either per entity, or collectively. For example, the minimum and maximum water levels could be extracted for a given year for each groundwater well, or the average water level of all wells

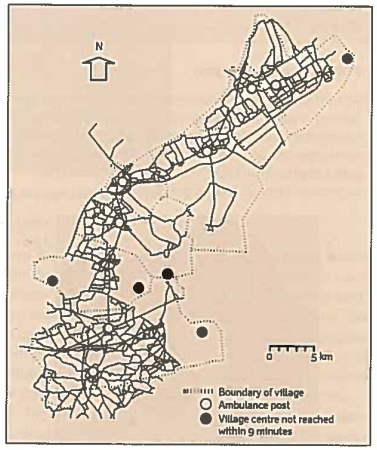


Figure 7.10 The results of a transport time analysis to see which suburban areas can be reached by ambulance within 9 minutes from the ambulance posts. Black lines show roads that can be reached within 9 minutes from the ambulance posts (open circles). Black circles show local centres that cannot be reached in that time. Pecked lines show outlines of urban areas.

could be computed. The result of these computations can be used to tag the enclosing polygon, which can be displayed with a new colour, shading, or label (Figure 7.2, lower example).

Other examples of applications of this kind of analysis are: from archaeology, 'determine the number of late Iron Age burial sites in parish A,' or 'retrieve all passage graves and determine the kinds of soil and landscape position where they occur'; or from soil science: 'find all soil profiles located in unit S_1, and compute the mean value and standard deviation of the clay content of the topsoil.'

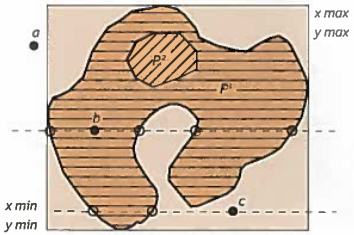
Entity overlap and intersection

In certain cases of '*A* contains *B* and *A* is contained by *B*' and with '*A* crosses *B* and *A* overlaps with *B*', where *A* and *B* are lines or polygons of different forms, the first steps in logical retrieval are to define new areas or line segments. With polygons the process is known as **polygon overlay and intersection**, and it leads to the creation of new spatial entities. Operations which transfer attributes from both input layers are overlay operators, while those

Box 7.3 The point-in-polygon problem and its solution

Point-in-polygon search

At least two separate steps in the creation of the polygon network involve the general problem of *point-in-polygon* search: the checks to see if a small polygon is contained by a larger one, and the association of a given polygon with a digitized text label. The figure below shows two aspects of the point-in-polygon algorithms.



First, a quick comparison of the coordinates of the point with the extents of the polygon quickly reveals whether a point is likely to be inside it or not. So point (*a*) may easily be excluded because it is outside the polygon's minimum bounding rectangle, but (*b*) and (*c*) cannot. To check if points (*b*) and (*c*) are in the polygon, a horizontal line is extended from the point. If the number of intersections of this line with the polygon envelope (in either direction) is odd, the point is *inside* the polygon.

To check if an island polygon, P^2, is inside the larger polygon, P^1, first check the extents; P^1 is then divided into a number of horizontal bands and the first and last point of each band is treated in the same way as the point (*b*) above. If the number of intersections for each line is odd, then the polygon P^2 is completely enclosed.

Problems may occur if any segment of a boundary is exactly horizontal and has exactly the same Y coordinate as the point X, but these may be easily filtered out.

which simply cut out part of one layer using boundaries contained in another layer are *cookie cutters*.

Figure 7.11 shows three different results, depending on whether the operation is spatial Boolean **union** (Figure 7.11a), **covering** (Figure 7.11b), or the **clip** cookie cutter (Figure 7.11c); further examples of overlay and cookie cutter operators are given below. Polygon overlay is used to answer questions such as 'Find the area of the City of Westminster that is covered by parks,' in which the first overlay may show the boundaries of the administrative areas of urban London and the second is an overlay delineating different kinds of land cover. An example of the clip cookie cutter operation (Figure 7.11c) in a physical application is that if map A is a map of soil types, and

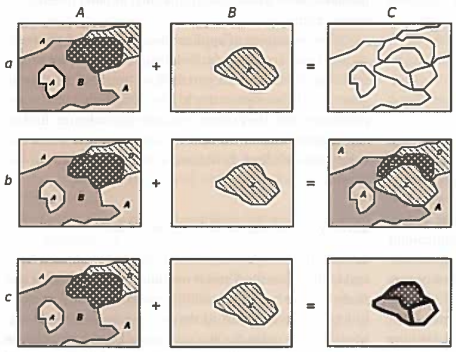


Figure 7.11 Polygon overlay leads to an increase in the number of entities in the database: (a) union overlay—all boundaries are retained; (b) second map covers the first and changes the map detail locally; (c) the covering map is used to cut out a small part of the first map (clip operator)

