Chapter 4: Basic GIS Functionality: querying, integrating and manipulating spatial data

4.1 Introduction

This chapter introduces the core functionality of GIS. It describes the first steps that a user can undertake once they have their data in a GIS software package. This is the last chapter that approaches GIS in this way, the remainder of the book focuses on issues rather than functionality. As described in Chapter 1, GIS software combines computer-mapping functionality that handles and displays spatial data, with database management system functionality to handle attribute data. This chapter describes the basic tools that this provides to the user that are not available in other types of software packages. The basic functionality is as follows:

- *Querying* both spatially and through attribute
- Manipulating the spatial component of the data: for example, through changing *projections*, *rubber sheeting* to join adjacent layers of data together, and calculating basic statistics such as areas and perimeters of polygons
- *Buffering* where all locations lying within a set distance of a feature or set of features are identified
- Data integration, either informally by simply laying one layer over another, or formally through a mathematical *overlay* operation
- Areal interpolation.

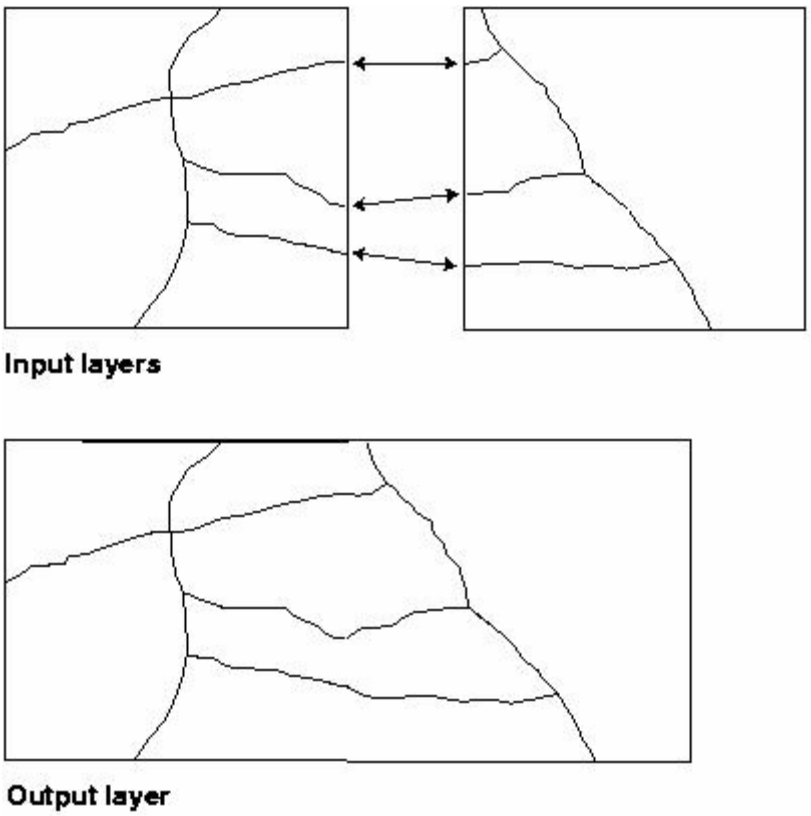
4.2 Querying data

As with all database systems, one of the core parts of GIS functionality is the ability to query the data. With GIS software there are two basic forms of querying: spatial and attribute. *Spatial querying* asks the question ‘what is at this location’? This is usually done by simply clicking on a feature and listing its attributes. More complex spatial queries could select all the features within a box or a polygon, or ask ‘what is near to this feature’? Queries of this type often require the use of buffering or overlay techniques as described later in the chapter. *Attribute querying* asks the question ‘where does this occur’? If a user has a layer consisting of the locations of churches with some information about each church, an attribute query could select all the churches whose denomination is Catholic and then draw them with a certain symbol. The user could then query the database to select all Protestant churches and draw these with a different

symbol to compare the patterns. There is much more on visualisation in Chapter 6; the purpose of introducing it here is simply to show how querying and mapping are inextricably linked.

4.3 Manipulating and measuring spatial data

Most GIS software packages come with a suite of options that allow the user to manipulate spatial data. One of the most basic of these is simply to change the *projection system* used. This can significantly alter the appearance of maps, particularly maps of the world, and can also make it possible to integrate data from layers that use different projection systems. In Britain, this could be used to take a variety of early maps on different projections and re-project them onto the National Grid. This would allow comparisons between different early maps, as well as with modern ones. Putting adjacent map sheets onto the same projection allows their digital representations to be joined to form a single layer. Where there are distortions to the sheets the maps will have to be *rubber-sheeted* (or "edge-matched") to ensure that the edges of the two sheets make a perfect join. This involves telling the software where certain key points are on the layer and where they actually should be. The entire layer will then be distorted using these references. Figure 4.1 shows an example of this.

