PART A

The context for spatial data analysis

1

Spatial data analysis: scientific and policy context

Seen from the perspective of the scientist or the policy maker, analytical techniques are a means to an end: for the scientist the development of rigorous, scientifically based understanding of events and processes; for the policy maker the strategic and tactical deployment of resources informed by the application of scientific method and understanding. This chapter describes various areas that raise questions calling for the analysis of spatial data.

The chapter is organized as follows. Section 1.1 identifies how location and spatial relationships enter generically into scientific explanation and section 1.2 briefly discusses how they enter into questions in selected thematic areas of science and general scientific problem solving. Section 1.3 considers the ways in which geography and spatial relationships are important in the area of policy making. Section 1.4 gives some examples of how problems and misinterpretations can arise in analysing spatial data.

1.1 Spatial data analysis in science

All events have space and time co-ordinates attached to them – they happen somewhere at sometime. In many areas of experimental science, the exact spatial co-ordinates of where experiments are performed do not usually need to enter the database. Such information is not of any material importance in analysing the outcomes because all information relevant to the outcome is carried by the explanatory variables. The individual experiments are independent and any case indexing could, without loss of information relevant to explaining the outcomes, be exchanged across the set of cases.

The social and environmental sciences are observational not experimental sciences. Outcomes have to be taken as found and the researcher is not usually able to experiment with the levels of the explanatory variables nor to replicate. In subsequent attempts to model observed variation in the response variable, the design matrix of explanatory variables is often fixed both in terms of what variables have been measured and their levels. It follows that at later modelling stages model errors include not only the effects of measurement error and sampling error but also various forms of possible misspecification error.

In many areas of observational science, recording the place and time of individual events in the database will be important. First, the social sciences study processes in different types of places and spaces – the structure of places and spaces may influence the unfolding of social and economic processes; social and economic processes may in turn shape the structure of places and spaces. Schaeffer (1953) provides an early discussion of the importance of this type of theory in geography and Losch (1939) in economics. Second, recording where events have occurred means it becomes possible to link with data in other databases – for example linking postcoded or address-based health data and socio-economic data from the Census. A high degree of precision might be called for in recording location to ensure accurate linkage across databases.

Spatial data analysis has a role to play in supporting the search for scientific explanation. It also has a role to play in more general problem solving because observations in geographic space are dependent – observations that are geographically close together tend to be alike, and are more alike than those which are further apart. This is a generic property of geographic space that can be exploited in problem-solving situations such as spatial interpolation. However this same property of spatial dependence raises problems for the application of 'classical' statistical reference theory because data dependence induces data redundancy which affects the information content of a sample ('effective sample size').

1.1.1 Generic issues of place, context and space in scientific explanation

(a) Location as place and context

Location enters into scientific explanation when geographically defined areas are conceptualized as collections of a particular mix of attribute values. Ecological analysis is the analysis of spatially aggregated data where the object of study is the spatial unit. In other circumstances the object of study might comprise individuals or households. Analysis may then need to include not only individual-level characteristics but also area-level or ecological attributes that might impact on individual-level outcomes.

'Place' can be used to further scientific understanding by providing variability in explanatory variables. The diversity of places in terms of variable values consitutes a form of 'natural' laboratory. Consider the case of air pollution levels across a large region which contains many urban areas with contrasting economic bases and as a consequence measurable differences in levels and forms of air pollution. Data of this type combined with population data can be used for an ecological analysis of the relationship between levels of air pollution at the place of residence and the incidence of respiratory conditions in a population, controlling for the effects of possible 'confounders' (e.g. age, deprivation and lifestyle). The Harvard 'six cities' study used the variability in air pollution levels across six cities in the USA to examine the relationship between levels of fine particle matter in the atmosphere and the relative risk of disease (Dockery et al., 1993).

Explaining spatial variation needs to disentangle 'compositional' and 'contextual' influences. Geographical variations in disease rates may be due to differences between areas in the resident population in terms of say age and material well being (the compositional effect). Variation may also be due to differences between areas in terms of exposure to factors that might cause the particular disease or attributes of the areas that may have a direct or indirect effect on people's health (the contextual effect).

Contextual properties of geographical areas may be important in a number of areas of analysis. Variation in economic growth rates across a collection of regional economies may be explained in terms of the variation in types of firms and firm properties (the compositional effect). It may be due to the characteristics of the regions that comprise the environments within which the firms must operate (the contextual effect). Regional characteristics might include the tightness of regional labour markets, the nature of regional business networks, wider institutional support and the level of social capital as measured by levels of trust, solidarity and group formation within the region (Knack and Keefer, 1997). The contextual effect may operate at several scales or levels. Hedonic house price models include the price effects of neighbourhood quality and also the quality of *adjacent* neighbourhoods (Anas and Eum, 1984). Brooks-Gunn et al. (1993) in their study of adolescent development comment: 'individuals cannot be studied without consideration of the multiple ecological systems in which they [the adolescents] operate' (p. 354). The contextual effect of 'place' can operate at a hierarchy of scales from the immediate neighbourhood up to regional scales and above. Neighbourhoods influence behaviour, attitudes, values and opportunities and the authors review four theories about how neighbourhoods may affect child development. Contagion theory stresses the power of peer group influences to spread problem behaviour. Collective socialization theory emphasizes how neighbourhoods provide role models and monitor behaviour. Competition theory emphasizes the effects on child development of competing for scarce neighbourhood resources whilst relative deprivation theory stresses the effects on child development of individuals evaluating themselves against others. Pickett and Pearl (2001) provide a critical review of multilevel analyses that have examined how the socio-economic context provided by different types of neighbourhood, after controlling for individual level circumstances, can affect health outcomes. Jones and Duncan (1996) describe generic contextual effects in geography.

The introduction of 'place' raises the generic problem of how to handle scale effects. 'Place' can refer to areal objects of varying sizes – even within the same analysis. In most areas of the social sciences properties of areas are scaled up from data on individuals or smaller subareas (including point locations) by the arithmetic operation of averaging – that is by implicitly assuming additivity. This seems to be a consequence of the nature of area-level concepts in the social sciences (e.g. social cohesion, social capital and social control; material deprivation) which allows analysts to adopt any reasonable operational convention. In environmental science a similar form of change of scale problem arises in change of support problems where data measured on one support (e.g. point samples) are converted to another (e.g. a small area or block) through weighted averaging. But not all change of scale problems in environmental science are linear and can be handled in this way, as discussed for example in Chilès and Delfiner (1999, pp. 593–602) in the case of upscaling permeability measurements. There is detailed discussion of upscaling and downscaling problems and methods in environmental science in Bierkens et al. (2000).

(b) Location and spatial relationships

The second way location enters into scientific explanation is through the 'space' view. This emphasizes how objects are positioned with respect to one another and how this relative positioning may enter explicitly into explaining variability. This derives from the interactions between the different places that are a function of those spatial relationships. This generic conception of location as denoting the disposition of objects with respect to one another introduces relational considerations such as distance (and direction), gradient or neighbourhood and configuration or system-wide properties which may play a role in the explanation of attribute variability. The roles that these influences may play in any explanation are ultimately dependent on place attributes and in particular on the interactions that are generated as a consequence of these place attributes and their spatial distribution. We consider different ways spatial relationships construct or configure space: through *distance* separation, by generating *gradients* and by inducing an area-wide *spatial organization*.

Distance can be defined through different metrics – for example straight line physical distance, time distance (how long it takes to travel from *A* to *B*), cost distance, perceived distance. Distance can be defined in terms of networks of relationships and in qualitative terms: near to, far from, next to, etc. Distance becomes part of a scientific explanation when attribute variability across a set of areas is shown to be a consequence of how far areas are from a particular region that possesses what may be a critical level of some causal factor. The geography of economic underdevelopment reflects variation in levels of absolute disadvantage in terms of endowments, including lack of natural resources, poor land quality and disease. However it also appears to reflect distance from the core economic centres because distance affects prices and flows of new technology (Gallup et al. 1999; Venables, 1999). The incidence of cancer of the larynx might be linked to certain types of emissions and disease counts by area might be linked to distance from a particular noxious facility (Gatrell and Dunn, 1995). The measurement of distance might need to allow for such characteristics as prevailing wind direction and topographic attributes that could affect the direction of spread and amount of dilution of the emissions. In situations where outcomes are a product of interaction between individuals or groups then the level of an attribute in one area may influence (and be influenced by) levels of the same attribute in other nearby areas. High levels of an infectious disease in one area may through social contact and the greater risk of an infected individual contacting a non-infected individual lead to high levels in other nearby areas. Proximity also acts as a surrogate for the frequency with which individuals visit an area and become exposed to a highly localized causal agent. In various ways the relative proximity of areas, providing a surrogate for the intensity of different types of social contact, becomes integral to how geographic space becomes a consideration in accounting for the spatial variability of the incidence of the disease.

A gradient is a local property of a space, for example how similar or how different two neighbouring areas are in terms of variable characteristics. Measured surface water at a location after a rainstorm reflects not only the water retention characteristics of the location but also neighbourhood conditions that affect runoff levels and hence surface water accumulation rates. The economic gradient between two adjacent areas as measured by unemployment rates or average household income levels may influence crime rates, inducing an effect in both neighbourhoods that is not purely a consequence of the characteristics of the two respective neighbourhoods. Rather it reflects the fact that two areas of such contrasting economic circumstances are close together (Bowers and Hirschfield, 1999). Block (1979) remarked in the context of property crime: ‘it is clear that neighbourhoods in which poor and middle class families live in close proximity are likely to have higher crime rates than other neighbourhoods’ (p. 52). This was ascribed to a sharpened sense of frustration on the part of the have-nots combined with routine activity and opportunity theories that describe motivated offender behaviour. Johnstone (1978) encountered a similar neighbourhood effect in a study of adolescent delinquency.

The overall spatial organization of attributes of the study region may be important. In some instances the overall spatial distribution, how a totality of events in an area are distributed in relation to each other, may influence outcomes and overall, system-wide, properties. In the surface water example, levels of accumulation at a location will reflect not only local conditions and neighbourhood conditions but will also be affected by the overall configuration of wider system attributes such as the size, shape and topography of the catchment. Explanations of trading levels between two areas may be based not only on the economic characteristics of the two regions (which affects what they can supply and levels of demand) and their distance apart (which affects transport costs) but also on the nature of 'intervening opportunities' for trade. This can produce different levels of trade between pairs of regions that in terms of economic characteristics and distance apart are otherwise identical (Stouffer in Isard, 1960, p. 538). Faminow and Benson (1990) discuss how the spatial structure of markets changes the nature of tests for market integration.

Health may be related to social relativities rather than absolute standards of living (Wilkinson, 1996). The spatial distribution of material deprivation within a city, the extent to which deprived populations are spatially concentrated or scattered and thus experience different forms of relative rather than absolute deprivation may have an influence on the overall health statistics for a city (Gatrell, 1998). The geography of deprivation may influence the sorts of social comparisons people make. This in turn may influence their health via psychological factors and health-related behaviours (MacLeod et al., 1999). To what extent is persistent inter-generational poverty amongst certain ethnic groups in the USA a consequence of their spatial concentration in certain types of ghettos, spatially enlarged by processes of selective migration and characterized by high levels of poverty and long-term unemployment (Wilson, 1997)? Are areas with high levels of violent drug-related crime embedded in deprived areas of a city which are extensive enough to create special problems for policing (Craglia et al., 2000)?

The importance of spatial relativities in explaining attribute variation is scale dependent – that is the role of such relativities is dependent on the scale of the spatial unit through which events are observed and measured, in relation to the underlying processes. What may be a relational property in understanding why particular houses are burgled in an individual level analysis (for example, whether there are street lights outside the house or not) becomes a property of the place in an ecological analysis (quality of street lighting). If there is some crime displacement from areas where street lighting is good to neighbouring areas where street lighting is poor this will not be evident in the data if the spatial scale of the analysis is such as to average areas of contrasting street lighting or is larger than the scale at which any displacement effect occurs. Moving up the spatial scale of analysis, what may call for the inclusion of relational properties when analysis is in terms of urban census tracts may be analysed as a pure place effect at county or state levels of analysis. What will be an economic spillover of consumer expenditure from one area to another if the areas are small will be a local multiplier if the scale of the geographic areas exceed the scale of consumer travel behaviour. There may be neighbourhood effects in voting behaviour at the tract level as a result of interaction linked to 'the communication process, bandwagon effects, reference group behaviour, or other forms of "symbolic interactionism"' (Dow et al., 1982, p. 170). When comparing voting behaviours across larger regions, such effects are likely to become absorbed within the aggregate measure or become a contextual effect linked to variation in intra-area social interaction. At this scale other variables, such as socio-economic attributes, may assume greater significance.

1.1.2 Spatial processes

Certain processes, referred to as 'spatial processes' for short, operate in geographic space, and four generic types are now discussed: diffusion processes, processes involving exchange and transfer, interaction processes and dispersal processes.

A *diffusion* process is where some attribute is taken up by a population and, at any point in time, it is possible to specify which individuals (or areas) have the attribute and which do not. The mechanism by which the attribute spreads through the population depends on the attribute itself. Conscious or unconscious acquisition or adoption may depend on inter-personal contact, communication or the exerting of influence and pressure, as in the case of voting behaviour or the spread of political power (Doreian and Hummon, 1976; Johnston, 1986). In the case of an infectious disease, like influenza, the diffusion of the disease may be the result of contact between infected and non-infected but susceptible individuals or the dispersal of a virus as in the case of a disease like foot and mouth in livestock (Cliff et al., 1985). The density and spatial distribution of the population in relation to the scale at which the mechanism responsible for the spread operates will have an important influence on how the attribute diffuses and its rate of diffusion.

Urban and regional economies are bound together by processes of mutual commodity exchange and income transfer. Income earned in the production and sale of a commodity at one place may be spent on goods and services elsewhere. Through such processes of *exchange and transfer* the economic fortunes of different cities and regions become inter-linked. The binding together of local spatial economies through wage expenditure, sometimes called wage diffusion, and other 'spillover' effects may be reflected in the spatial structure of the level of per capita income (Haining, 1987).

A third type of process involves *interaction* in which outcomes at one location influence and are influenced by outcomes at other locations. The determination of prices at a set of retail outlets in an area may reflect a process of price action and reaction by retailers in that market. Whether retailer *A* responds to a price change by another (*B*) depends on the anticipated effect of that price shift on levels of demand at *A*. This may influence whether any price reaction at *A* needs to fully match the price shift at *B* or not. The closer the retail competitor at *B* is the more likely it is that *A* will need to respond in full (Haining, 1983; Plummer et al., 1998). Such interaction seems to be affected by the spatial distribution of sellers, including their density and clustering (Fik, 1988, 1991).

In a diffusion process the attribute spreads through a population and the individuals in the population have a fixed location. The final type of process, a *dispersal* process, represents the dispersal of the population itself. Such processes of dispersal may involve, for example, the dispersal of seeds from a parent plant or the spread of physical properties like atmospheric or maritime pollution or the spread of nutrients in a soil.

1.2 Place and space in specific areas of scientific explanation

The need for rigorous methods for spatial data analysis will be felt most strongly in those areas of thematic science where geographic space has entered directly into theorizing or theory construction. It will also be felt in areas of study where the identification of any regularities in spatial data is taken to signal something of substantive interest that justifies closer investigation. The next subsection discusses definitions and this is followed by a few brief examples.

1.2.1 Defining spatial subdisciplines

The recognition of the importance of location in a thematic discipline is signalled when subfields are defined prefixed with words such as 'geographical', 'spatial', 'environmental' or 'regional': geographical and environmental epidemiology (Elliott et al., 1992, 2000), spatial archaeology (Clarke, 1977) and spatial archaeometry, environmental criminology (Brantingham and Brantingham, 1991), regional economics (Richardson, 1970; Armstrong and

Taylor, 2000). Geography has systematic subfields which may overlap with the above with labels like: medical geography, historical geography, the geography of crime, economic geography. To the extent that there are real differences between these two approaches, geography as a synthetic discipline is often most interested in understanding particular places, drawing on the ideas and theories of the thematic disciplines (to which geographers themselves may contribute) in order to construct explanations or develop case studies. On the other side the thematic fields draw on place and space for the reasons discussed above – to develop understanding of the processes underlying disease incidence, pre-historic societies, the occurrence of crime and victimization, wealth creation.