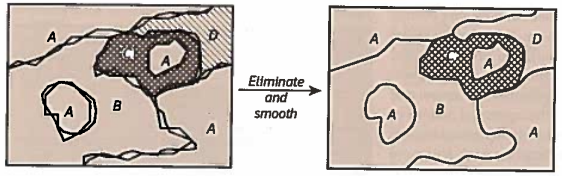


Figure 7.12 Polygon overlay can lead to a large number of spurious small polygons that have no real meaning and must be removed

map B shows the boundary of a catchment, the result is a soil map of that catchment alone.

In some situations, polygon overlap leads to the creation of so-called **spurious polygons** (Figure 7.12). This is because of small errors in the creation of boundaries that are supposed to lie in the same place. The errors can result from digitization errors or from errors made during surveying. There are several solutions. The first is to designate the boundaries on one feature layer as the dominant boundaries to which all others must defer. The second is to examine all the spurious polygons and eliminate all those which have an area smaller than some critical threshold; here it must be decided to which of the larger polygons the area covered by the spurious polygons should be added. The third is to pass a smoothing window over all the coordinates of spurious polygons along the conjugate boundary zones and to compute a new, average boundary. This is frequently over-defined and can be simplified using the Douglas–Peucker algorithm (Douglas and Peucker 1973) or other means, and smoothed for computing the boundaries of the new polygon entities and for display (Figure 7.12).

The application of overlay and **cookie cutter** operations is illustrated below using a single example. Figure 7.7 showed railways in Wales, while Figure 7.13 shows areas of woodland. A two-band buffer was computed around the railways to distances of 2.5 km and 5 km (Figure 7.14).

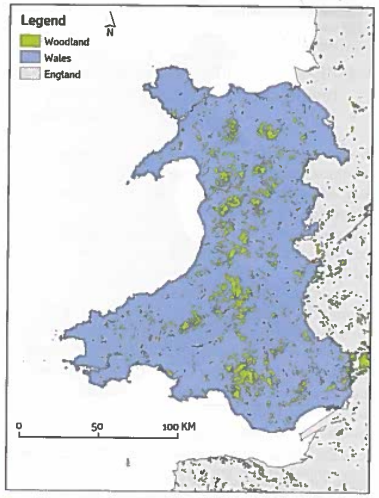


Figure 7.13 Woods in Wales (Contains Ordnance Survey data © Crown copyright and database right 2014)

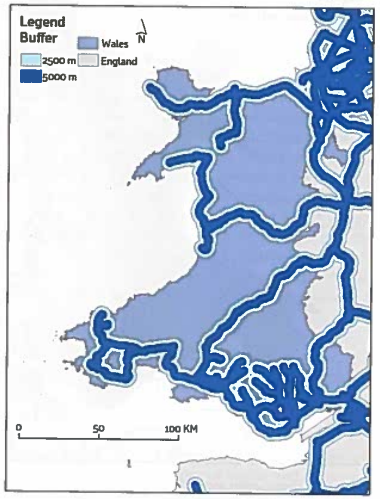


Figure 7.14 Buffer for railways in Wales: 2.5 km and 5 km (Contains Ordnance Survey data © Crown copyright and database right 2014)

The outcome of the union, **intersect**, and **identity** overlay operators with these two inputs (i.e. buffers and woods) are shown in Figures 7.15 (union), 7.16 (intersect), and 7.17 (identity). The outputs of the cookie cutter operations erase and clip are shown in Figures 7.18 (**erase**) and 7.19 (clip). Table 7.1 summarizes the inputs and outputs of each of the operations. With the overlay operators (union, intersect, and identity), all attributes for both layers are retained for the output areas. For the cookie cutter operations (erase and clip), attributes *only* for the input layers are retained—the erase and clip layers are used simply to cut out a part of the input layer. It is not always straightforward to select which overlay or cookie cutter operation is most appropriate, and often more than one operator may be suitable, in principle, for a particular task.