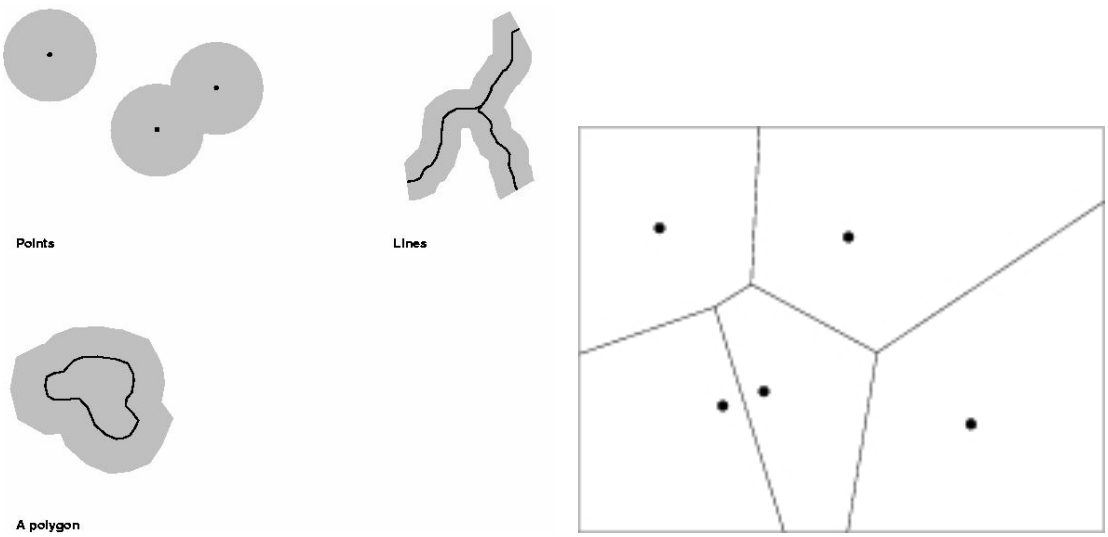


[Rubber sheeting to join two layers together]

Most GIS software packages will also calculate basic statistics about their spatial features. Typical examples include calculating the length of lines, the area and perimeter of polygons, and the distances between points. There are many examples of why these basic measures can be useful. These include measuring distances along a transport network, the use of areas to calculate population densities, and calculating the distances between settlements.

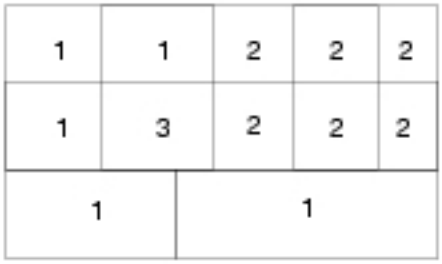
4.4 Buffering, Thiessen polygons and dissolving

There are times when rather than simply being interested in the locations of a type of feature, a user is interested in the locations within a set distance of a feature. Examples of this might include wanting to know all areas within (or outside of) 1km of a hospital, areas within 10km of a railway line, or within 5km of an urban area. Where information of this type is required a *buffering* operation is used. Buffering takes a point, line, or polygon layer as input and produces a polygon layer as its output, as shown in Figure 4.2.

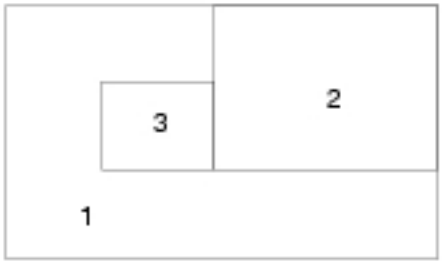


[Buffers around points, lines and a polygon and Thiessen polygons]

A user may also want to allocate catchment areas to a point dataset, and this can be done by generating *Thiessen polygons* (also known as Voronoi polygons). This creates a polygon layer in which the polygon boundaries are lines of equal distance between two points. This means that a polygon is the area that is nearest to the point that generated it, as is shown in Figure 4.3. This is a simple form of *interpolation*, whereby data are allocated from one set of spatial units to another.



Input layer



Output layer

[Dissolving to aggregate polygons]

Another option occurs where a user wants to create aggregate polygons from a more detailed layer. For example, a user might have a polygon layer where each polygon represents a farmer's field with attribute data that includes crop type. If the user is only interested in where particular crops are grown, then many field boundaries represent redundant information that can be removed. This is done by what is called a *dissolve* operation whereby the boundaries of adjacent polygons with the same crop type are removed to form aggregate polygons. This is shown in Figure 4.4.