Epidemiology distinguishes between geographical and environmental epidemiology. Geographical epidemiology focuses on the description of the geography of disease at different scales, ecological studies and the effects of migration on disease incidence (English, 1992). It is concerned with examining the factors associated with spatially varying levels of incidence, prevalence, mortality and recovery rates of a disease after controlling for age and sex. Environmental epidemiology seeks to model area-specific relative risk, after controlling for population characteristics and socio-economic confounders, arising from exposure to environmental risk factors such as naturally occurring radiation, air pollution or contaminated water. The study of geographical patterns and relationships help our understanding of the causes of disease, if not directly then at least by suggesting hypotheses that may then be pursued by other forms of investigation.

Swartz (2000) in his review defines environmental criminology as concerned with micro-level research which focuses on 'individual locations, and attempts to explain the relationship between site-specific physical features, social characteristics and crime' (p. 40). This is distinguished from the 'ecological tradition' in criminology which is 'confined to relatively large aggregations of people and space' (p. 40). Bottoms and Wiles (1997) use the term environmental criminology which they define as: 'the study of crime, criminality and victimisation as they relate, first to particular places, and secondly to the way that individuals and organisations shape their activities spatially and in so doing are in turn influenced by place-based or spatial factors' (p. 305). The term environment is used more broadly than in epidemiology and the definition allows for both the micro level and ecological levels of spatial analysis.

Clarke (1977) defines spatial archaeology as 'the retrieval of information from archaeological spatial relationships and the study of the spatial consequences of former hominid activity patterns within and between features and structures and their articulation within sites, site systems and their environments' (p. 9). Clarke identifies the key features of the subfield as the retrieval of useful archaeological information from the examination of the geography of archaeological data; the examination of archaeological data at a range of different geographical scales; the use of the map as a key tool in the process of extracting information. Hodder (1977) identifies the key stages of spatial analysis in archaeology as going from mapping, to the construction of summary descriptions of mapped distributions to the identification of map properties and local anomalies. Geo-coding data which have been collected from different field surveys and other disparate data sources provides a particularly useful way to link and cross check data sets. When combined with appropriate spatial analysis techniques this may assist with classification and the identification of heritage areas (for an example in the case of dialect studies see Wilhelm and Sander, 1998).

The field of regional economics as defined by Richardson (1970, p. 1) is concerned with the role of 'space, distance and regional differentiation' in economics. It has been broadly concerned with two classes of problem. Location theory focuses on explaining the location of economic activity and why particular activities are located where they are. The field of study originated with the work of Von Thunen who in the 19th century considered the problem of the location of agricultural production. This area of regional economics developed through the work of a succession of 19th-century and later theorists concerned principally with industrial location theory. The other main area of study is the regional economy and is concerned with explanations of economic growth at the regional scale, the causes of poor economic performance at the regional level and associated policy prescriptions.

Five areas of thematic science have been selected to illustrate the role of place and space within them.

1.2.2 Examples: selected research areas

(a) Environmental criminology

Early work in environmental criminology examined the links between urbanization, industrialization and crime and how and why different urban–industrial places generated different crime patterns. There is interest in the criminological implications of the shift towards the post-industrial city. The decline of traditional shopping areas and the changing nature of the inner city, the creation of new out-of-town shopping centres and new forms of residential housing with new forms of occupancy are generating new offence geographies (Bottoms and Wiles, 1997). Changes in the use of space within an urban area together with new patterns of mobility and new life styles are

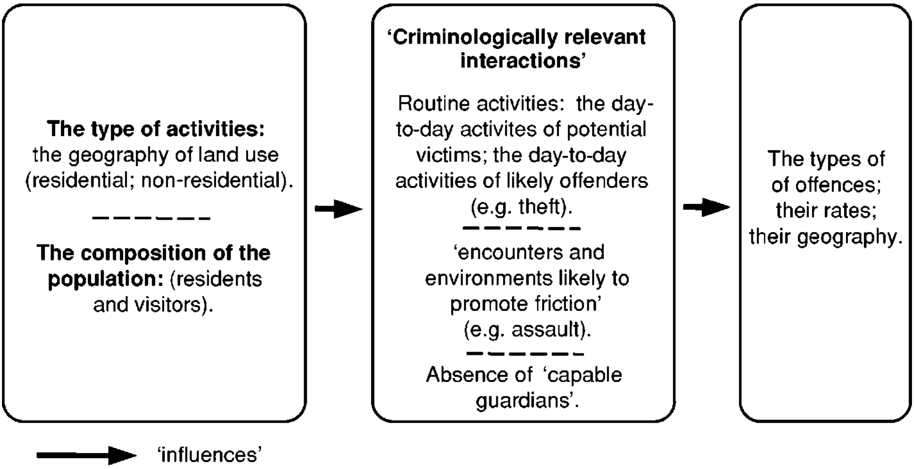

The type of activities: the geography of land use (residential; non-residential).
The composition of the population: (residents and visitors).
'Criminologically relevant interactions'
Routine activities: the day-to-day activites of potential victims; the day-to-day activities of likely offenders (e.g. theft).
'encounters and environments likely to promote friction' (e.g. assault).
Absence of 'capable guardians'.
The types of of offences; their rates; their geography.
'influences'

Figure 1.1 Wikström' s (1990) model for the geography of offences (after Bottoms and Wiles, 1997)

inducing changes in offence patterns and the emergence of new geographical concentrations of offences (Ceccato et al., 2002). Wikström's (1990) model of where offences occur is based on variations in land use within the urban area and the forms of social interaction taking place within the urban area (figure 1.1). Offences take place where criminal opportunities intersect with areas that are known to the offender because of their routine use of that space.

The lack of 'neighbourhood organization' or social cohesion or the co-existence of certain types of social organization and disorganization within a neighbourhood are possible explanatory variables in understanding where high offence rates occur and where offenders come from (Shaw and McKay, 1942; Bursik and Grasmick, 1993; Wilson, 1997; Hirschfield and Bowers, 1997). Explanations of where offenders come from often lay emphasis on area-level attributes and emphasize housing type and neighbourhood socialization processes that may provide too few sanctions on juvenile delinquent behaviour in certain areas (Bronfenbrenner, 1979; Martens, 1993). Wikström and Loeber (2000) identify neighbourhood socio-economic context as having a direct impact on the late onset of offending for certain groups of young offenders. Sampson et al. (1997) identify the role of collective efficacy, defined as a combination of social cohesion and a willingness for individuals to act on behalf of the common good to explain area-level variation in victimization rates.

Area-level contextual influences linked to social organization and processes of informal social control within neighbourhoods play a role in explanations of the causes of offending and victimization. These influences when analysed at an aggregate level are measured at the area level. Variables include:

socio-economic variables, demographic variables linked to family structure and residential mobility, ethnicity variables measuring the degree of ethnic heterogeneity found in an area and environmental variables linked to the nature and, in the case of housing, the density of the built form. By contrast there appear to be few analyses that introduce spatial processes or spatial relationships explicitly into explanations in criminology. Messner et al. (1999) suggest that the distribution of violent crime in the USA may be linked to the dynamics of youth gangs so that the geography of youth violence may be the expression of a spatially contagious process linked to social networks and other forms of communication. Cohen and Tita (1999) suggest that on the question of whether there is a diffusion process going on: 'the jury is still out' (p. 376).

Swartz (2000) distinguishes micro from ecological (or macro) traditions of analysis within criminology. A shift to the micro scale requires that socio-economic, demographic and environmental variables, which are still relevant at this scale of analysis, are defined appropriately. Now, for example, measures of the quality of the environment at the micro scale (e.g. in terms of lighting, street width, presence of cul-de-sacs) become important. And they are important in both the opportunities they may offer for crime and for the effect they may have on the formation of social networks. However, perhaps the more significant shift, in the context of spatial analysis, is that the city can now be treated in a more fragmented, local way. New modes of analysis assume particular importance such as the detection and investigation of crime 'hot spots' and other forms of localized patternings of offences, victims and offenders. As spatially fine-grained offence and victim data have become available through police recording systems so there has been an increase in micro-scale analyses in criminology.

(b) Geographical and environmental (spatial) epidemiology

Geographical studies examining disease variations do not generally shed an unambiguous light on the causes of disease because exposure to a risk factor and disease outcome are not measured on the same individuals. Environmental risk factors and neighbourhood contextual effects may have quite small impacts which are overwhelmed by individual circumstances or lifestyle factors. In ecological analyses, regression and correlation techniques are used to explore and test for relationships between attributes, dose levels and disease outcomes but it can be difficult to separate out compositional from contextual effects. International scale studies can often provide the most insight because differences on a global scale can be large, such as in the case of the link between exposure to sunlight and the incidence of rickets (English, 1992). The important consideration is not the geographic scale, per se, but whether there is adequate variation in the risk factor and the populations are sufficiently distinct in terms of their exposure levels (Lloyd, 1995).

In small-area studies exposure levels may be more homogeneous but the interpretation of geographical variation is made difficult by the effects of population movements and migration, the size of the population at risk and errors in population estimates (English, 1992). Area estimates of the level of a risk factor, such as an areal estimate of the level of air pollution, are used to impute levels of exposure experienced by individuals. This may be the most cost-effective way of examining the impact of environmental risk factors. Measuring exposure at an individual level is often both costly and potentially unreliable. Analyses can be strengthened by selecting subgroups of the population (such as by age or race) and by controlling for potentially confounding variables such as socioeconomic factors (Jolley et al., 1992).

However geographical studies can suggest causal hypotheses so that within epidemiology as a whole, geographical and environmental epidemiology represents a form of exploratory analysis (see chapters 6 and 7). Cuzick and Elliott (1992) classify the several types of small-area studies: investigations of clusters where there is no putative source of a risk factor; investigations of incidence rates around possible point sources of a risk factor of a given type; investigations of clustering as a general phenomenon; ecological studies; mapping disease rates. Epidemiologists appear to be divided on the value of these different small-area studies. The search for a sound methodology to undertake cluster detection has led to numerous techniques appearing in the medical and statistical literature, whilst at the same time drawing criticism that their contribution to establishing links between risk factors and health outcomes has been fairly limited. Swerdlow (1992) cites studies where the levels of raised incidence of malignant nasal conditions were traced to occupational hazards identified in areas of England with local boot and shoe and furniture making. Small-area studies may also be helpful in pointing to the specific source of an outbreak when the risk factors are understood, as in the case of an outbreak of toxoplasmosis in Greater Victoria, Canada which was linked through mapping to one reservoir in the water distribution system (Bowie et al., 1997).

The study of infectious diseases raises questions about the origins of an outbreak, how it develops through time, the geographical form and extent of its spread and the conditions under which a small outbreak may turn into a major epidemic in which a large proportion of the population becomes infected. Predicting the course and geographic spread of an infectious disease is critical to trying to control it, but each of the individual questions raises wider questions about the role of place and space, and these have influenced mathematical modelling and empirical investigations of infectious diseases (Bailey, 1975).

For example, certain characteristics of places have been identified as important in understanding the origins of an outbreak. In the case of common infectious diseases like measles, the origins of outbreaks have been linked to urban centres of sufficient size and in which the disease is endemic with an epidemic occurring when the conditions for spread are right (Bartlett, 1957, 1960).

The Hamer–Soper model is basic to deterministic and stochastic modelling of the course of an epidemic. Although there are important variants the focus is on transition rates (in the case of deterministic models for large populations) or transition probabilities (in the case of stochastic models for small populations) which are specified for each of the three states of an individual. *Susceptibles* are individuals not yet infected but who are members of the population at risk; *infecteds* are individuals with the disease and at a stage when they might pass it on to a susceptible; *removals* are individuals who have been vaccinated or had the disease and are no longer infectious nor susceptible. Early work assumed a population with homogeneous mixing so that all individuals were assumed to have the same-sized acquaintance and kinship circles. For example, the transition rate or probability for a susceptible to become infected in a given interval of time was modelled as proportional to the numbers of infecteds and susceptibles and the length of the time interval. From the set of transitions it was then possible to derive threshold conditions under which a small outbreak would become an epidemic (see Bailey, 1975 for a review).

The multi-region version of the Hamer–Soper model in Cliff et al. (1993, p. 363) was used to model measles outbreaks in Iceland and allowed homogeneous mixing within regions. However inter-regional transmission of the disease was the result of inhomogeneous mixing. Infection was passed from region j to region i through an inter-regional transition process. This process was a function of the number of susceptibles in region i and the number of infecteds in j, with a parameter that was modelled as an inverse function of the distance between the centroids of the regions.

The multi-region Hamer–Soper model limits the number of parameters to be estimated, by assuming that the inhomogeneous mixing between the N regions, rather than generate $N(N-1)$ parameters, is a function of distance so that only a single parameter needs to be estimated. Large numbers of parameters create problems for model estimation and inference, and the models could become still more complex if it becomes necessary to add more information to capture the internal characteristics of the regions. The model adopts a top–down approach to the analysis of complex systems, partitioning the study area into a pre-determined number of regions or zones.

Other modelling approaches have adopted a bottom–up approach representing the process in terms of a large population of individuals. In these models it is interaction at the micro level that defines the dynamics and the geography of the spread of the epidemic. An early example of this is Bailey (1967) who studied the spread of a disease on a lattice of individuals, each classified as either a susceptible, an infected or a removal, and where the spread starts from a single infected individual at the centre of the lattice. The model is a stochastic model of disease spread. At any given time, a susceptible only has a non-zero probability of becoming an infected if spatially adjacent to an infected. In a model of this type susceptibles change their states according to local (neighbourhood) transition rules. The model contains no mechanism for the infection to 'jump' and in particular there can be no transmission between spatially separated populations since there is no migration. This is an early example of the application of cellular automata theory (Couclelis, 1985; Phipps, 1989). System-wide properties emerge from micro-scale interactions. Bailey analysed the threshold conditions under which an outbreak would become a pandemic. He used a regular lattice for his simulations, but more complex spatial inhomogeneities can be incorporated through the spatial configuration of the population, as Hagerstrand (1967) employed in his models of innovation diffusion – an even earlier example of this type of modelling.

(c) Regional economics and the new economic geography

The subdiscipline of regional economics is positioned at the intersection of geography and economics and overlaps with the field of regional science. The nature of regional science and its original links with economics and geography can be gauged from Isard (1960). The current emphasis within both these areas of research can be judged from journals including the *Journal of Regional Science*, *Papers of the Regional Science Association*, *International Regional Science Review* and *Regional Studies*. The field of regional economics is principally concerned with regional problems and the analysis of economic activity at the subnational scale. Research in this area focuses on case studies and the mathematical modelling of economic growth at regional scales – models which have to reflect the different economic circumstances applying at the regional as opposed to the national scale (Armstrong and Taylor, 2000).

Early approaches to understanding regional growth differences focused on the role of the export sector and led to an approach to modelling based on predefined regions between which factors of production would move as well as flows of goods in response to levels of regional demand. Regional econometric and input–output modelling were characterized by a top–down approach in which inter- and intra-regional relationships were specified usually in terms of large numbers of parameters. One purpose was to develop regional forecasting models to track how economic change in particular sectors in particular regions would transmit effects to other sectors in other regions through the export sector.

A long-standing interest in regional economics is the extent to which there is convergence or divergence in per capita income growth rates between regions in the same market (Barro and Sala-i-Martin, 1995; Armstrong and Taylor, 2000). Inter-regional and inter-sectoral flows of labour and capital, responding to wage and profit differences, were seen as important in inducing convergence. However, migration of inputs, drawing on neo-classical arguments, is only one type of spatial mechanism that could induce convergence. Baumol (1994) identifies the role of technology transfers and the spatial feedback effects arising from productivity growth. In addition to these spatial mechanisms there are other geographical aspects to the modelling. These include the effects of spatial heterogeneity (regional differences in resource endowments, labour quality, local government and institutional policies) and the effects of local spillovers for example (Rey and Montouri, 1999; Rey, 2001; Moreno and Trehan, 1997; Conley, 1999).

Economists 'new economic geography' is concerned with regional growth and with understanding how the operation of the economy at regional scales affects national economic performance (Krugman, 1995; Porter, 1998) and trade (Krugman, 1991). This field, according to Krugman (1998), has served 'the important purpose of placing geographical analysis squarely in the economic mainstream' (p. 7), although its content and overall direction has drawn criticism from some economic geographers (Martin and Sunley, 1996; Martin, 1999).