Overlay and cookie cutter operators are key tools in GIS multicriteria decision analysis (GIS-MCDA) (Carver 1991; Malczewski 1999; Bell et al. 2007). GIS-MCDA relates to GIS-based processes for decision-making using multiple data sources. Thus, overlay of multiple layers is needed to identify areas of overlap and extract subareas which meet criteria determined using combinations of these layers. The examples above relate to woodlands and road buffers, which can be combined in several different ways, allowing the selection of areas which are woodland or otherwise, and which fall within or outside the buffer zones. Wise (2002) shows how many key GIS algorithms work 'behind the scenes', and this includes discussion of methods for conducting overlay procedures.

Table 7.1 Overlay and cookie cutter operations for woods and buffers

Operator	Layer labels	Layers	Attributes
Union	Layers the same	All for all areas	All for all areas
Intersect	Layers the same	All for common areas	All for common areas
Identity	Input: buffers Identify: woods	All buffers and all woods inside buffers	For all buffers and all woods inside buffers
Erase	Input: woods Erase: buffers	All woods in buffer area	For woods inside buffer areas
Clip	Input: woods Clip: buffers	All woods in buffer area	For woods outside buffer areas

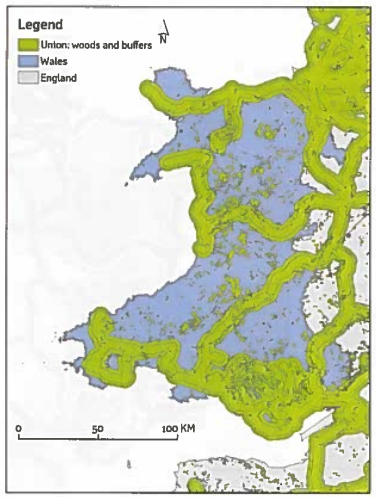


Figure 7.15 Union: woods and buffers—retains all features in both input layers (Contains Ordnance Survey data © Crown copyright and database right 2014)

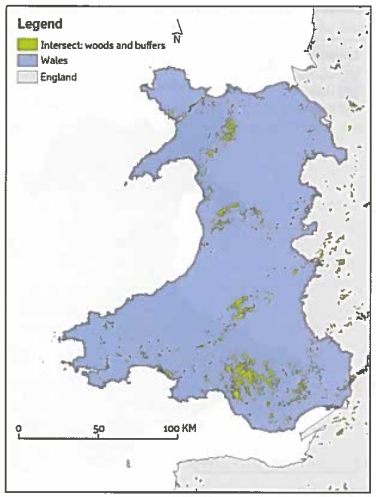


Figure 7.16 Intersect: woods and buffers—retains all features in overlapping areas (Contains Ordnance Survey data © Crown copyright and database right 2014)

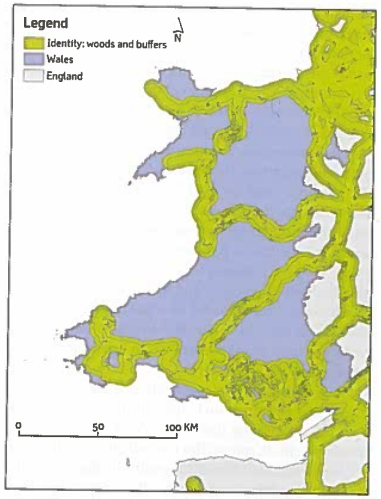


Figure 7.17 Identity: woods and buffers—retains all buffers and woods within buffers (Contains Ordnance Survey data © Crown copyright and database right 2014)

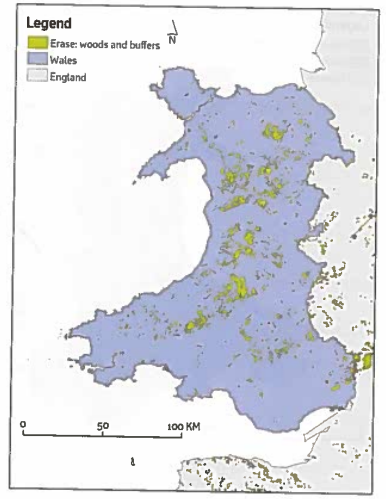


Figure 7.18 Erase: woods and buffers—retains all woods outside buffers (Contains Ordnance Survey data © Crown copyright and database right 2014)

Operations on one or more entities that are linked by directed pointers (object orientation)

Logical operations on lines and polygons that result in entities being split or removed from the database impose heavy computational costs on a spatial database because the number of entities may depend on the operations being carried out. In practical terms, if two simple polygons intersect to create a third, then the third polygon and the attributes it inherits from the two original polygons must be added to the database. If reclassifying two adjacent polygons results in the removal of a common, unnecessary boundary, then two polygons must be removed and one added to the database. In practice, the number of polygons added or removed may be large and indeterminate, so it is difficult to say just how great this overhead is. To save modifying the original database, the changes may be computed on a subset of the original data, and the results stored in a separate file or folder.