4.5 Bringing data together to acquire knowledge

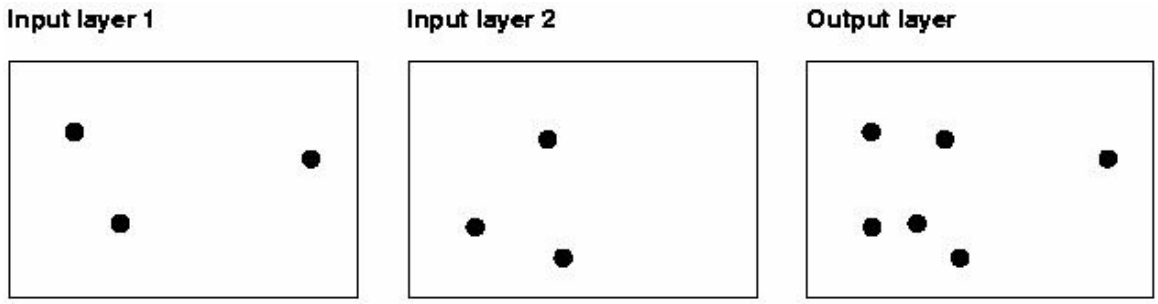
Manipulating the spatial component of a single layer of data is useful, but the full potential of GIS lies in its ability to integrate data from a variety of layers. At a basic level this merely involves combining layers on-screen to compare patterns. This might be as simple as taking a raster scan of a map and placing a vector layer over the top. The raster layer provides a spatial context for the features in the vector layer. Another option is to lay one vector layer on top of another; for example to compare the pattern of roads with the location of farms to see which farms lie near the major roads. Field boundaries might be a third layer added to this. This approach goes beyond basic mapping, as querying the underlying attribute database allows a detailed understanding of a multi-faceted study area to be developed. In this way an integrated understanding of the problem can be derived from many (possibly highly disparate) sources.

Healey and Stamp provide an example of this in their study of regional economic growth in the north-eastern United States (Healey and Stamp 2000). They have created a large and comprehensive database that contains industrial plants, such as blast furnaces, foundries and coalmines, represented by points. This is combined with polygon data showing the boundaries of natural resource deposits and, in areas of more detailed study, land parcels. To study the economic development of these they also needed a database of the transport system, and this is reproduced by layers of lines that include the railroad and canal networks, the turnpike roads and the rivers. This can be combined with aggregate, background information, such as that extracted from both the population and industrial censuses. Pearson and Collier (1998) use a similar approach in their study of agricultural productivity. As was described in Chapter 2, they combine environmental information in the form of raster grids and terrain models with information from the tithe surveys and census data represented by polygon layers. Siebert (2000) brings together information on

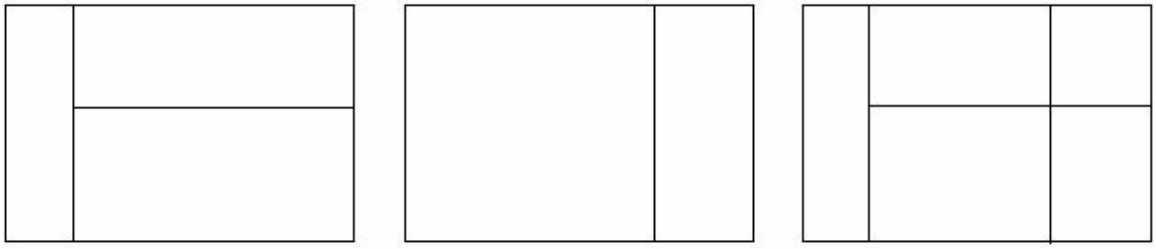
changing land-use, changing transport systems, administrative units, and population distribution to explore the urban development of Tokyo.