Porter's theory, whilst not cast in formal terms, is concerned with the positive externalities (the contextual benefits) that a firm enjoys by being located where the environment confers competitive advantage on its operations. The theoretical underpinings to this advantage are captured in 'Porter's diamond', a conceptual model consisting of four components: factor conditions, demand conditions, firm strategy and the role of related and supporting industries. Geographical proximity strengthens and intensifies the interactions within the diamond and Porter (1998, p. 154) argues that competitive advantage accrues most effectively to a firm from a combination of the right system-wide (or national) conditions combined with intensely local conditions that foster industry clusters and geographical agglomerations.

A central feature of Krugman's modelling is the 'tug of war between forces that tend to promote geographical concentration and those that tend to oppose it – between "centripetal" and "centrifugal" forces' (Krugman, 1998 p. 8). The former includes external economies such as access to markets, and natural advantages. The latter includes external diseconomies such as congestion and pollution costs, land rents and immobile factors. Models are general equilibrium and spatial structure, for example an uneven distribution of economic activity across locations emerges from assumptions about market structure and the maximizing behaviour of individuals. At the centre of new economic geography models is a view of the space economy as a complex, self-organizing, adaptive structure: complex in the sense of large numbers of individual producers and consumers; self organizing through 'invisible-hand-of-the-market' processes; adaptive in the sense of consumers and producers responding to changes in, for example, tastes, lifestyles and technology. Where neo-classical theory is based on diminishing returns in which any spatial structure (such as the creation of rich and poor regions) is self cancelling (through convergence), the new economic geography is based on increasing returns from which spatial structure is an emergent property (Waldrop, 1992). Model outputs are characterized by bifurcations so that shifts from one spatial structure to another can result from smooth shifts in underlying parameters.

(d) Urban studies

Krugman's deterministic models appear to share common ground with multi-agent models used in urban modelling. In multi-agent models active autonomous agents interact *and* change location as well as their own attributes. Individuals are responding not only to local but also global or system-wide information. Again, spatial structure in the distribution of individuals is an emergent property, and multi-agent models, unlike those of the regional approach to urban modelling developed in the 1970s and 1980s, are not based on pre-defined zones and typically use far fewer parameters (Benenson, 1998).

These stochastic models have been used to simulate the residential behaviour of individuals in a city. They have evolved from cellular automata modelling approaches to urban structure (see section 1.2.2(b)). They describe a dynamic view of human interaction patterns and spatial behaviours that contrasts with the more static relational structures found in cellular automata theory (Benenson, 1998; Xie, 1996). In Benenson's model the probability of a household migrating is a function of the local economic tension or cognitive dissonance they experience at their current location. These tensions are measured by the difference between their economic status or their cultural identity and the average status of their neighbours. The probability of moving to any vacant house is a function of the new levels of economic tension or cognitive dissonance they would experience at the new location. If the household is forced to continue to occupy its current location, cultural identity can change.

A point of interest with both multi-agent and cellular automata models is how complex structures, and changes to those structures can arise from quite simple spatial processes and sparse parameterizations (White and Engelen, 1994; Portugali et al., 1994; Batty, 1998; Benenson, 1998). The inclusion of spatial interaction can lead to fundamentally different results on the existance and stability of equilibria that echo phase transition behaviour in some physical processes (Follmer, 1974; Haining, 1985). It is the possibility of producing spatial structure in new parsimonious ways (rather than assuming regional structures), together with the fact that the introduction of spatial relationships into familiar models can yield new and in some cases surprising insights, that underlies at least some of the current interest in space in certain areas of thematic social science. This interest, as Krugman (1998) for example points out, is underpinned by new areas of mathematics that make it possible to model these systems. In addition modern computers make it possible to simulate models that are not amenable to other forms of analysis.

Local-scale interactions between fixed elementary units, whether these are defined in terms of individuals or small areas, can affect both local properties and system-wide properties as illustrated by cellular automata theory. This effect is also demonstrated through certain models of intra-urban retailing where pricing at any site responds to pricing strategies at competitive neighbours. This can yield fundamentally different price geographies depending on the form of the profit objective and the spatial structure of the sites in relation to the choice sets of consumers (Sheppard et al., 1992; Haining et al., 1996). Multi-agent modelling adds another, system-wide level to the set of interactions, allowing individuals to migrate around the space and change type as a function of local circumstances, global conditions and local conditions in other parts of the region. However, all these forms of modelling raise questions about how model expectations should be compared with observed data for purposes of model validation. One aspect involves comparing the spatial structure generated by model simulations with observed spatial structures and this calls directly for methods of spatial data analysis (Cliff and Ord, 1981).

(e) Environmental sciences

Wegener (2000) provides a classification by application area of the large range of spatial models in environmental sciences drawing on Goodchild et al. (1993) and Fedra (1993). Atmospheric spatial modelling includes general circulation models and diffusion models for the dispersion of air-borne pollutants. Hydrological models includes surface water and ground water modelling. Land process spatial modelling includes models for surface phenomena such as plant growth or soil erosion and models for subsurface phenomena such as geological models and models of subsurface contamination (through waste disposal or infiltration). Biological and ecological spatial modelling includes vegetation and wildlife modelling – models of forest growth, fish-yield models, models for the spread of diseases through natural or farmed populations and models for the effect of resource extraction (like fishing) on stock levels. Finally the classification includes integrated models which involve combinations of the above groups such as atmospheric modelling and the transport of air-borne infectious diseases to livestock populations. To the earlier classification, Wegener adds environmental planning models such as those for noise in an urban area.

Space in environmental modelling is often continuous and spatial relationships are defined in terms of distance – either straight line or in terms of network structure as in the case of rivers in a catchment. Biological and ecological models of the spread of disease may introduce problems of short- and long-distance migration of the modelled populations. This needs to be accommodated in order to represent population mixing within the spatial model. Examples of different types of spatial modelling in the environmental sciences can be found for example in Goodchild et al. (1993) and Fotheringham and Wegener (2000).

Table 1.1 provides a summary of different ways place and space enter generically into the construction of scientific explanation in the examples cited in this section. The table identifies the different generic classes and selects an illustrative example for each. The two 'views' of geography are not mutually exclusive, as illustrated in the bottom row of the table.

1.2.3 Spatial data analysis in problem solving

There is a similarity to nearby attribute values in geographic space and Tobler (quoted for example in Longley et al., 2001) has referred to this property as the 'First Law of Geography'. Fisher (1935) noted in the context of designing agricultural field trials: 'patches in close proximity are commonly more alike, as judged by yield of crops, than those which are further apart' (p. 66). Processes that determine soil properties operate at many different spatial scales from large-scale earth movements that are responsible for the distribution of rock formations to the small-scale activities of earth worms. The consequence is a surface of values that displays spatial dependence in variable values and may contain different scales of spatial variation (Webster, 1985).

The same is true even when values represent aggregates with respect to an areal partition. That socio-economic characteristics tend to be similar between adjacent areas has often been noted (see Neprash, 1934, for an early observation of this). As Stephan (1934) remarked: 'data of geographic units are tied together . . . we know by virtue of their very social character, persons, groups

Table 1.1 *A summary of the generic treatment of geography in scientific explanation (see text for details)*

	Location as place and context			Location as relationships between places			
Classification	Individual level (individual units, e.g. people or households, as the objects of analysis).	Ecological level (spatial aggregates of individual units as the objects of analysis).		'Top–Down' inter-regional models.	'Bottom–Up' interaction models.		
	Relationship between individual-level response and individual-level characteristics, exposures and the contextual effect of area(s).	Relationship between a response and compositional effects and exposure effects.	Relationship between a response and compositional effects and spatial contextual effects.		Distance influences.	Neighbourhood (e.g. local gradient) influences.	Neighbourhood + system-wide (e.g. configuration) influences.
Examples	Relationship between individual experiences of victimization and personal characteristics, neighbourhood characteristics and higher-level spatial contextual influences.	Relationship between rates of a disease and environmental exposures after allowing for confounders, and compositional effects.	Relationship between regional economic growth rates and aggregate characteristics of firms and area measures of social capital.	(1) Hamer–Soper model of epidemics. (2) Regional econometric models. (3) Regional models of urban structure.	Disease incidence as a function of distance from a possible source of pollution.	(1) Cellular automata. (2) Differences in deprivation levels between adjacent areas as a factor in understanding crime rates.	(1) Multi-agent models. (2) Krugman models.
	Spatial variation in offender rates as a function of aggregate household attributes, local neighbourhood attributes (place and context) adjacent neighbourhood opportunities to offend (space). Spatial variation in uptake rates of a health service as a function of aggregate household attributes, neighbourhood attitudes (place and context) and physical access to service as a function of location (space).						

and their characteristics are inter-related and not independent' (p. 165). There is strong correlation between the adjacent pixels on a remotely sensed image of land cover. The fact that events close together in geographic space tend to be more alike than those further apart can be exploited to handle a number of scientific and technical problems. We briefly list some examples. In all cases a knowledge of the underlying spatial dependence can be exploited to tackle the problem.

[1] A researcher wants to sample a surface to estimate a parameter (such as the average level of a variable) to a pre-specified level of precision. The aim is to do so as efficiently as possible – that is by devising a sampling plan that will ensure the desired level of precision without taking an unnecessarily large sample size (Dunn and Harrison, 1993).

[2] Samples have been taken across a surface and the analyst wishes to interpolate the value at a location that has not been sampled. Or the analyst wants to draw a map of surface variation which also reflects the uncertainty associated with the different parts of the map as a consequence of the sampling plan adopted (Isaaks and Srivastava, 1989).

[3] A spatial database has been assembled for a set of variables. The database contains values that are missing 'at random', in the sense that there is no underlying reason (such as suppression for confidentiality) why the particular values are missing. The analyst wants to obtain estimates of the missing values (Griffith et al., 1989). A variant of this problem is as follows. The analyst wishes to use the same database to model the relationship between a response variable and a set of explanatory variables. Rather than discard every case for which there is one or more missing values which could seriously reduce the data set, it is possible to fit the model using all the collected data whilst making allowance for the missing values (Little and Rubin, 1987).

[4] A group of artefacts have been found as a result of archaeological field research. It is not clear which can be classified as belonging to the same type. Information on the attributes of the artefacts combined with information on the location where they were found may be used to provide a classification (Barcelo and Pallares, 1998).

The spatial nature of data, in particular the spatial dependence in the data, can be exploited to help solve technical problems of the sort described. One of the main problems is to represent that spatial dependence in order to use it in the estimation problem.

There is another group of problems where the 'replication' provided by having observations on many geographical areas or the fact that each area is embedded within a larger set of other areas can be exploited to yield solutions to certain problems.

[1] One role for ecological inference is to use ecological data to learn about the behaviour of individuals within aggregates. Suppose there are *n* spatial units each with a cross-tabulation for which only marginal totals (row and column sums) are known. The objective is to make inferences about the cells of each of the tables. This problem may arise for example analysing data on race (black or white) and voting behaviour (voted or did not vote), where the totals of each race voting are known as are the totals of who voted and who did not. The real interest however lies in the cells within the cross tabulations, for example the proportion of voting-age black people who vote in a given electoral area or precinct – values which are unknown (King, 1997).

[2] A survey has allowed estimates of the unemployment rate to be obtained for each of a number of small areas in a region. It is known that, whilst the small-area estimators that have been used are unbiased estimators of the small-area unemployment levels, they have low precision (high estimator variance) because of the small number of samples falling in each area. Low precision makes it difficult to detect real differences between the small areas. However, the estimator for the regional level of unemployment has a much higher precision, but as an estimator for any of the small-area levels of unemployment it will be biased.

In both of the problems cited, an approach lies through applying methods in which the small-area estimator 'borrows information' or 'borrows strength'. In the first problem a statistical model may be specified that draws on information from all the other *n* individual tables in the data set in order to estimate cell values. In the second, the small-area estimator can borrow information from the region-wide estimator. In some applications the procedure of borrowing information focuses on a geographically defined neighbourhood around the small area and spatial dependence is built into the estimation procedure (Mollie, 1996).

1.3 Spatial data analysis in the policy area

The policy maker is concerned with the strategic and tactical deployment of resources. A framework for such deployment is as follows: identify the areas of need according to specified criteria and set objectives, target the intervention, manage and monitor the intervention and evaluate the outcomes in relation to the objectives set. There are many forms of resource targetting and *geographical* targetting is one of them. Geographical targetting means directing resources at specific areas. These might be large areas such as regions entitled to bid for funds under different objectives within the European Union's Structural Funds programme or small areas such as local community-based initiatives taken by the police in partnership with local community groups to reduce crime in a neighbourhood. Projects have different time horizons from the short-term tactical to the long-term strategic.

Implementating a programme of intervention at the local or regional level initiates a process involving many different participants, requiring the sharing and integration of relevant data sets, including geographically referenced data sets, underpinned by relevant analytical tools. Where modelling is an element of any analysis, the inclusion of variables that can be manipulated by policy instruments is usually an important element in model specification.

'Intelligence-led' intervention is informed by different types of knowledge from understanding general processes operating at different spatial scales to highly localized knowledge of places and circumstances and the likely effectiveness of different courses of action. McGuire (2000) illustrates this from the perspective of the New York City Police Department's computerized crime statistics (COMPSTAT) process which has been credited with playing a significant role in crime reduction in New York in the mid to late 1990s. In the UK, the 1998 Crime and Disorder Act has placed a statutory requirement on police to undertake crime and disorder audits with local partners and to produce strategies based on this work. The focus is particularly on high-volume crime such as burglary and car theft. This Act together with the availability of geo-coded offence, offender and victim data have played a significant role in the growth of crime mapping and analysis (Hirschfield and Bowers, 2001). Gordon and Womersley (1997) discuss the ways GIS mapping capability and spatial analysis can support local health service planning.

The distinction between tactical and strategic deployment of resources is important for methodological reasons. Tactical deployment is often focused on a very narrow and specific set of objectives. This might include dealing with a sudden upsurge in street robberies in an area of a city or, in the case of a health authority, a sudden outbreak of a disease in a particular area or the identification of an area of abnormally raised incidence. Data sets underlying the formulation of a tactical response usually refer to short periods of time. There may be a need for rapid data collection followed by relevant data processing, perhaps 'hot spot' analysis, to support the case that something unusual is happening, where it is happening and to prioritize amongst competing demands (Craglia et al., 2000).

In the context of disease, 'cluster investigations, initiated in response to reports of apparent disease excess in a locality, are often demanded by public concern, but are difficult to interpret . . . a balance (has) to be struck between generating unwarranted public concerns and identifying genuine health risks as early as possible' (Wilkinson, 1998, p. 185). In the case of health, controversy may surround whether a cluster is statistically significant and whether it really does signal something of substantive significance that calls for special investigation or intervention – particularly when the number of cases is small and no cause can be identified. In such cases it may not be clear what action ought to be taken, if any. Wilkinson takes the view that the use of geographical 'surveillance' techniques – GIS-based systems for computing disease rates and applying statistical tests to look for clusters – will grow as the technology advances but that 'the interpretation of their output . . . requires expert judgement and considerable circumspection' (p. 186).

Strategic deployment of resources is based on long-term data series and on analyses that have identified if not causes then at least strong associations between, for example, socio-economic and environmental attributes, crime, disease or ill health. Strategic deployment in the case of a health or police authority may be associated with a (re)focusing of mission arising from what are seen as shortcomings in current levels of performance in relation to priorities. From this may follow decisions on implementation that result in new geographical patterns of strategic resource targetting that may distinguish between those elements of spatial variation due to compositional effects and those due to area-level contextual effects. The Acheson (1998) Enquiry into Inequalities in Health reviewed the 'evidence on inequalities in health in England . . . as a contribution to the development of the Government's strategy for health, to identify areas for policy development likely to reduce these inequalities' (p. xi). One of the recommendations was 'a review of data needs to monitor inequalities in health and their determinants at a national and local level' (p. 120). The recently completed ESRC Health Variations Programme focused on understanding different aspects of health variation in the UK, including geographical variation and the importance of place (ESRC, 2001).