In hybrid-relational GIS, adding and removing polygons means modifying both the spatial data and the attribute data separately. Modifying the spatial data is more than just adding or deleting an entry in a table, because all the topological connections need to be recomputed. The advantage of network-relational hybrid databases is that, in principle, there is no limit to the kind and number of analysis queries that can be defined. Object-oriented GIS attempt to get around these computational problems by incorporating a large amount of information to structure the data in such a way that data volumes do not change greatly as queries are carried out. This means that the most common data retrieval and analysis options need to be thought out beforehand, which is why constructing an object-oriented database can take a great deal of time.

7.8 General aspects of data retrieval and modelling using entities

Spatial entities can be retrieved and new attributes can be computed by a wide range of logical and numerical methods. The numerical procedures can also be applied to inclusion and intersection problems and for proximity

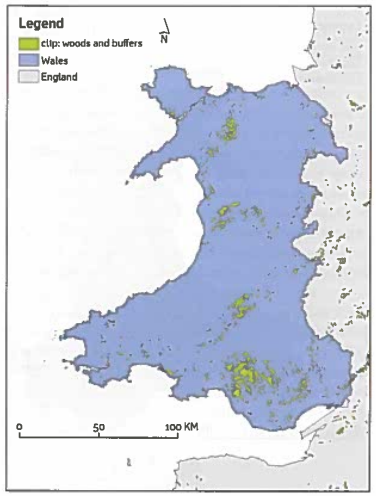


Figure 7.19 Clip: woods and buffers—retains all woods within buffers (Contains Ordnance Survey data © Crown copyright and database right 2014)

analysis, and for analysis of relations over topological connections. The methods can be combined to create complex models for addressing many different kinds of spatial problem. Note that many data-analysis operations are not commutative so the sequence in which the commands are executed is very important. While it can be very informative to sit in front of the computer browsing through a database to see what is there or how different procedures work (e.g. with exploratory data analysis), informal procedures are best for simple data retrieval and transformations. However, when a complex series of commands must be used frequently to retrieve and transform data it is sensible to create a structured command file that can be reviewed, modified, and used by several people (the R programming environment offers a wide range of packages which implement spatial analysis methods using command lines). Such a set of commands constitutes a 'model' or a 'procedure' which can be stored in the GIS, referenced directly by an icon or a name, and used on other databases to carry out the same set of operations.

None of the methods of analysis presented in this chapter pay any attention to data quality or errors; there is a tacit assumption that all data and all relations are known exactly. In spite of this (we will return to this topic in Chapter 12), spatial modelling with GIS has great value for exploring different scenarios. The tools described in this chapter have been widely used for site selection and assessment of alternative scenarios in planning contexts. GIS enables users to assess and present different possible outcomes which can then be visualized in flexible ways (see Chapter 5) to allow a wide range of interested parties to compare and contrast these alternatives.

7.9 Summary

In this chapter, the analysis of discrete features (points, lines, and areas) has been the focus. Attribute operations, such as reclassification of areas, were outlined along with explicitly spatial operations such as measurement of distances from objects, overlay operators, and cookie cutter operations. The kinds of methods detailed in this chapter allow users of GIS packages to answer a wide range of questions, yet there is much more to spatial analysis, as the following chapters show.

Questions

1. Develop a simple entity-based model to analyse the effects of land use change annually.

2. Work out a GIS-based system for the optimum location of (a) fire stations, (b) banks, and (c) health care services in cities.
3. Explore the advantages and shortcomings of using entity-based models for ecological modelling.
4. Develop a GIS system for helping to manage the demand for building materials required for constructing a new suburb.
5. Explore the advantages of entity-based GIS in (a) real estate management, (b) hydrological modelling, and (c) archaeological site investigations.

Further reading

- Atkinson, P. M. and Tatnall, A. R. (1997). Introduction: neural networks in remote sensing. *International Journal of Remote Sensing*, 18: 699–709.
- Davis, J. C. (2002). *Statistics and Data Analysis in Geology*. 3rd edn. John Wiley & Sons, New York.
- Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge University Press, Cambridge.
- Lloyd, C. D. (2010). *Spatial Data Analysis: An Introduction for GIS Users*. Oxford University Press, Oxford.
- Malczewski, J. (1999). *GIS and Multicriteria Decision Analysis*. Wiley, New York.
- Wise, S. (2002). *GIS Basics*. Taylor & Francis, London.