4.6 Formally integrating data through overlay

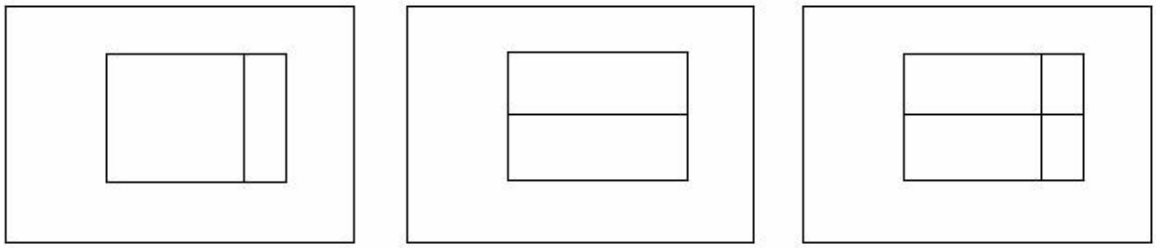
In addition to comparing layers visually, layers can be physically combined to produce new layers through geometric intersections. This is called *overlay*. Any of the three types of *vector data* can be overlaid with any of the others, as is shown in Figure 4.5. An overlay operation combines not only the spatial data but also the attribute data. This has many potential uses. For example, a user has a polygon layer containing data about administrative units, such as Irish counties, and wants to find out what proportion of each county was covered by water using a polygon layer showing lakes. An overlay operation would produce a new polygon layer that combined the attributes of both polygon layers, as shown in Figure 4.6, and thus each new polygon would have a combination of the county attributes and the attributes from the water layer. It is also possible to combine point or line layers with polygon layers using overlay. For example, as inputs we might have a point layer representing towns and a polygon layer representing counties and we want to determine which towns lay in which county. Overlaying the two layers would produce a point layer with the county polygon attributes added to each point, thus giving us the required information.