In the UK the strategic deployment of resources for tackling social exclusion focuses on neighbourhood renewal and requires the geographical coordination of many small-area data sets (crime, health, education, housing, employment). Craglia et al. (2002) report an example of a children in need audit by enumeration district in Sheffield that involves bringing together many different local data sets. The next step is the co-ordination of the corresponding services with the aim of narrowing the gap between the most deprived areas and others (Minister for the Cabinet Office, 1999).

In the context of policing Swartz (2000) remarks that ecological or macro-level data analysis (see section 1.2.1) is likely to be of most interest to politicians who are concerned with the levels of strategic resource allocation across the different areas of the country. The link between deprivation and ill health has provided a justification for weighting resource allocation to Regional Health Authorities in England to reflect geographical variation in deprivation (see, e.g., Martin et al., 1994). There are dangers with ecological analyses (see chapter 4) and Fieldhouse and Tye (1996) warn that directing expenditures at areas with high levels of deprivation may not be targetting a high proportion of deprived people.

National politicians are unlikely to be interested in insights from highly localized, micro-level, spatial data analysis but those with responsibility for local problems and local-scale resource allocation will be (Swartz, 2000). Local government, grass roots organizations and neighbourhood associations as well as local police officials may find insights from micro-scale analyses helpful in identifying persistent high crime areas as well as helpful in pointing the way in terms of how resources might best be targetted. Once the areas are clearly delimited, 'the solution to the problem might be as simple as improving street lighting . . . or as complex as improving the living conditions of local residents' (Swartz, 2000, p. 44).

The technique of geographic profiling is 'a strategic information management system designed to support serial violent crime investigations' (Rossmo, 2000, p. 211). In contrast to 'hot spot' techniques that are used to analyse volume crimes like burglary and car theft at the local scale, geographic profiling is used to narrow down the range of possible anchor points (e.g. home address, work address) of a criminal engaged in serial rape or sexual assault. Geographic profiling uses a spatial analysis technique, informed by assumptions about offender activity spaces. The latter is based on the work of Brantingham and Brantingham (1991) and geographic profiling is applied to local data sets to produce 'a probability surface showing the likelihood of offender residence within the hunting area' (Rossmo, 2000, p. 197).

The neighbourhood is an important scale for strategic resource targetting. The neighbourhood environment may be important in influencing or even shaping individual, preventative health behaviours and attitudes towards health (Sooman and Macintyre, 1995). Neighbourhood characteristics (social, material, physical) are likely to be important in determining the extent to which elderly members of a population are able to cope independently.

Examples of the need to be sensitive to individual-level characteristics and neighbourhood characteristics are not limited to health interventions. Wikström and Loeber (2000) note in the context of preventative strategies aimed at reducing youth offending that it is important to base strategy on 'knowledge about the interaction of "kinds of individuals" in "kinds of contexts"' which 'have higher potentials to be effective than strategies that pinpoint either the individual or the context' (p. 1111). Strategy aimed at addressing problems of disadvantage needs to separate out those aspects of an underlying problem that relate to individual circumstances and which might be addressed using individually targetted support from those aspects that relate to area-level group effects and which should be addressed by area-level programmes.

The case for localized geographical targetting is strengthened when micro-level data analysis is underpinned by an understanding of likely cause. The link between coronary heart disease and certain types of diet is well established so that when areas of high incidence have been identified within a city it is often clear what needs to be included in any form of intervention. In other circumstances where specific causes may not be understood, such as in the case of low uptake of a screening programme, if 'cold spot' analysis identifies such areas in a city, it may be necessary to undertake individual-level surveys within the areas to try to discover why.

1.4 Some examples of problems that arise in analysing spatial data

This final section provides a few examples where attention paid to the spatial nature of the problem may guard against drawing unwarrented conclusions or following inefficient procedures. The properties of spatial data have an impact on many aspects of data analysis from description and exploration of data sets through to modelling.

1.4.1 Description and map interpretation

Before drawing conclusions from mapped data about geographical variation it is important to try to ensure that values are 'equally robust'. Some elements of spatial variability may be an artefact of the data arising from errors that have propagated through a data set as a result of carrying out arithmetic or logical operations on data contaminated by error (Arbia et al., 1999). It may have arisen as a consequence of the small number problem (Gelman and Price, 1999). In mapping disease rates by area, sample sizes may be small, sampling variability large so that structures in the data, particularly extreme rates but also spatial trends, may be a statistical artefact. As fine-grained spatial data becomes more readily available the risk of making this type of error increases. Mollie (1996) provides an illustration using cancer data for French *departements* of how the most extreme rates tend to be associated with areas with the smallest populations.

1.4.2 Information redundancy

The presence of spatial dependence means the information content of a sample used to estimate a population parameter is less than would be the case if the *n* observations were independent. In the terminology of Clifford and Richardson (1985) the 'effective' sample size is less than the number of cases sampled because data points near to one another carry 'duplicate' information about the parameter. Statistical testing needs to recognize the degrees of freedom that are actually available for carrying out tests using spatial data where observations are not independent.

Because data points close to one another carry duplicate information this needs to be recognized in designing sampling plans (Dunn and Harrison, 1993). This feature of spatial data as well as the specific configuration of the observed data points needs to be recognized in spatial interpolation (Isaaks and Srivastava, 1989).

1.4.3 Modelling

A regression model may provide a good fit but a map of the residuals (the differences between the observed values of the response and the model predictions or fits) may show clear evidence of spatial structure that is confirmed by a test statistic. This violates one of the assumptions of regression which undermines the validity of inferences drawn from the model. In addition, notwithstanding the goodness of fit of the model, such a result suggests the model can be improved perhaps by inclusion of new covariates in the model specification.

1.5 Concluding remarks

This chapter has considered the importance of spatial data analysis in a number of different areas, distinguishing between the role of spatial data analysis as a tool of science and as a tool of the policy maker. Spatial data analysis is concerned with studying patterns in variables and associations between variables but in the absence of time series data it is usually not possible to identify causal relationships; this means neither the components of any causal system nor the directions of causation in those cases where there is ambiguity about the direction of causation. At best spatial analysis can point to possible causal relationships which can then be followed up by other methods.

Large quantities of fine-grained geographically referenced data are available, and can be linked on a common spatial reference with the help of a geographic information system. Geographic variability can today be mapped and analysed at a level of detail and spatial extent that in the past was difficult to undertake. The availability of such geo-coded data provides an opportunity to construct new views of old problems and to explore views of new spatial data prompting new understanding and new insights. Such a claim rests at least in part on the quality of the data and later chapters (particularly chapter 2) will discuss issues of data quality. However at this point it is worth remarking on two further points. First, fine-grained geo-coded data means precise information on location but clearly the locational reference must be relevant to the area of enquiry. Place of residence is important in analysing household burglary patterns but place of residence may be less useful in analysing geographies of chronic diseases with long latency times. Second, the greater the level of spatial detail the higher the level of noise there is likely to be in the data and the greater the need to draw on methods that distinguish between ‘noise’ and ‘signal’ in pattern analysis and in the analysis of relationships.

2

The nature of spatial data

This chapter discusses the nature of spatial data. All the analytical techniques in this book use a space-(time-) attribute data matrix (see the introduction). This matrix is the end product of a process of construction that starts from a conceptualization of geographical reality. What is it necessary to know about the relationship between that reality and the data matrix as a representation of that reality so far as the conduct of analysis is concerned and the interpretation of results?

There are a number of steps involved in specifying this relationship (see, e.g., Longley et al., 2001; Mark, 1999). First there is a process of *conceptualizing* the real world. This process extends to the identification of those fundamental properties that are relevant to the application. Such fundamental properties relate both to entities ('things in the real world') and the spatial relationships between entities. In the context of spatial data analysis spatial dependence in attribute values is considered as a fundamental property. Second, a data matrix acquires properties that may distance it from the real world as a consequence of *representational* choices. These are the decisions made about what to include in the data matrix and in what form, usually for the purpose of storing the data in a computer. Decisions must be taken, for example on spatial scale or level of spatial aggregation and the geometric class (points, lines, areas or surfaces) used to represent geographical entities. Third, when attributes and spatial properties are measured, this introduces another source of uncertainty between the real world and what is contained in the data matrix. Inaccuracies in the *measurement process* is an important source of uncertainty that affects the spatial data matrix.

Section 2.1 considers conceptualization and representation issues as they relate both to geographic space and the attributes captured in a data matrix. There is discussion of the nature of spatial dependence. Section 2.2 defines the data matrix, and makes observations on the relationship between it and the geographic reality it is attempting to represent. Section 2.3 considers different dimensions of data quality. The implications of data quality for the conduct of spatial data analysis are discussed in chapter 4.

The final section 2.4, describes methods for obtaining quantitative measures of spatial dependence. Although spatial dependence is a fundamental property it is important to see it as one of many data properties (some inherited from the chosen representation rather than fundamental) to be thought about in analysing spatial data. Discussions about the importance of spatial dependence and the methods of section 2.4 need to be seen in the context of broader issues that are raised particularly in sections 2.2 and 2.3.

2.1 The spatial data matrix: conceptualization and representation issues

2.1.1 Geographic space: objects, fields and geometric representations

Modelling geographic reality means the process of capturing the complexity of the real world in a finite representation so that digital storage is possible. This abstracting of a 'real, continuous and complex geographic variation' (Goodchild, 1989, p. 108) into a finite number of discrete 'bits' involves processes that include generalization and simplification. *Objects* and *fields* represent two fundamental conceptualizations of the entities that comprise geographic reality (Goodchild, 1989; Salgé, 1995; Longley et al., 2001). The difference is most easily expressed through examples. Variables such as temperature, snow depth or height above sea level are appropriately conceptualized as fields. A house (point), road (line) or political unit (area) are usually conceptualized as objects. Objects refer to things in the world whilst a field refers to a single valued function of location in two-dimensional space (see figure 2.1). Usually one or other of these two accord better with our mental perception of the real world and may also provide a better basis for efficient computation (Mark, 1999).

Four classes of digital objects for representing geographic phenomena are usually identified – *points, lines, areas* and *surfaces* (as contour lines for example). Object space is represented digitally by points, lines or areas. A town may be represented as an area (using its administrative boundary to delimit the area) or at another scale of representation as a point. As an area its representation may be refined using census tract-level data (e.g. wards or enumeration districts). Each enumeration district may be represented as an area object using the administrative boundary or as a point object by identifying the area or population-weighted centroid. The population of the town can be

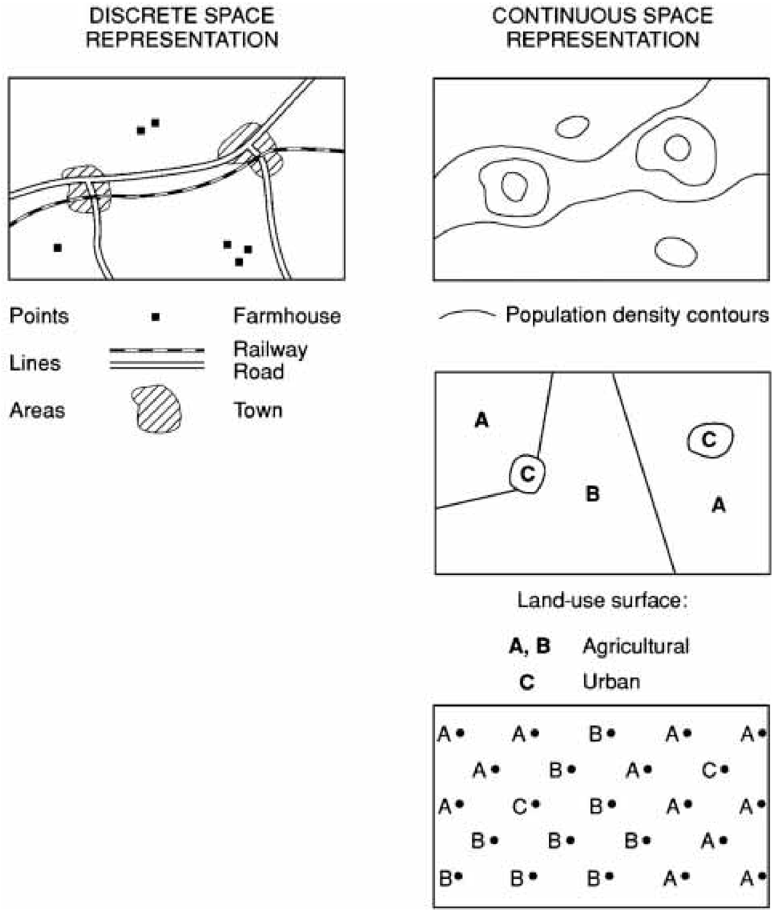


Figure 2.1 Discrete and continuous space representations

represented as address point objects at the scale of individual households. It follows that census tracts represent a *spatial aggregation* of these fundamental entities. A town or a forest are represented as area objects and a boundary line drawn, even though in reality the boundary may be ambiguous and 'fuzzy'.

In the case of a field, data values associated with attributes are possible at each of an infinite number of point locations on the surface. Storing data about a field in a data matrix requires it to be made finite. The field as a surface can be represented by using contour lines. Representing a field using areas often means dividing the region into small regular spatial units called *pixels*. Pixel size specifies the *spatial resolution* of the representation and is the field equivalent of spatial aggregation. To represent a field by points means choosing sample locations. A point measure may be literally sufficient as in the case of a measure of soil or snow depth. In cases like air pollution any measure is a function of the size of block ('support') used to define the quantity.

In the case of area objects and area representations of a field, either the areas are defined independently of data values as in the case of census tracts and image pixels or their boundaries reflect a change in data values. In the first case the areas are said to be *intrinsic*. In the second case the areal partition is imposed *after* analysing data values and the partition defines homogeneous (or quasi-homogeneous) areas or *regions*. Fields can be *segmented* into blocks of pixels with the same or similar values. Census tracts can sometimes be *aggregated* into larger groupings of contiguous tracts that are similar at least with respect to a small number of variables.

There are situations, however, where there is a choice of conceptualization. Population distribution can be conceptualized as object or field. If conceptualized in terms of objects then the representation may be in the form of points (e.g. by residence) or counts by regular areas (pixels) or irregular areas (e.g. census tracts). If conceptualized in terms of a field then the representation may be in the form of a density surface, or by spatially distributing population counts using kernel density smoothing or by interpolation. Bithell (1990) and Kelsall and Diggle (1995) construct relative risk surfaces for disease using kernel density methods to convert population count and disease count data to density surfaces (see chapter 7).

The implications of these representational choices are considered in section 2.3. For further perspectives on spatial representational issues see Raper (1999).

2.1.2 Geographic space: spatial dependence in attribute values

The presence of spatial dependence means that values for the same attribute measured at locations that are near to one another tend to be similar, and tend to be more similar than values separated by larger distances. By 'similar' is meant that if an attribute value is large (small) then nearby values of the same attribute will tend to be large (small). This characteristic has parallels with time series data. Values for the same variable close in time tend to be similar and more similar than values separated by longer time periods. The nature of this similarity may be independent of where (in space) or when (in time) values are measured. For spatial data this implies, no matter where one looks on the map, that the nature of that similarity is the same. The dependency structure in this case is said to be *stationary*. By contrast, if the structure of dependency varies across the map so that any measure of similarity depends on which part of the map is analysed, it is said to be *non-stationary* or that the structure of dependency is *heterogeneous*.

However there are important differences between space and time in respect of dependency which is why different spatial statistical techniques are needed to quantify and analyse spatial as opposed to temporal data (see section 2.4). It is also the reason why different statistical models are needed to describe spatial as opposed to temporal variation (see chapters 5 and 9). First, time has a unidirectional flow. The past may influence the present but the future can only influence the present in the sense of an expectation of that future, not in the sense of an actual realization of that future. Space has no equivalent to the trilogy of past, present and future. Second, spatial dependency is complicated by extending over two dimensions, not one, and because the structure of that dependency need not be the same along the two axes (north/south; east/west). If the dependency structure is the same on both axes it is called an *isotropic* dependency structure, if it is not it is called *non-isotropic* or *anisotropic*. Finally, periodicity which is often encountered in time series data (seasonal effects, business cycle effects; daily and weekly effects) is not often encountered in spatial data.