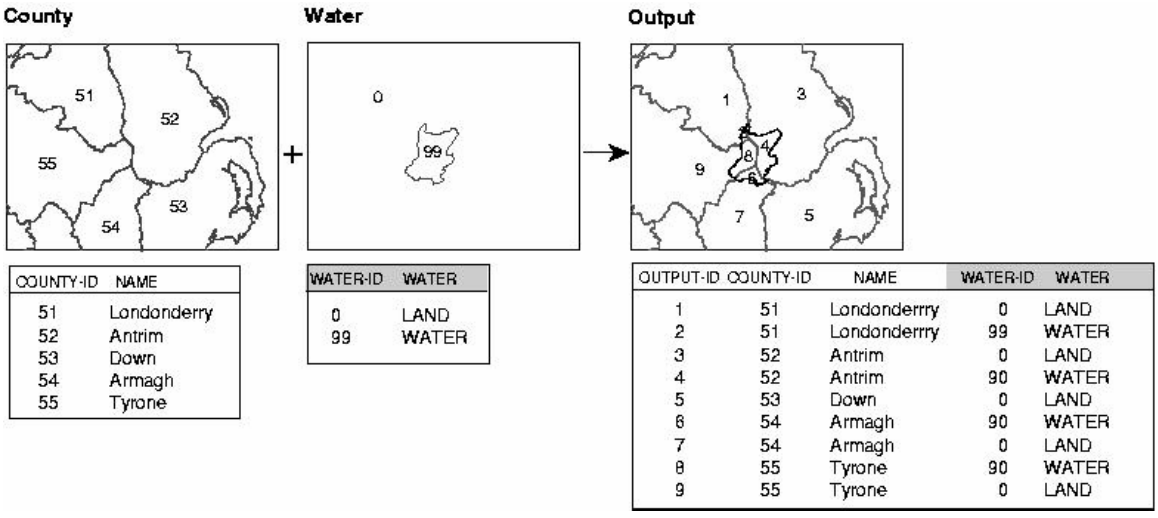
a. Points onto points



b. Lines onto lines



c. Polygons onto polygons



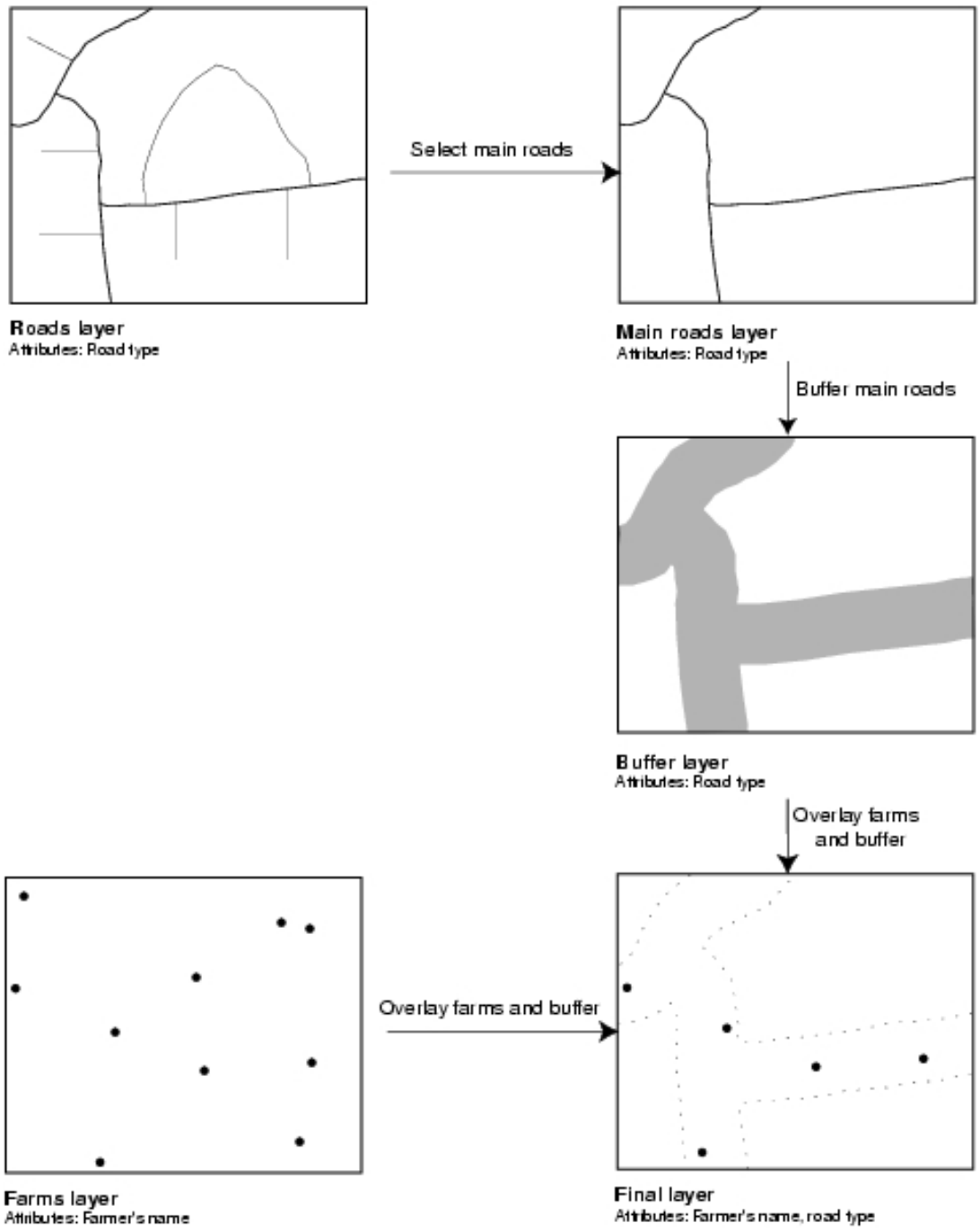
COUNTY-ID	NAME
51	Londonderry
52	Antrim
53	Down
54	Armagh
55	Tyrone

WATER-ID	WATER
0	LAND
99	WATER

OUTPUT-ID	COUNTY-ID	NAME	WATER-ID	WATER
1	51	Londonderrry	0	LAND
2	51	Londonderrry	99	WATER
3	52	Antrim	0	LAND
4	52	Antrim	90	WATER
5	53	Down	0	LAND
6	54	Armagh	90	WATER
7	54	Armagh	0	LAND
8	55	Tyrone	90	WATER
9	55	Tyrone	0	LAND

[Different types of overlay operations and Spatial and attribute data being combined using an overlay operation]

Combining buffering and overlay allows complex spatial queries and operations to be performed. For example, with a line layer showing the road network and a point layer containing farm locations, a user may want to calculate which farms lie within 1km of a major road. This can be done as shown in Figure 4.7. First, the user selects only the major roads from the road layer and copies these to a new line layer. A buffer is then placed around the new layer so that a polygon layer is created with polygons representing areas within 1km of a major road. If only farms within 1km of a major road are required, then a ‘cookie cutter’ overlay can be used. In this only farms on the input layer lying within a polygon on the buffer layer will be copied to the final layer. The final layer only contains five of the original farms and will contain all the attributes of both the farms and roads source layers.



[Spatial manipulation to solve problems]

Perhaps more than any other GIS operation, overlay tests the accuracy of the input layers. If a point layer is overlaid with a polygon layer then inaccurate polygon boundaries can easily lead to a point lying near a boundary being allocated to the wrong polygon. Accuracy is tested further where two polygon layers intersect. As was discussed

in Chapter 3, it is unlikely that any two operators will digitise a curve in exactly the same way even if the same source map is used. If this happens then the overlay operation will lead to *sliver polygons* being formed. These are very small polygons formed in the manner shown in Figure 4.8. This may seem like a trivial problem but is in fact the bane of vector overlay operations. It is possible to attempt to remove sliver polygons automatically. They tend to be long and thin and thus have a small area compared to the length of their perimeter. While this can be used to identify slivers, deleting them can still be problematic as it requires a decision about which boundary should be deleted. The problems caused by sliver polygons will depend on the scale and accuracy of the two sources, and the accuracy of the digitising. If the two layers have both been digitised to a high standard of accuracy from high quality source maps of similar scales, then the problems are likely to be minimal and can usually be solved by automated procedures within the software. If any of these three criteria are not met there is likely to be a significant job tidying the resulting output layer.