2.1.3 Variables

Attribute characteristics can refer to the spatial objects themselves or to entities that are associated with or attached to the spatial objects but not directly dependent on them. Attribute characteristics that refer to spatial objects such as the size or spatial extent of an area object raise conceptual and representation issues – what is the length of a coastline or the area of a forest? Attribute characteristics attached to spatial objects such as the number of cases of an offence, the number of plant species are also subject to conceptualization and representation issues. Conceptualization refers to the definition and meaning of the attribute (an offence, a disease, an economic sector, deprivation). Representation refers to how an *attribute* is operationalized into *variables* for the purpose of acquiring and storing data on the attribute (e.g. how deprivation is measured) and to enable analysis to be undertaken. Analysis is undertaken on data collected with respect to one or more variables that measure attributes associated with geographic reality that is typically represented in the form of spatial objects.

Conceptualization and representation issues of attributes are specific to each particular application. Two generic issues that can be discussed here relate to or have implications for attribute representation and data analysis. The first is how variables are classified and the second is the level of measurement of a variable. Classification identifies the place of each variable in an analysis; level of measurement defines what arithmetical operations are permissible.

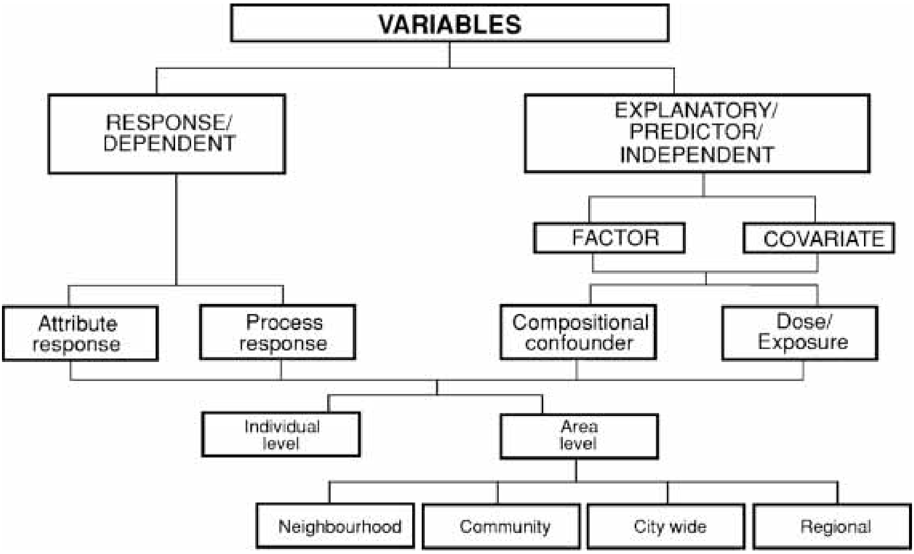


Figure 2.2 Classification of variables

(a) Classifying variables

Depending on the context, variables are divided into groups and given different labels (see figure 2.2). In the case of data modelling, Y represents the variable whose variation is to be explained. In multiple regression modelling where the purpose is to explain variation in Y in terms of other variables, Y is called the *dependent* or *response* variable and the other variables are called *independent* or *explanatory* or *predictor* variables ($X_1, X_2, \ldots$). If a predictor variable is measured at the nominal or ordinal level it is often called a *factor* and if measured at the interval or ratio level it is often called a *covariate* (see below for the definition of these level of measurement terms).

When modelling, it is useful to distinguish between variables in a further sense. Some variables measure *individual-level attributes*, such as the age or sex of an individual, or the number of bedrooms or floor space in a house. In some forms of spatial modelling these quantites may be aggregated over all individuals (people, houses) within the area but they can still be referred back to the attributes of individuals located within the spatial unit. These types of variables can be referred to as measuring *aggregated individual-level attributes* – for example the proportion of the resident population in the age group 16–64. These can be used to make an assessment of compositional effects on a response that refers to a spatial aggregate.

Attributes can also be attached to areas to refer to area-level or group-level properties. The Townsend index of material deprivation and the DETR index of local need are area-level measures of deprivation (Townsend et al., 1988; DETR, 2000). An individual, who may or may not be individually deprived as a consequence of a particular combination of personal or household characteristics, might be said to be *exposed to* or in their daily life *experience* the level of deprivation in the area where they live. Social capital, social cohesion and social control or guardianship are attributes defined at the area level to capture economic, sociological or power structure attributes of areas (e.g. Putnam, 1993). Individuals might be said to be *exposed* to or *experience* the level of social cohesion present in the area where they live (Hirschfield and Bowers, 1997). Collective efficacy, defined as 'social cohesion among neighbours combined with their willingness to intervene on behalf of the common good' (Sampson et al., 1997) is a neighbourhood-level attribute that in combination with aggregate demographic characteristics of individuals may help to explain geographical variations in certain types of offending. In the context of spatial modelling, variables measuring such attributes are measuring *area-level contextual attributes*. Such attributes may exist at different levels or scales ranging from the neighbourhood or ED level, to the level of the ward or some grouping of EDs through to city-wide, regional and higher spatial scales. Typically variables that measure such attributes, whilst they might be constructed from variables measuring individual-level attributes (derived from the census), involve combining variables into an index which is meant to quantify (or operationalize) a group-level concept.

In environmental epidemiology, covariates may be included that measure exposure to possible environmental risk factors – such as particulate matter in the incidence of respiratory conditions. These variables are called *exposure* or *dose* variables. Individual-level attributes (e.g. age and sex composition of area populations) must be accounted for in explaining spatial variation in disease incidence because of area variation in demographic groups (compositional effects). There are other types of covariates which must be accounted for. The term *confounder* variable is used to denote a variable that may also influence the level of a response variable (e.g. the rate of some disease) and which if not allowed for may obscure the true nature of a dose–response relationship. Deprivation is a confounder variable in the relationship between exposure to air pollution and the rate of a respiratory condition. This is because deprived populations often live in areas that suffer from higher rates of air pollution and deprived populations may suffer from higher rates of respiratory problems for reasons not directly linked to air quality (such as poor housing or poor diet). Cigarette smoking levels should also be considered as a confounder in such an analysis, and to have a more direct association than deprivation with respiratory health.

A regression model can be classified as either an *attribute–response* or a *process–response* model. This classification derives from the nature of the model and whether explanatory variables relate to underlying processes or not. For example, a regression model may be used to explain variation in house prices across a region in terms of other individual- and area-level attributes including housing characteristics and area characteristics. In this case the attribute–response model is analysing relationships between attributes and is not constructed in terms of the processes responsible for determining house prices, which presumably should include market processes and the maximizing behaviours of buyers and sellers. In a process–response model the explanatory variables include variables that link directly to the underlying process mechanisms. The previously cited epidemiological models that include dose or exposure variables fall into this category.

(b) Levels of measurement

Specifying the level of measurement of a variable is important because it specifies the formal properties of the number system underlying the measurement and determines what arithmetic operations are valid and hence what statistical procedures can be employed. In the case of nominal data, cases can only be said to belong to the same class or not, such as land-use type or lifestyle category. At the ordinal level the classes must have an order or ranking as in the case of road status (motorway, A class, B class, minor) or income category (under £5000 p.a; £5000 to under £10 000 p.a.; £10 000 to under £15 000 p.a. etc.). Nominal and ordinal data consist of counts by categories (categorical data) which are called discrete variables. The term 'qualitative variable' is also sometimes used.

At the continuous scale, observations may fall anywhere on a continuum. There are two levels of measurement: interval and ratio. An important difference between the interval and ratio scales is that at the interval scale it is not possible to assert whether, for example, one number is twice as big (or small) as another, because there is no natural origin. In the case of the Townsend index of material deprivation a ward with an index of 4.0 is not twice as deprived as a ward with an index of 2.0. This is because an index value of 0 does not mean an absence of deprivation.

Nominal data can be summarized and compared using modes and frequency distributions. Data at the ordinal level can be summarized and compared using medians and boxplots as well. At interval and ratio levels of measurement, means and standard deviations can also be used. Any variable measured at the ordinal level or higher can be reduced to a lower level of measurement but this results in a loss of information. However there are circumstances when this might be appropriate, as for example when the measurement process is known to contain bias. Interval- and ratio-level data might be degraded to ordinal-level data. The analyst is expressing confidence in the ordering of values, but not the actual data values themselves. This may occur in the analysis of some small-area official crime statistics, for example where the analyst is aware of undercounting but it is not such as to invalidate the real differences that exist between areas. A variable that is not of direct interest (such as a confounder variable in an epidemiological regression model) may have a non-linear effect on the predictor. One way to handle such a variable, for example an interval-valued deprivation index, is to reduce it to ordered classes so that it appears as a set of dummy variables in the model.

Even when data are retained at the ratio level, the analyst may prefer to use statistics that assume a lower level of measurement (the median to measure the centre of a distribution of values rather than the mean). This is because such statistics are robust to possible errors or extreme values that might be present in the data (Hampel et al., 1986).

Table 2.1 gives examples of spatial data classified by the type of spatial object to which a variable refers and the level of measurement of the variable attached to the spatial object. As noted above, attribute values may also be attached to the objects themselves, such as area (in the case of an area object) or length (in the case of a line object).

Variable values are associated with map objects. In order to be able to specify permissible map operations applied to the map objects it is necessary to distinguish between variables that are *spatially extensive* and those that are *spatially intensive*. Quantities such as counts by area are termed spatially extensive and when two areas are merged the corresponding counts can be summed to give the quantity for the newly created map object. Rates, densities and proportions are area dependent. The denominator refers to some attribute of the area (size, population, population at risk) and the variables are called spatially intensive (Goodchild and Lam, 1980). To arrive at the correct value of a spatially intensive variable after aggregation the numerator and denominator must be aggregated separately. This distinction has implications for areal interpolation (see section 4.2.2(b)) and for visualization and statistical analysis (see chapters 6 and 7).

2.1.4 Sample or population?

The spatial objects and the attributes that are present across geographic space may either be conceptualized as comprising the whole population or a single realization from some ‘superpopulation’. This is an important aspect of conceptualization that has implications for how spatial data are statistically analysed (the type of sampling theory that is relevant) and the nature of inference.

In some applications the reality that is observed (in terms of objects and attributes) is considered to be the only possible state. If samples are taken then

Table 2.1 *Classification of spatial data by level of measurement and type of spatial object*

Level of measurement	Spatial representation			
	Point (P)	Line (L)	Area (A)	Surface (S)
Nominal (=)	House: burgled/not	Road: under repair/not	Census tracts classified by lifestyle	Land-use type
Ordinal (≥ ; ≤)	Preference rankings of towns in a region by quality of life	Road classification (Motorway; *A*, *B*, . . . class)	Census tracts assigned to income classes	Soil texture (coarse/ medium/fine)
Interval (≤ ; ≥ ; ±)	Townsend index* for town	Length using Greenwich Meridian as reference	Townsend index* for wards	Ground temperature (°C)
Ratio (≤ ; ≥ ; ± ; × ; /)	Output from a factory p.a.	Freight tonnage p.a.	Regional per capita income	Rainfall (cm); snow depth (cm)

Notes: *Townsend index of material deprivation (see Townsend et al., 1988).
Permissible operations and relationships between numbers are given in brackets.

the inferences that are made apply to properties of that observed state. If the particular state is observed in its entirety there would be no need to undertake statistical inference except perhaps in relation to, for example, other possible spatial configurations of what has been observed. Statistical inference is used to decide whether a pattern of attribute values can be classified as random or not (the randomization hypothesis). We call this the deterministic case and either the location of the spatial objects or the attribute values or both might be conceptualized as the outcome of a deterministic process.

In other applications the reality that is observed is only one of a theoretically very large (perhaps infinite) number of possible states. In analysing the realization that has occurred (the actual counts of offences or numbers of cases of a disease) the analyst may be more interested in drawing conclusions about the underlying process responsible for it. The analyst is less interested in analysing the realized map of disease incidence or offences and more interested in specifying the generating model, estimating its parameters and testing hypotheses. These parameters, for example the relative risk of a disease or of houses being burgled, are the parameters of real interest which the actual counts of cases in any one period are used to estimate. We call this the stochastic or random case and again either the locations of the spatial objects or the attribute values or both might be conceptualized as the outcome of a stochastic process.

These contrasting conceptualizations of what is observed do not divide neatly into subject areas or types of attributes. In constructing a superpopulation view the analyst may invoke the existence of other regions from the same population, existing at the same point in time, to justify the use of statistical modelling and statistical inference (e.g. geostatistical applications in geology). In other areas the superpopulation view is underpinned by the idea of the process replicating over time in the same location (air pollution maps for different years) making the assumption that the underlying process has not changed in the intervening time period. Patterns of crop yield by small areas in any given year might be considered as one possible realization of values. The process responsible for actual yields in any particular year is complex, involving a very large number of small effects and hence viewed as stochastic in nature. Sometimes the superpopulation view is not justified in any of these ways but used as a conceptual device to avoid the criticism of placing too much emphasis on the particular data set that has been collected (such as a national population census) and on which analysis is then performed. There is often a vaguely defined superpopulation comprising, in the case of census data, the set of other national population assemblages like the one observed at the particular time when the census was taken. Sampling theory is invoked to attach confidence intervals to reported estimates.

The spatial objects and attribute values may both be stochastic or both deterministic. The spatial objects may be stochastic and the attributes deterministic. The spatial objects may be deterministic and the attributes stochastic valued. The situation is further complicated by recalling from section 2.1.3 that the term attribute may refer either to an attribute defined on the object (number or proportion of cases of a particular event attached to the object) or a spatial attribute of the object itself (such as its size or maximum spatial extent in the case of an area; length in the case of a line). This potentially gives rise to a 2 (object location) × 2 (object attribute) × 2 (attribute defined on the object) or three-way typology with eight types.

A map showing diseased trees in a region might be conceptualized as stochastically located objects (the trees) with stochastic attributes (diseased or not diseased). The location of vegetation clumps of a particular species might be conceptualized as the outcome of a stochastic process. Although the attribute referring to species type is fixed, the attribute referring to the size of each area might be conceptualized as the outcome of a stochastic process. Census areas are treated as deterministic spatial objects but the number of events occurring within the areas might be treated as the outcome of a stochastic process.

This classification is considered further in the next section.

2.2 The spatial data matrix: its form

Spatial data are classified by the type of spatial object to which variables refer and the level of measurement of these variables. Let $Z_1, Z_2, \ldots, Z_k$ refer to k variables (which may be differentiated in some applications as described in section 2.1.3(a)) and **S** to the location of the point or area. The spatial data matrix is represented generically as:

$$\begin{array}{c c}
\text{Data on the } k \text{ variables} \qquad \text{Location} & \\
\begin{bmatrix}
z_1(1) & z_2(1) & \cdots & z_k(1) & \mathbf{s}(1) \\
z_1(2) & z_2(2) & \cdots & z_k(2) & \mathbf{s}(2) \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
z_1(n) & z_2(n) & \ldots & z_k(n) & \mathbf{s}(n)
\end{bmatrix} & \begin{array}{l} \text{Case 1} \\ \text{Case 2} \\ \vdots \\ \text{Case } n \end{array}
\end{array}$$

which can be shortened to:

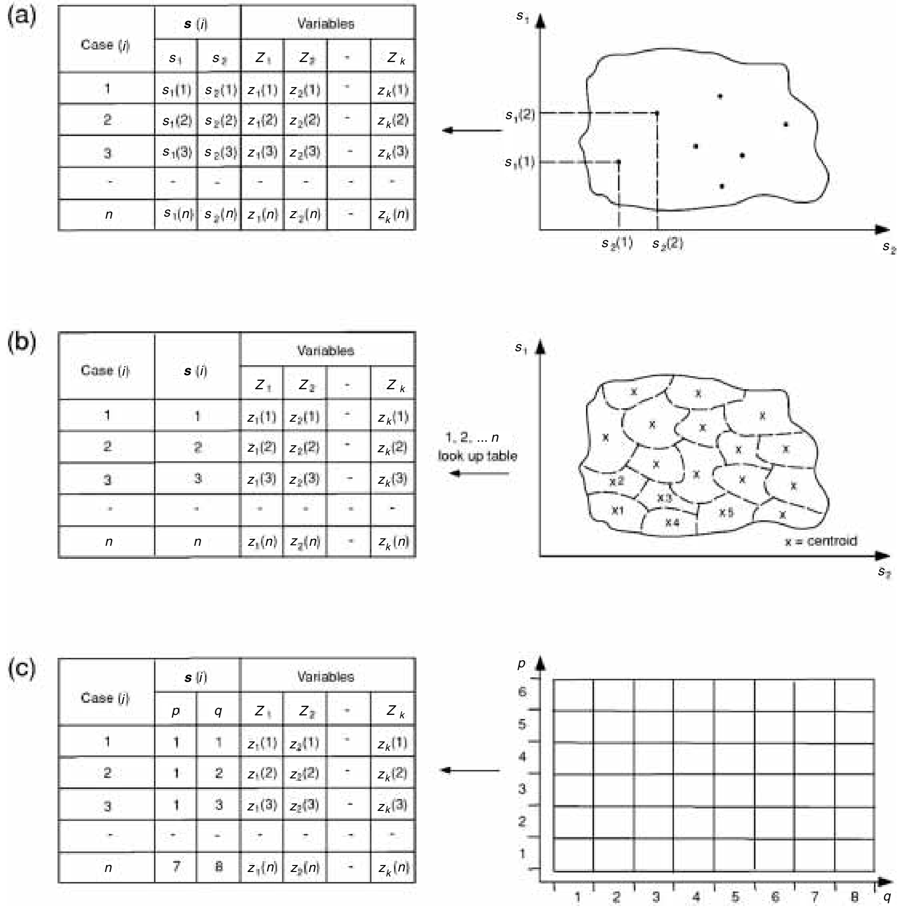
$$\{z_1(i), z_2(i), \ldots, z_k(i) \mid \mathbf{s}(i)\}_{i=1,\ldots,n} \tag{2.1}$$

The use of the lower case symbol on z and **s** denotes an actual data value whilst the symbol inside the brackets references the particular case. Attached to a case

(i) is a location $\mathbf{s}(i)$ which represents the location of the spatial object. The bold font on $\mathbf{s}(i)$ identifies this as a vector. Time is implicit in the specification because all observations should be compatible in the time period they refer to. There may be several data matrices each referring to a different time period.

In the case of data referring to point objects the location of the ith point is given by a pair of co-ordinates as illustrated in figure 2.3(a). The axes of the co-ordinate system will usually have been constructed for the particular data set but a national or global referencing system may be used. For some modelling applications the axes are scaled to the unit square. This system of referencing is appropriate whether the locations refer to the points of a discrete space or point samples on a continuous surface.

In the case of data referring to areas the 'location' of each object needs to satisfy an agreed convention. If the areas are irregular shapes then one option



(a)

Case (i)	$\mathbf{s}$ (i)		Variables			
	s_1	s_2	Z_1	Z_2	-	Z_k
1	$s_1(1)$	$s_2(1)$	$z_1(1)$	$z_2(1)$	-	$z_k(1)$
2	$s_1(2)$	$s_2(2)$	$z_1(2)$	$z_2(2)$	-	$z_k(2)$
3	$s_1(3)$	$s_2(3)$	$z_1(3)$	$z_2(3)$	-	$z_k(3)$
-	-	-	-	-	-	-
n	$s_1(n)$	$s_2(n)$	$z_1(n)$	$z_2(n)$	-	$z_k(n)$

(b)

Case (i)	$\mathbf{s}$ (i)	Variables			
		Z_1	Z_2	-	Z_k
1	1	$z_1(1)$	$z_2(1)$	-	$z_k(1)$
2	2	$z_1(2)$	$z_2(2)$	-	$z_k(2)$
3	3	$z_1(3)$	$z_2(3)$	-	$z_k(3)$
-	-	-	-	-	-
n	n	$z_1(n)$	$z_2(n)$	-	$z_k(n)$

(c)

Case (i)	$\mathbf{s}$ (i)		Variables			
	p	q	Z_1	Z_2	-	Z_k
1	1	1	$z_1(1)$	$z_2(1)$	-	$z_k(1)$
2	1	2	$z_1(2)$	$z_2(2)$	-	$z_k(2)$
3	1	3	$z_1(3)$	$z_2(3)$	-	$z_k(3)$
-	-	-	-	-	-	-
n	7	8	$z_1(n)$	$z_2(n)$	-	$z_k(n)$

Figure 2.3 Assigning locations to spatial objects

Table 2.2 *Typologies of spatial data*

Types of data+		Model or 'scheme'		Example	
GISc	Cressie (1991, p. 8)	Variable value	Spatial index	Variable	Space
Point or area object data	Lattice	Variables (discrete or continuous valued) are random variables	Point or area objects to which the variables are attached are fixed	Crime rates Land use Disease rates Prices	County Urban tracts Census tracts Retail sites
Continuous-valued field data	Geostatistical	Variable is continuous valued function of location	Variable is defined everywhere in the (two-dimensional) study region	Soil pH Surface Temp (°C)	Watershed Area of water
Randomly located point-object data	Point patterns	(i) Given attribute (ii) Variable is a random variable	Randomly located point objects in the study region	(i) Trees (ii) Trees: diseased or not (i) Hill forts (ii) Hill forts: classified by type	Forest area Forest area Archaeological research area Archaeological research area
Random area-object data	Objects	Spatial extent of each area object is a random variable	Location of area objects (e.g. their centre or origin point) in the study region is a random variable	Lichen patches Vegetation clumps	Moorland Field

Note: + Types of data as suggested by the Geographical Information Science literature (GISc) and Cressie (1991).

is to select a representative point such as the area or population-weighted centroid and then use the same procedure as for a point object to provide **s**(i). Alternatively, each area is labelled and a look-up table provided so that rows of the data matrix can be matched to areas on the map (figure 2.3(b)). If the areas are square pixels as in the case of a remotely sensed image they may be labelled as in figure 2.3(c).

It will be necessary to have a method for keeping track of spatial relationships, particularly adjacencies in the case of area data and this will be discussed in section 2.4.

The classification of spatial data by type of object and level of measurement is a *necessary* first step in specifying the appropriate statistical technique to use to answer a question. That the classification is not *sufficient* is because the same spatial object may be representing quite different geographical spaces (points are also used to represent areas for example). In addition spatial objects and the attribute values attached to the objects may be the outcome of deterministic or stochastic processes as described in section 2.1.4. Table 2.2 provides a typology of spatial data. The table uses the terminology of this chapter but also links to the terminology used by Cressie (1991) for classifying different spatial statistical models.

In describing the nature of spatial data it is important to distinguish between the discreteness or continuity of the space on which variables are measured and the discreteness or continuity of the variable values themselves. If the space is continuous (a field), variable values must be continuous valued since continuity of the field could not be preserved under discrete-valued variables. If the space is discrete (object space) or if a continuous space has been made discrete (e.g. by segmenting it, see section 2.1.1), variable values may be continuous valued or discrete valued (nominal or ordinal valued).

2.3 The spatial data matrix: its quality

The relationship between the real world and the data matrix, including the inheritance of fundamental properties such as spatial dependence, is influenced by the two phases of a mapping from reality to any specific data matrix. These are, first decisions taken on the choice of representation (in terms of both the representation of geographic space and the attributes to be included and how they are to be measured) and second by the accuracy of measurements (on both geographic co-ordinates and attribute values) *given* the chosen representation. Figure 2.4 depicts this relationship and the terms often used to characterize it (see, e.g., Longley et al., 1999). The chosen representation constitutes the *model* of the real world that is employed. Any data matrix can be assessed in terms of the quality of this model so that the first stage of assessment of a

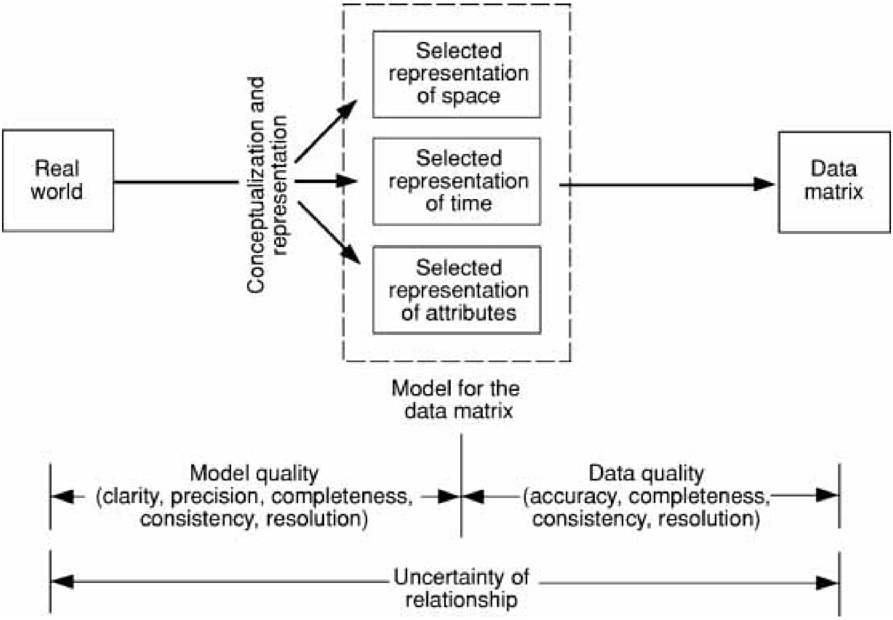

Real world
Conceptualization and representation
Selected representation of space
Selected representation of time
Selected representation of attributes
Data matrix
Model for the data matrix
Model quality (clarity, precision, completeness, consistency, resolution)
Data quality (accuracy, completeness, consistency, resolution)
Uncertainty of relationship

Figure 2.4 From geographical reality to the data matrix

data matrix can be in terms of *model quality*. Model quality may be assessed in terms of the precision (as opposed to vagueness) of a representation, its clarity (as opposed to ambiguity), its completeness (in terms of what is included) and its consistency (in terms of how objects are represented). The level of resolution or spatial aggregation is also a representational issue.

The second stage of assessment is in terms of *data quality* given the model. Importance attaches here to the accuracy of (or lack of error in) the data and completeness in the sense of coverage for example. The overall relationship between the space-(time-) attribute data matrix and the real world it is meant to capture (arguably the most important relationship) is sometimes specified in terms of the *uncertainty* of the mapping. The form of uncertainty associated with any matrix is therefore a complex combination of these two stages, model definition and data acquisition through measurement, that are associated with moving from geographic reality to the spatial data matrix. We now consider the two dimensions of model and data quality.

2.3.1 Model quality

Model quality refers to the quality of the representation by which a complex reality is captured. As described in the previous sections this involves discretizing reality in terms of a finite collection of spatial objects, spatial relationships and variables. Assessment of model quality includes whether, for example, data on the necessary spatial objects and variables, measured appropriately and sufficiently current, are available and have been correctly encoded according to the set of representation rules (Salgé, 1995). Model quality involves assessment of the appropriateness of the spatial representation of an object and the level of detail provided (including spatial detail in terms of resolution or aggregation of fields or objects). Buttenfield and Beard (1994) suggest the use of the term accuracy to reflect the correspondence between a representation or conceptualization and what the analyst wishes to measure. The term error they suggest should be used in the context of measurement processes (see below).

(a) Attribute representation

Consider first the quality of the representation of an attribute and consider a particular example. The specification for the database may include the need to have a measure of deprivation. Measurement of deprivation starts with a conceptualization of what deprivation is from which may follow a specific operationalization in the form of an index. It may be difficult to speak of error in measuring deprivation when there is no unambiguous (true) value and many ways of operationalizing the measurement of the concept both in terms of what variables to include and which types of arithmetic or logical operations to use in the construction of the index (Lee, 1999a and b; DETR, 2000). Now, suppose it is concluded that the Townsend index should be obtained for each of many spatial units. If the study area includes rural and urban areas, is an index based on the four dimensions of overcrowding, unemployment, housing tenure and car ownership as appropriate to rural as urban deprivation? In rural areas, deprived populations do not own their own house and live in overcrowded conditions but they do tend to own a car and historically at least have suffered less from unemployment.

Model quality also calls for consideration of the extent to which surrogate data represent valid proxies for what the analyst would like to include in the database. Road traffic-count data or even numbers of vehicles per unit area provide estimates of relative levels of exposure to road traffic pollution and NO_2. Wikström (1991) not only defined 'best' denominators but also what he considered the 'best practicable' for computing area offence rates in Stockholm. Swerdlow (1992) lists the minimum set of variables required to undertake an analysis of small-area cancer incidence data (see also Wakefield and Elliott, 1999). These studies help to ensure model completeness in specifying the contents of a data matrix (Brassel, 1995).

(b) Spatial representation: general considerations

Decisions on how to represent attributes in geographic space are influenced by the specific application and the spatial scale of the analysis. The choice may be made on pragmatic grounds (data are only available on a certain spatial framework or there is a need to find a common spatial framework for integrating different data sets); methodological grounds (more powerful analytical methods are available for some data representations than others); or theoretical grounds (a population density surface may better reflect the mobile nature of population in studying an infectious disease). Martin (1999, p. 75) suggests disease incidence can be represented in terms of either points, lines, areas or surfaces. Choice of representation has implications not only for the type of digital object used to capture the phenomena but also for the choice of analytical techniques and the form of visualization subsequently used. Martin (1999) notes there is no single 'right way' to represent socio-economic phenomena with the implication that 'considerable onus is placed on users to fully understand the implications of the representation strategy which they choose to adopt' (p. 78). The choice of representation should not be made uncritically for in any given circumstance the choice may raise 'technical, conceptual and ethical difficulties' (p. 79).

If analysis can disregard internal differentiation and if the geographical scale of the analysis means that detailed estimates of spatial relationships such as distance or areal configuration are not needed then the representation of an areal object such as a town by a point is appropriate. Individuals are often given fixed-point locations according to place of residence. This may raise conceptual issues. Fine grained geo-coded data provide precise information on location but it needs to be confirmed that the precision of the locational datum is appropriate to the study. In the study of non-infectious diseases, precisely locating an individual (by their address) may not be relevant in identifying exposures to certain types of environmental factors. Precisely locating individuals and then obtaining proximity or adjacency measures may not be relevant when analysing an infectious disease where social interaction rather than spatial proximity may hold the key to understanding the spread of the disease. Individual-level data (for the purpose of research or because of government or commercial interests) also raise ethical issues.

For the types of analyses described in this book, exactness of definition is required: points must have specific locations and areas must have both a precise location and spatial extent. Ambiguity or fuzziness in these definitions is not permitted except as part of a sensitivity analysis when the implications of other representations might be explored. So, the analyst must be able to justify any choice and be explicit on the criteria used to assign such a precise geometry in those cases when this is not a feature of the object. Examples of this include mapping the boundary of a forest, the use of a point to represent the 'location' of a census tract, or locating an individual by their place of residence.

(c) Spatial representation: resolution and aggregation

Pixel resolution or the level of aggregation determines the amount of spatial detail that is present in the data matrix. Decisions taken on the scale of any partition or the specific location of boundaries (particularly in the case of aggregating point objects such as households into census tracts) introduce uncertainty into the relationship between the contents of the data matrix and reality. This uncertainty arises because geographic space does not form natural units. The extent to which these discrete representations adequately reflect underlying properties of the real world depend for example on such aspects of the representation as the size of the areas or the density of the samples in relation to the spatial variability on the field. A field which varies considerably over short distances will need a denser sample of points to represent it or a more complex set of line contours than one which shows little variation. Areas used to represent a field act as filters, smoothing out variation up to the scale of the areal unit. Objects aggregated arbitrarily and modifiably into areas, fields partitioned into pixels of a given size, lose any variability that is present up to the scale of the spatial unit. Wards (five to six thousand households on average) and enumeration districts (comprising on average 150 households) conceal socioeconomic heterogeneity because real variation in such attributes is usually at a smaller scale. Aggregations that differ in terms of scale or partition (at a given scale) will affect how a fundamental property such as spatial dependence is captured in the data matrix.

The level of aggregation will also affect how 'noisy' a mapped data set is. Maps of rates constructed from aggregating small rather than large populations are more affected by counting errors so that statistics are less reliable as a basis for making area comparisons and likely to be less stable from time period to time period. Ward rates are more reliable than enumeration district rates but less reliable than for the town or city as a whole. If the underlying process is random (e.g. rates of a disease) the error variances are bigger the smaller the population.