a. Input layer 1 b. Input layer 2 c. Output layer

[The creation of sliver polygons]

Overlay can also be performed on *raster data* providing they use the same *pixel* sizes. This is sometimes referred to as *map algebra* as two or more input layers are used to create an output layer whose cell values are calculated based on a mathematical operation between the input layers. An example of this is shown in Figure 4.9 where cell values on the two input layers are added to calculate values on the output layer. Other mathematical operations such as subtraction and multiplication can also be used, as can a wide range of other logical operations such as Boolean algebra.

Input 1

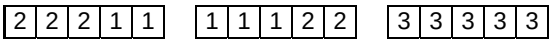
1	1	1	1	1
1	1	1	1	1
2	3	3	1	1
2	2	2	1	1

Input 2

0	0	0	2	2
0	1	3	2	2
0	1	3	2	2
1	1	2	2	2

Output

1	1	1	3	3
1	2	4	3	3
2	4	6	3	3
3	3	4	3	3



[Map algebra with raster data]

When two layers are combined using an overlay operation, the resulting layer will be at best as accurate as the less accurate layer. Unfortunately, the result is likely to be more inaccurate than this as error will be cumulatively added from both layers. This is termed *error propagation* and means that, as layers are combined, errors and uncertainty can multiply surprisingly quickly. This means that when multiple overlays are performed this must be done with the limitations of all the source layers being borne in mind.

An example of the use of overlay in historical research is provided by Lee (1996). The 19th century censuses of Ireland published data using baronies. These were relatively large spatial units and Lee wanted to estimate the internal population distribution of baronies in County Antrim to provide a more realistic representation of the population distribution. She believed that the distribution was likely to be affected by the presence or absence of large water features, altitude, and the proximity and function of nearby settlements. Her GIS consisted of:

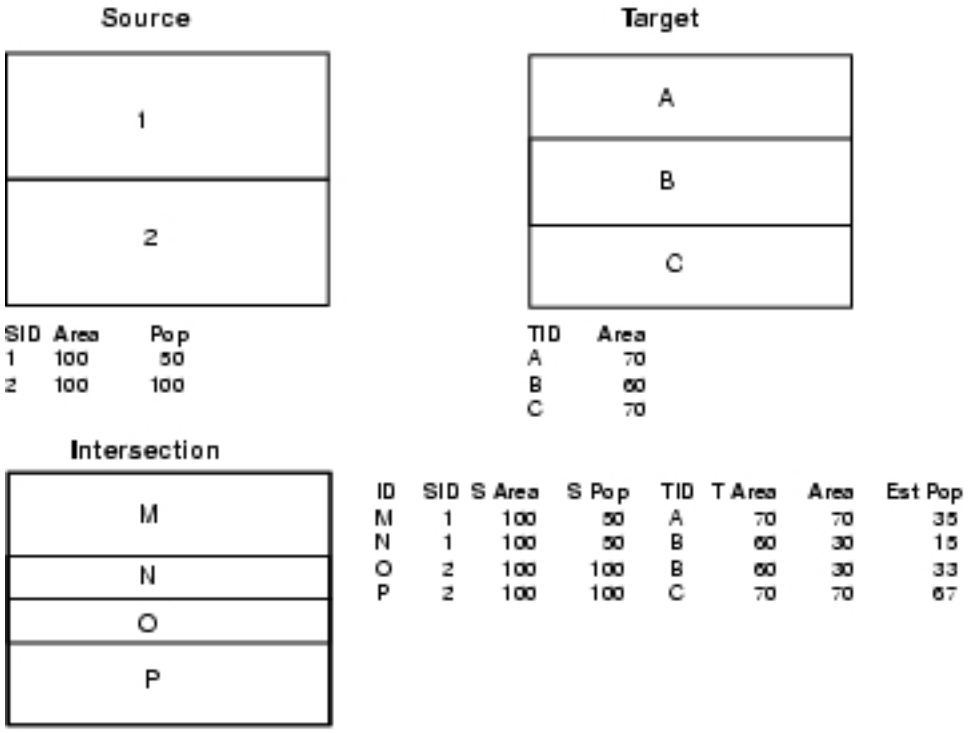
- A polygon layer of the 16 baronies in County Antrim
- A polygon layer of large water features
- A polygon layer showing altitude (for example, land less than 200 feet, 200 to 400 feet, etc.)
- A point layer showing the location of settlements with attribute data concerning their size and function.

The barony, water feature, and altitude layers were overlaid to produce an output layer with 263 polygons. She then overlaid the settlement layer onto the centroids of these polygons so that distances from each settlement to each of the 263 centroids could be calculated. Finally, she used a rather arbitrary model to allocate the barony populations to each of the derived polygons based on the barony population, whether the polygon was covered in water, the polygon's altitude, and the distance from the polygon's centroid to nearby settlements. This shows how integrating a variety of disparate datasets can be used to generate new datasets.

4.7 Integrating incompatible polygon data through areal interpolation

A final issue to be discussed in this chapter is *areal interpolation*. This commonly occurs where there are two or more polygon layers of socio-economic data and the

polygons represent administrative units, which a user wants to integrate. Where the two sets of polygons nest perfectly because the administrative units used were identical, this is a simple operation. Where they do not, for example if we are comparing census data published using registration districts with election data published for constituencies, then overlay does not provide the complete answer as it is uncertain how to allocate data to the resulting polygons. The traditional response to this is aggregation, which results in a loss of spatial detail, something that a GIS approach should attempt to avoid (see also Chapter 7).



[Areal interpolation]

Areal interpolation can be used instead. First we overlay the layer containing the input data onto the layer we want to estimate the populations for. These are termed the source and target layers respectively. The overlay generates the ‘zones of intersection’ between the two layers. The problem is then to estimate what proportion of the data from a source polygon to allocate to each zone of intersection. The simplest method of estimation is termed ‘areal weighting’. This is shown in Figure 4.10. The source and target units are overlaid to form the zones of intersection M, N, O and P. These polygons have all the attributes of the source and target polygons plus the areas of the new polygons as calculated by the overlay. The final column of attribute data, ‘Est Pop’, is

added by the user. Its values are estimated based on the area of the zone of intersection compared to the source polygon. Polygon N has an area of 30 while its source polygon, 1, had an area of 100. As a result 30% of polygon 1's population is allocated to polygon N giving 15 people. The final stage is to aggregate the newly estimated data to target zone level so we estimate that target polygon 2 has a population of 15+33=48.

The assumption of even population distribution is obviously extremely unrealistic. For example, registration districts in England and Wales usually consisted of a market town and its hinterland, hardly a likely candidate for an even population density. Various techniques have been devised to work away from this assumption, mainly by using further knowledge that allows us to estimate where within the source zones the population is likely to be concentrated. This type of functionality is rarely properly incorporated into GIS software, and if it is then the technique and its limitations are likely to be hidden from the user. This means that users are likely to have to implement the appropriate procedure for themselves.

An example of the use of areal interpolation is provided by Gregory *et al*. (2001). They wanted to compare three quantitative indicators of poverty: infant mortality, overcrowded housing, and unemployment, as they changed in England and Wales from the late-19th century to the late 20th. To do this they compared data from four time periods: the late 19th century, the 1930s, the 1950s and the 1990s. All the data were available as polygon data from the census or *Registrar General's Decennial Supplements*, but used significantly different reporting geographies. The late 19th century data were published using approximately 630 registration districts, while the two dates from the mid-20th century used approximately 1,500 local government districts (however, even these were difficult to compare as the system was extensively reformed between the two dates). Modern data were available at much more spatially detailed levels, with as many as 100,000 units. To allow direct comparison, all the data were interpolated onto the least spatially detailed units, the 630 registration districts. This resulted in the loss of significant amounts of spatial detail from the later data and also introduced some error to the results, but it did enable a consistent time-series to be generated that allowed them to compare the changing patterns of inequality over time at a geographically consistent scale.

4.8 Conclusions: information from spatially detailed, integrated databases

GIS software provides extensive functionality that allows a user to approach his or her dataset in a way that combines the spatial and attribute components of their data. This functionality leads to added value being extracted from an existing dataset. All datasets have limitations, and the extra functionality provided by GIS software allows us to use the data in ways that their creator would never have envisaged. As a result it is important to consider the limitations of all layers when manipulating them with the GIS. It is also important to consider the limitations of the techniques used on the data, particularly those that integrate data. As long as the results of spatial operations are understood within these limitations, GIS software provides new functionality that should allow new understanding to be derived from spatially referenced data.

In this chapter we have started to see the usefulness of the combined spatial and attribute data model used by GIS. This allows data to be queried and integrated in ways that no other approach can manage. The key advantage of this is that it allows the complexity of the data to be handled without undesirable simplification of the data.