2.3.2 Data quality

Data quality refers to the performance of the data set given the specification of the model. Any assessment of data quality from the users or producers perspective is in terms of how closely data values represent reality *given* the chosen model for representing that reality (Salgé, 1995). As far as the user is concerned both model and data quality affect the databases' fitness for purpose, but, whereas model quality assessment is specific to the application, data quality assessment involves generic criteria (Guptill and Morrison, 1995).

Assessment of data quality is of particular importance in a field where there is considerable reliance on secondary data sources. The analyst using secondary data must be satisfied they meet acceptable scientific standards but there may have to be a trade-off between data quality and the cost of acquiring better data. The data arising from some lifestyle surveys for example, may be useful in piecing together a picture of an urban environment when taken with other data but may be too noisy or contain too many errors to justify the use of rigorous analytical techniques and statistical tests.

Any assessment of spatial data quality must include both the variable (attribute) values and the spatial objects. Nor are these two dimensions of data quality independent. The right measurement assigned to the wrong location may lead to errors in counts for areas and distance measures (Griffith, 1989). Positional error in defining the extent of a vegetation region may be the result of attribute error in those cases where regional boundaries are defined in terms of attribute variation (Goodchild, 1995). Since data also have time co-ordinates, errors in recording the timing of events have implications for the quality of a spatial data set. A spatial data set involves the recording of attribute values and their co-ordinates in space and time. All three have implications for spatial data quality and errors in any one can have implications for the quality of the others.

Data quality may be spatially heterogeneous, that is the error structure may vary across the map. Location error in remotely sensed data is not uniform across a map even after geometric rectification (Borgeson, Baston and Keiffer, 1985; Ford and Zanelli, 1985; Welch et al., 1985). Heterogeneity of error can arise from the interaction between the process of measurement and the underlying geography being measured. For example, population census surveys and crime surveys usually provide more accurate counts of the number of people or number of crime events in suburban areas than they do in inner-city areas. The errors on remotely sensed images differ by sensor type and according to the nature of the topography.

Guptill and Morrison (1995) identify seven dimensions of spatial data quality: data lineage (description of the history of a data set); positional and attribute accuracy; completeness; logical consistency; temporal specification; semantic accuracy (the accuracy with which features, relationships and attributes are encoded or described given the rules for representation). Veregin and Hargitai (1995), in the context of digital cartography and geographic information systems, emphasize: data accuracy (the opposite of data error), resolution (or precision as it relates to data measurement), consistency and completeness. These four categories are discussed here.

(a) Accuracy

According to Taylor (1982), data accuracy is the inverse of data error and is defined as the difference between the value of a variable, as it appears in the database for any case, and the true value of that variable. Taylor remarks that it is 'very convenient to assume that every physical quantity does have a true value . . . We can think of the true value of a quantity as that value to which one approaches closer and closer as one makes more and more measurements, more and more carefully. As such the 'true value' is an idealization, similar to the mathematician's point with no size or line with no width' (p. 109). All measurement must entail some error because it arises from the inevitable imprecision of the process of taking a measurement together with the definitional problem that measurements in the real world are not well-defined quantities (Taylor, 1982). It follows that the type of error described by Taylor does not carry the connotation of a mistake and something that can therefore be corrected. Improved processes of experimentation and taking measurements both in terms of the quality of the instrumentation and the skills of the person taking the measurements can reduce this type of uncertainty, although can never eliminate it.

There are other practical variants to the definition of data accuracy apart from the concept of an idealized 'true value'. In some applications, error is defined as the discrepancy that exists with respect to a more accurate but expensive process of measurement, as in the case of certain types of soil measurements (Heuvelinck, 1999). The most accurate process may be impractical – for example obtaining individual-level exposures to air pollutants for large populations. In some cases reality is unobservable because it refers to historical events.

Within the definition provided by Taylor (1982) and in addition to the types of errors or uncertainties he specifies are what might be termed 'real' or 'gross' errors. By a gross error is meant an error arising for example from a failure associated with the process of measurement or, at a later stage the processes of storing, manipulating, editing or retrieving data in the database. A gross error arises from a failure to utilize the level of precision that is possible by the measurement device or the database storage device. Suppose the accuracy allowed by a measuring device in locating an object on a given map is ±1.0m when translated on to the ground. Take this to mean that skilled technicians repeatedly using the device arrive at a value within ±1.0m 95% of the time. Any user of the device who takes a measurement that is found to be, let us say, greater than 1.5m from the true value on the ground might be deemed to have generated a measurement containing a gross error. Defining an error as a 'gross' error will be more convincing if evidence, external to the measurement process, can be assembled to demonstrate that the measurement is wrong or that the experiment has been corrupted in some way – for example that a rainguage has been tampered with. Unlike the first type of error, every effort must be made to detect gross errors, eliminate or revise the corrupted data values and improve processes in the future – whilst not falling into the trap of simply discarding data values because they are out of line with expectations. We now consider some of the main types of measurement error in spatial databases.

Point location error can arise from errors in laying down or geo-coding ground markers such as environmental monitoring points. In the case of taking data from a map there can be digitizing errors linked to operator eyesight, patience and hand movement as well as the technology (Dunn et al., 1990), and there have been trials to estimate these errors (Maffini et al., 1989). There are also mapping errors associated with the construction of the source map itself (particularly in the case of older maps) and its scale (and hence the precision with which objects can be represented given map parameters such as pen size). The quality of data from ground surveys will depend for example on the precision of the theodolite. Source map errors arise from expressing a curved surface on a flat sheet of paper and shrinkage and distortion effects associated with the paper. Drummond (1995) discusses the accuracies attained by different national mapping agencies together with their method of reporting them.

The root mean square errors (RMSEs) in the two directions (north/south (s_1-direction); east/west (s_2-direction)) are used to represent positional error (Drummond, 1995). The RMSE is appropriate because measurements are at the interval or ratio scale. The RMSE for a map can be based on the discrepancy between the measured co-ordinates $(s_1(i), s_2(i))$ for n objects, and their true co-ordinates $(s_1(i;\text{ true}), s_2(i;\text{ true}))$ that have been obtained from a higher-quality measurement system. So:

$$RMSE = [(1/n)\Sigma_{i=1,\ldots,n}[(s_1(i) - s_1(i;\text{true}))^2 + (s_2(i) - s_2(i;\text{true}))^2]]^{1/2}$$

which can be decomposed into the error in different directions.

Gross errors in point data can be subtle and difficult to spot but in other cases, when they give rise to logical inconsistencies, obvious. Errors in a data set that locates only addresses within a city can be screened by superimposing city boundaries. In an analysis of road traffic accidents in the north-east of England, Raybould and Walsh (1995) found events geo-coded in the North Sea! In some systems for recording car thefts, because of inherent uncertainty in fixing the last location of the car, or because of the method of recording, thefts may get assigned to the nearest main road intersection giving a false picture of the geography of car thefts. Swerdlow (1992) cites the following sources of gross errors in health data: coding the place of cancer treatment as if it were the place of residence; using the addresses of hotels or embassies in the case of foreigners who come for treatment; using the nearest post office or cancer registry when the place of residence is unknown. 'As a result, even though such registrations are few, a small-area analysis might well show very high risk of cancer apparently relating to residence in post offices or in the registry itself!' (p. 57).

A continuous boundary is approximated in a digital database by a set of line segments. The selection of the endpoints of each line segment is the outcome of a sampling process, and the size of the errors associated with the boundary will depend on the sampling scheme and in particular on the density of sample points in relation to the complexity of the line. It has been suggested *line location error* be represented by an epsilon band on either side of the line identifying the error bounds on the line (Chrisman, 1989). 'When the width epsilon of the band is set to the deviation of the uncertainty of the line, the sausage represents some form of mean error in the area' (pp. 25–6). However, error along the line is unlikely to be independent or uniform and there appears to have been little empirical work on the shape of this band.

The errors associated with the position of the line segments of an area have implications for other forms of error. First there is likely to be error associated with derived attribute measurements taken from the measured object, such as its area and the length of its boundary. Second, spatial relationships between areas may be affected, for example whether two areas are adjacent or not. Third, there is likely to be attribute error such as errors in counts because point events are allocated to the wrong area.

Attribute error can arise as a result of the processes of collecting, storing, manipulating, editing or retrieving attribute values and as noted errors associated with the spatial objects can induce error in these measurements. Attribute error can arise from the inherent uncertainties associated with the measurement process and definitional problems, including specifying the point or period of time a measurement refers to (Taylor, 1982; Buttenfield and Beard, 1994). There are some special problems that may introduce errors into attributes associated with spatial objects. The measurement of particulate matter and other forms of atmospheric pollution at a location are subject to effects associated with monitoring sites which are positioned with respect to objects that affect the flow of air. UK census data at the enumeration district level and above are altered by the quasi-random addition of $-1, 0$ or $+1$, to counts – a process known as barnardization.

There are several ways of quantifying attribute errors. For variables recorded at the interval or ratio level the RMSE is again useful. If $z_1(i)$ is the measurement for variable Z_1 for case i, then if the true value is denoted $z_1(i;\text{true})$:

$$RMSE\,(Z_1) = [(1/n)\Sigma_{i=1,\ldots,n}\,(z_1(i) - z_1(i;\text{true}))^2]^{(1/2)}$$

In the case of nominal and ordinal data, the misclassification matrix (**M**) is used particularly for remotely sensed data, where each spatial unit is of the same size (Congalton, 1991). The columns of **M** refer to ground truth (perhaps obtained from a field survey) and the rows to the classification assigned using remotely sensed data. The diagonal elements of the matrix $\{m_{i,i}\}_i$ identify the number of correctly classified pixels so that the sum of these diagonal values divided by the total number of pixels ($m_{..}$) is the proportion of correctly classified pixels (PCC):

$$PCC = \Sigma_{i=1,\ldots,n}\,(m_{i,i}/m_{..})$$

In any application not only will misclassification rates reflect data errors but they will also be a function of how difficult it is to distinguish between classes and so error rates are affected by the degree of disaggregation into different classes. Goodchild (1995, p. 73) provides an illustration.

The Kappa (κ) coefficient (Stehman, 1996) evaluates accuracy as the discrepancy between the actual PCC and that value of PCC which would be expected if there was a random allocation of pixels to classes ($E(PCC)$):

$$\kappa = [PCC - E\,(PCC)]/[1 - E\,(PCC)]$$

where:

$$E\,(PCC) = (\Sigma_{i=1,\ldots,n} m_{i,.} m_{.,i})/m_{..}{}^2$$

where $m_{i,.}$ is the sum of values on row i of the matrix **M** and $m_{.,i}$ is the sum of values in column i of the matrix **M**. This coefficient, which formulates accuracy as a sampling problem, allows for the fact that a certain number of correct classifications will occur purely by chance. There are other measures of accuracy. The users measure is the proportion of pixels that appear to be in class i which are correctly classified ($m_{i,i}/m_{i,.}$); the producers measure is the proportion of pixels truly in class i that are correctly classified ($m_{i,i}/m_{.,i}$) (Congalton, 1991; Veregin, 1995). When areas are not of equal size other methods based on calculating proportions are used (see for example Court, 1970; Wang et al., 1997).

Where data recording methods are automated and large volumes of data are collected the risk of undetected gross error increases perhaps to the extent of casting doubt on the wisdom of analysing such data statistically at all (Hampel et al., 1986; Leonard, 1983). Careful screening of the data prior to any analysis is essential. It cannot be assumed that the sheer volume of data will overwhelm any problems of data quality that might exist. Even in carefully assembled databases such as the US Census there are recorded examples of serious, indeed bizarre, errors (Coale and Stephan, 1962; Fuller, 1975). Openshaw (1995, pp. 401–5) reviews data problems associated with the 1991 UK Census. Hampel et al. (1986) suggest that as a matter of routine between 1% and 10% of all values in a data set will contain gross errors. Rosenthal (1978) found error rates of between 0% and 4% (with an average in a very skewed distribution of 1%) in 15 data sets in psychology. Lawson (2001, p. 37) reports findings that suggest a discrepancy rate of between 12% and 18% between cause-of-death certification and necropsy. He notes that this is likely to lead to particular problems when dealing with rare diseases.

The common assumption in error analysis that attribute errors are independent (see for example Taylor, 1982) is likely to hold less often in the case of spatial data. Location error may lead to overcounts in one area and undercounts in adjacent areas because the source of the overcount is the set of nearby areas that have lost cases as a result of the location error. So, count errors in adjacent areas may be negatively correlated. Farm boundaries do not correspond to administrative boundaries so allocating farm land use to the administrative area within which the main farmhouse is situated will produce patterns of over- and undercounting which will get worse over time as farm size increases. Coppock (1955) examines the relationship between farm and parish boundaries in the UK.

In the case of remotely sensed data, the values recorded for any pixel are not in one-to-one relationship with an area of land on the ground because of the effects of light scattering. The form of this error depends on the type and age of the hardware and natural conditions such as sun angle, geographic location and season (Craig and Labovitz, 1980). The point spread function quantifies how adjacent pixel values record overlapping segments of the ground so that the errors in adjacent pixel values will be positively correlated (Forster, 1980). The form of the error is analogous to a weak spatial filter passed over the surface so that the structure of surface variation, in relation to the size of the pixel unit, will influence the spatial structure of error correlation (Haining and Arbia, 1993). Linear error structures also arise in remotely sensed data (Craig, 1979; Labovitz and Masuoka, 1984; Short, 1999).

(b) Resolution

The most important aspect of resolution as it affects data quality is in terms of the spatial dimension of the data. We consider this first before briefly discussing resolution in terms of temporal aspects of the data and attribute values.

High-resolution data, small spatial units used for small-area mapping, contain high levels of noise so it may be difficult to identify underlying structure. Kennedy (1989) and Wilkinson (1998, p. 181) discuss this in the context of disease mapping. Precision in the spatial sense does not equate with precision in the statistical sense. Small-area disease rates often display high levels of variation that is an artefact of small counts with the smallest areas often showing the most extreme rates. The addition or subtraction of only a few cases has a larger impact on rates for areas with small populations than on rates for areas with large populations.

Spatial resolution has implications for how areas can be represented for the purpose of analysis. Areas may be represented by their area or population-weighted centroids, although these terms are not always used in any exact or consistent sense. In the UK, the population centroids of EDs are determined by eye. The smaller the areal object the more it is possible to make the centroid a meaningful representation of the area. The Ordnance Survey of Great Britain's Code-Point defines the centroid of unit postcodes to a precision of 1 metre, whilst its Address-Point attains a precision of 0.1m (in most cases) in defining the centre of an individual house (Harris and Longley, 2000).

There are particular problems when data sets on different spatial frameworks and at different scales of resolution are linked to a common spatial framework. The unit postcode contains an average of about 12 households, and health data, for example, are aggregated from this level to the enumeration district level in order to attach Census information for an ecological analysis. Since the 1991 UK Census, linkage is via an ED-postcode directory in which a postcode is assigned to an ED (its so called pseudo-ED) on the basis of which ED the majority of postcode households lie within. No problem arises if the unit postcode sits entirely within an ED. However unit postcode boundaries do not nest within ED boundaries so there are many occasions when the whole of the disease count for a postcode is attributed to an ED even though part of the unit postcode lies within another ED.