Further reading from chapter 4:

References giving in **bold** are key references.

Querying data and basic manipulation of spatial data

See good basic GIS texts (Chapter 2) and map projections (Chapter 3)

Buffering, Thiessen polygons, overlay, and error propagation

Chrisman, N.R., 1990. The accuracy of map overlays: a reassessment. In: D.J. Peuquet and D.F. Marble, eds., *Introductory readings in Geographic Information Systems*. London: Taylor & Francis, 1990, 308-320

Chrisman, N.R., 1991. The error component in spatial data. In: D.J. Maguire, M.F. Goodchild and D.W. Rhind *Geographical Information Systems: principles and applications*. Longman: Harlow, 1991, 165-174. Available online at: http://www.wiley.co.uk/wileychi/gis/volumes.html

DeMers, M.N., 2002. *GIS modelling in raster*. Chichester: Wiley.

Environmental Systems Research Institute, 1997. *Understanding GIS: the Arc/Info method*. 4th edition. Cambridge: GeoInformation International.

Flowerdew, R., 1991. Spatial data integration. In: D.J. Maguire, M.F. Goodchild and D.W. Rhind, eds. *Geographical Information Systems: principles and applications. Volume 1: principles.* Longman: Harlow, 1991, 375-387. Available online at: http://www.wiley.co.uk/wileychi/gis/volumes.html

Goodchild, M.F. and Gopal, S., 1989, eds. *The accuracy of spatial databases.* London: Taylor & Francis.

Heuvelink, G.B.M., 1999. Propagation of error in spatial modelling with GIS. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, eds. *Geographical Information Systems: principals, techniques, management and applications.* 2nd edition. Chichester: John Wiley, 1999, 207-217

MacDougall, E., 1975. The accuracy of map overlays. *Landscape planning*, 2, 23-30

Shepherd, I.D.H., 1991. Information integration and GIS. In: D.J. Maguire, M.F. Goodchild and D.W. Rhind, eds. *Geographical Information Systems: Principles and applications. volume 1: principles.* Longman: Harlow, 1991, 337-360. Available online at: http://www.wiley.co.uk/wileychi/gis/volumes.html

Tomlin, C.D., 1991. Cartographic modelling. In: D.J. Maguire, M.F. Goodchild and D.W. Rhind, eds. *Geographical Information Systems: principles and applications. Volume 1: principles.* Longman: Harlow, 1991, 361-374. Available online at: http://www.wiley.co.uk/wileychi/gis/volumes.html

Unwin D.J., 1995. Geographic Information Systems and the problem of 'error and uncertainty.' *Progress in human geography*, 19, 549-558

Unwin, D., 1996. Integration through overlay analysis. In: M. Fischer, H.J. Scholten and D. Unwin, eds. *Spatial analytical perspectives on GIS*. London: Taylor & Francis, 1996, 129-138

Areal Interpolation

Flowerdew, R. and Green, M., 1994. Areal interpolation and types of data. In: A.S. Fotheringham and P.A. Rogerson, eds. *Spatial analysis and GIS*. London: Taylor & Francis, 1994, 121-145

Goodchild, M.F., Anselin, L. and Deichmann, U., 1993. A framework for the areal interpolation of socio-economic data. *Environment & planning A*, 25, 383-397

Goodchild, M.F. and Lam, N.S.-N., 1980 Areal interpolation: a variant of the traditional spatial problem. *Geo-processing*, 1, 297-312

Gregory, I.N., 2002. The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, environment and urban systems*, 26, 293-314

Langford, M., Maguire, D. and Unwin D.J., 1991. The areal interpolation problem: estimating population using remote sensing in a GIS framework. In: I. Masser and M. Blakemore, eds. *Handling Geographical Information: methodology and potential applications*. New York: Longman, 1991, 55-77

Historical case studies

Gregory I.N., Dorling D. and Southall H.R., 2001. A century of inequality in England and Wales using standardised geographical units. *Area*, 33, 297-311

Healey, R.G. and Stamp, T.R., 2000. Historical GIS as a foundation for the analysis of regional economic growth: theoretical, methodological, and practical issues. *Social science history*, 24, 575-612

Lee, J., 1996. Redistributing the population: GIS adds value to historical demography. *History and computing*, 8, 90-104

Pearson, A. and Collier, P., 1998. The integration and analysis of historical and environmental data using a Geographical Information System: landownership and agricultural productivity in Pembrokeshire c. 1850. *Agricultural history review*, 46, 162-176. See also Chapter 2.

Siebert, L., 2000. Using GIS to document, visualize, and interpret Tokyo's spatial history *Social science history*, 24, 537-574. See also Chapter 2.