Collins et al. (1998) compared the allocation of new births in 1996 to Sheffield EDs using address matching and the allocation arising from the use of the ED-postcode directory. Address matching is an expensive but more reliable method of allocation because the Ordnance Survey's Address Point coordinate falls inside the permanent building structure of the address. Of all records 16.3% (532 out of 3264) were allocated to the wrong ED by the ED-postcode directory. In terms of overall counts there is some cancelling out. Figure 2.5 shows the geography of the net under- and overcounting. The geography of this misallocation process will reflect the geography of the mismatch

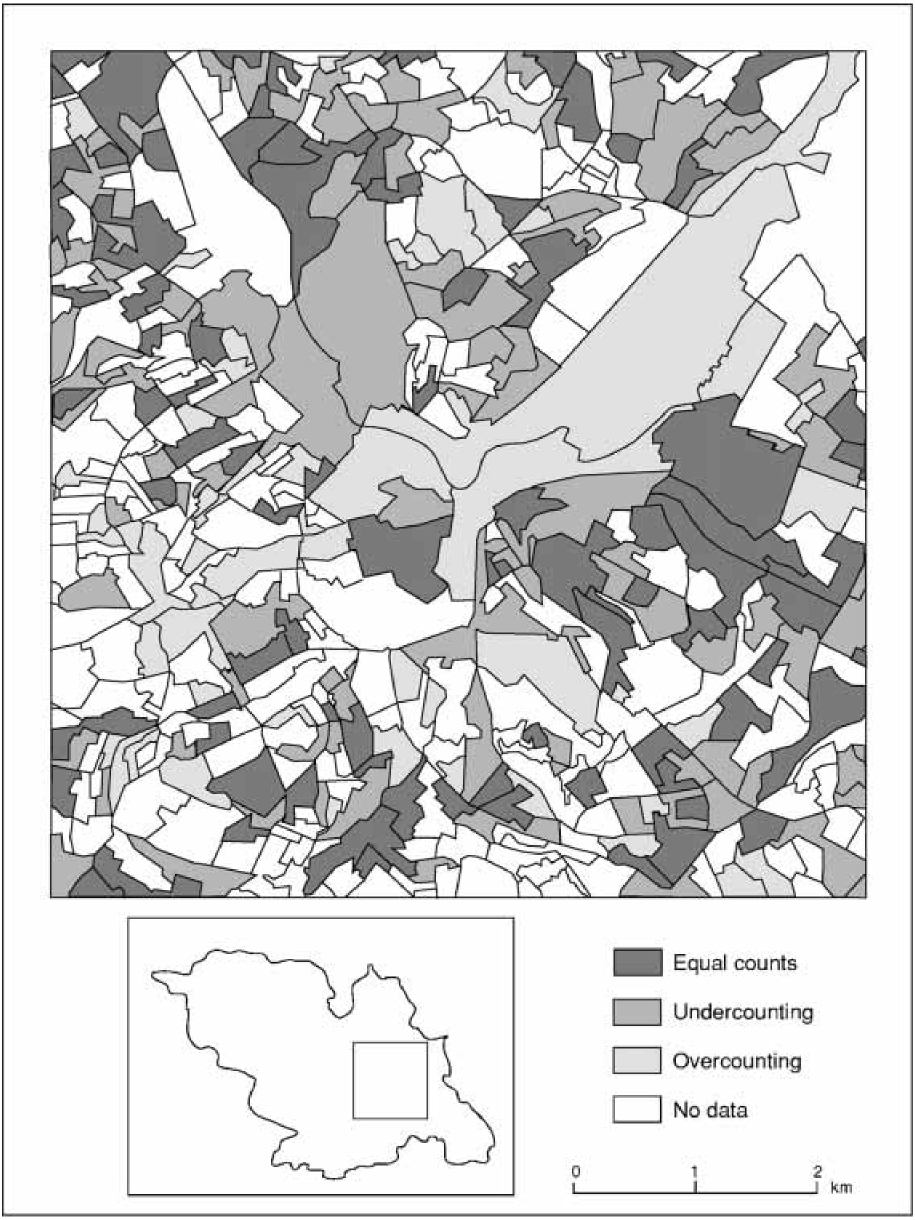


Figure 2.5 The geography of net over- and undercounting using the ED-postcode directory: new births in Sheffield 1996

between ED and unit postcode boundaries. EDs with overcounts will tend to be adjacent to EDs with undercounts since these are the EDs where the overcount is coming from. Of the 532 mismatched records the correlation between their 'correct' Townsend deprivation score (correct in the sense of the deprivation score of the ED they actually belonged to) and their allocated score was 0.495. Over a third of these records had Townsend scores greater than ±2 from the value in their correct ED (on a range from −5.4 to 13.4 across the 1057 Sheffield EDs). Such matching is used to impute deprivation scores to individual records in epidemiology and in evaluating health service targetting.

The last example illustrates the effect spatial resolution can have on the accuracy of attribute values. Counts by area contain errors when data sets that are either at different spatial resolutions or at identical but incompatible spatial resolutions have to be linked and it is not possible to return to a smaller spatial scale to ensure an accurate mapping. Craglia et al. (2001) in modelling violent crime areas in English cities in terms of socio-economic characteristics of the areas had to reconcile crime regions drawn on street plans with census areas. Even when the source data are of relatively high quality the process of data integration will result in some degrading of the quality of that data. Wherever possible ancillary data to aid the allocation should be used which is why the ED-postcode directory method (that uses household counts) was able to improve on the former OPCS Central Postcode Directory approach (that used a very crude postcode grid reference) for linking unit postcode events to enumeration districts. Theobald (1989) discusses how the spatial resolution at which elevations are measured relative to actual landform variation introduces errors into grid-based digital elevation models.

Temporal resolution refers to the period of time over which data values are aggregated and raises similar issues to those discussed in the context of spatial resolution. Aggregating counts by lengthening the time period for example may conceal small scales of temporal variation but alternatively will increase area counts making rates more robust, thereby helping to reduce the small-number problems cited previously in connection with health data and other forms of mapping.

Variable resolution refers to the precision with which attributes or other quantities are measured. In categorical data it refers to the fineness of the classification whilst in interval and ratio data it refers to how many significant numbers a value is recorded to. Neither should exceed the precision of the instrument used to collect data. Today, storage devices do not place limits on precision levels for variable values so at that stage the problem is usually to avoid spurious precision.

(c) Consistency

Consistency is defined as the absence of contradictions in a database. It refers to ‘the logical rules of structure and attribute rules for spatial data and describes the compatibility of a datum with other data in the data set' (Kainz, 1995, p. 109). Rules derive from mathematical theory and formal tests can be constructed to check that there is consistency both within and between different layers of the data set. Topological rules for spatial objects must not be violated (e.g. there is no more than one household at the same location, at the same time) and there are no contradictions in attribute values (e.g. there are no cases of a disease in a census block with no population).

Inconsistency can originate from data errors – such as geo-coding traffic accidents in the North Sea – but can be a particular problem when linking together different data sets for the same area (e.g. geological surveys and census data) or making comparisons between two time periods (Guptill and Morrison, 1995; Kainz, 1995). It is possible for each of two or more source maps to be consistent but to show inconsistencies when brought together (Kainz, 1995, p. 134). The USA's TIGER (Topologically Integrated Geographic Encoding and Referencing) system provides a structure for joining up geographic references. It describes the block structure used by the US Census and it ensures consistency in relating street patterns and other edges (rivers, roads, railways) to census and postal geographies. Such consistency is an essential underpinning to linking geographical data sets for purposes of spatial data analysis. Consistency should be retained over time to ensure that the same spatial identifiers at one census point refer to the same areas at the next and that any changes are carefully logged. In England and Wales, 70% of enumeration districts changed between 1981 and 1991. Census data in the UK is collected every ten years but health and crime data are collected on a much more frequent basis. Analysing year 2000 crime data against 1991 Census data is a form of data inconsistency that calls for further information on the geography of urban redevelopment or population migration since 1991 in order to interpret or qualify findings and to help guard against entering impossible attribute values into the database. An analyst who assembles a single national data set by linking together regional data sets needs to be sure that the regional data sets adopt the same method of classification and in the case of surveys are based on asking the same questions.

(d) Completeness

Model completeness includes whether all the variables which have been specified as necessary in order to undertake the analysis are available within the database and whether the spatial scale and geographic scope are sufficient (see section 2.3.1). The concept of model completeness is important in the context of designing a primary data collection programme and for evaluating secondary data sources. However, a data set can be 'model complete' but not 'data complete' and vice versa. Within any data set constructed to a 'model complete' specification there can be missing values, undercounts and overcounts. This is what is meant by 'data incompleteness' and there is clearly overlap in the last two types with data error. 'Spatially uniform' data incompleteness raises problems for analysis but spatial variation in the level of data incompleteness with, for example, undercounting more serious in some parts of the study area than others can seriously affect comparative work and the interpretation of spatial variation.

Swerdlow (1992) lists some of the reasons for data incompleteness in cancer data: errors in registration, including address errors, duplication and lateness in recording that can vary between the catchments of different hospitals or clinicians; errors arising from patients moving across registry boundaries; public awareness campaigns and differential access to healthcare; differences in the criteria used by pathologists in different areas when making diagnoses. In the UK, until a complete cycle of a call–recall screening programme has been undertaken there will be incompleteness because some invitations may not yet have been issued. In the case of mortality data, Lopez (1992) lists diagnostic 'fads' which lead to an overdiagnosing of certain causes of death such as cerebrovascular diseases, differences in medical training and cultural norms as causes of data incompleteness. Undercounting of certain causes of death such as sexually transmitted disease, suicide and alcoholism are due to 'diagnostic reluctance' and where there are specific sensitivities in certain localities this may lead to spatial variability in the level of undercounting. With people living longer, partly as a result of the decline of infectious diseases, the practice of only recording a single cause of death when there could be multiple pathologies will also lead to undercounting. All these factors can lead to forms of under- or overcounting and give rise to spatial variation that is an artefact of how the data were collected.

In the case of official criminal statistics, Bottoms and Wiles (1997) referring to the work of Farrington and Dowds (1984) note that drawing attention to geographical differences between large counties in England can be dangerous because of differences in police investigative and reporting practices. On the intra-urban scale, Bottoms and Wiles (1997) cite their own work and that of Mawby (1989) and others to conclude that in the case of reactive policing statistics, official crime and offender data often seem to reflect real differences between areas of a city. However they add that in any given case this should not be taken for granted and should be investigated. Separate surveys are occasionally carried out to supplement official statistics to try to estimate levels of underreporting of offences (Bottoms, Mawby and Walker, 1987). There are different levels of public reporting of offences between areas. Burglaries in suburban areas will, on the whole, be well reported for insurance purposes, but in some inner-city areas there may be underreporting either because there is no 'incentive' or because of fear of reprisals. Some crimes are more uniformly underreported: victims of sexual assault may be reluctant to report whilst there is probably substantial underreporting of domestic violence. Overreporting can be a problem with particular types of crime where there are inappropriate incentives (usually financial) to come forward with allegations.

The Census provides essential denominator data for computing small-area rates. However refusals to cooperate can lead to undercounting and the 1991 Census in the UK was thought to have undercounted the population by as much as 2% because of fears that its data would be used to enforce the new local 'poll tax'. Inner-city areas show higher levels of undercounting than suburban areas where populations are easier to track. Although there are ten-year gaps between successive censuses, population in- and out-flows in many areas may be such as to preserve the essential socio-economic and demographic characteristics of the areas. However some areas of the city may experience population mobility and redevelopment which result in marked shifts. For this reason other sources of population data have been investigated like the Family Health Service Authorities (FHSAs) patient register to track inter-census population shifts (Lovett et al., 1998).

In the USA missing data rates for Census questions ranged from 0% to 8% in 1990 but with few exceptions every housing unit reported at least one person. This is achieved by undertaking exhaustive follow-up surveys. These are expensive and were estimated to take up approximately 20% of the US Census' ten-year budget. A shift to a sampling approach to deal with non-response might result in a rate of housing unit non-response as high as 20% in addition to the anticipated 0%–8% non-response for specific questions.

Census (and other data) not only have problems of undercounting, values for areas can be missing. Missing data may be due to suppression for confidentiality reasons as in the case of those enumeration districts where there are very small numbers of households. Historical data may be missing because records have been lost or because of the stage reached in surveying an area. It can be difficult to know if some archaeological data sets are incomplete and if so the nature of the incompleteness (Hodder, 1977). In the case of remotely sensed data, some areas of the image may be obscured because of cloud cover. A distinction should be drawn between data that are 'missing at random' from data that are missing because of some reason linked to the nature of the population or the area. Weather stations may be temporarily out of action because of equipment failure, monthly unemployment records lost because of office closure or industrial dispute. These might be considered cases of data missing at random.

However mountainous areas will tend to suffer from cloud cover more than adjacent plains and there will be systematic differences in land use between such areas. This distinction has implications for how successfully missing values can be estimated and whether the results of data analysis will be biased because some component of spatial variation is unobservable.

Spatial data raise special completeness issues because if there is a geography to the incompleteness this undermines comparative work and the description and analysis of spatial variation. There are other forms of spatial data incompleteness. Map objects may refer to homogeneous vegetation types but change over time may result in boundary shifts and new vegetation regions which should be captured in the database by new spatial objects or adaptation or deletion of former map objects (Brassel et al., 1995, pp. 96–7). The geographic extent of the database may be too small to enable the analyst to detect large-scale trends or periodicities in the data (Horton et al., 1964; Burrough et al., 1985). The specification of the boundary of the study region for which data are collected proscribes the extent to which analysis can examine the role of external influences on events within the study region. This may call for the collection of data in a pre-defined zone extending beyond the strict boundary of the region of interest.

2.4 Quantifying spatial dependence

Spatial dependence is an inherent property of an attribute in geographic space because of the underlying continuity of space and the operation of various types of processes (see chapter 1). This property will be inherited by any data collected on the attribute. The form of the inheritance will depend on, for example, spatial resolution or sampling density in the case of a field space or the scale of spatial aggregation in the case of an object space (e.g. Chou, 1991).

The first step to quantifying the structure of spatial dependency in a data set is to define for any set of point or area objects the spatial relationships that exist between them. Many forms of spatial data analysis require this initial step. After this has been done there are several ways of measuring spatial dependence. These measures can be applied to the whole map to arrive at a single average measure of spatial dependence or to geographically defined subsets if heterogeneity is suspected.

(a) Fields: data from two-dimensional continuous space

Suppose point samples have been taken. Spatial relationships are typically defined on the basis of distance or distance bands. The $\mathbf{s}(i)$ in the data

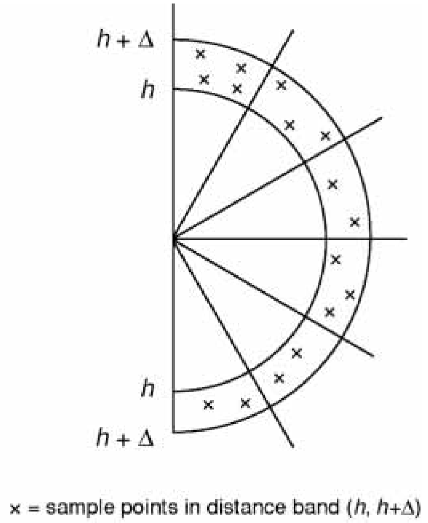


Figure 2.6 Neighbours of a single point within a distance band and by 30° segments

matrix provide sufficient information for computing inter-point distances $d(i, j)$ between any pairs of points i and j. Let the notation $[(i, j) \mid d(i, j) = h]$ denote the condition that j is selected providing it is a distance h from i. This condition may be relaxed so that j is selected providing it lies within a distance band $h \pm \Delta$ of i. Thus we write: $[(i, j) \mid d(i,j) = h \pm \Delta]$. Banding may be important to allow for uncertainty in data point locations and to ensure sufficient numbers of pairs from which to compute reliable statistics. However there may be many pairs in some bands and few in others so estimator precision will vary and ought to be allowed for in making comparisons between different bands. In the case of regularly distributed datapoints there may be no pairs in some distance bands, many in others. Where there are sufficient data, pairing can also be made to depend on direction so that j is selected providing it lies within distance band $h \pm \Delta$ and in segment k of the half circle to the east of i. Thus we write $[(i, j) \mid d_k(i, j) = h \pm \Delta]$. This is illustrated in figure 2.6 using 30° segments.

The following discussion is based on the simple case of $[(i, j) \mid d(i, j) = h]$ but can be easily generalized to the case of banding and/or segmenting. For any distance h, similarity of values can be assessed graphically using the bivariate scatterplot $\{(z(i), z(j)) \mid d(i, j) = h\}$. Similarity is indicated by a scatter which is upward sloping to the right and compact around the 45° line. If the scatter is widely spread out from the diagonal this is indicative that pairs are not similar which tends to occur as h increases. Isaaks and Srivastava (1989, p. 52) refer to plots taken in a specific direction as **h** scatterplots, where **h** is a vector denoting both distance and direction.

Numerical methods for assessing similarity in the case of variables measured at the ordinal, interval or ratio levels can be based on the squared difference $(z(i) - z(j))^2$ which will tend to be small if $z(i)$ and $z(j)$ are similar and large otherwise. A measure can also be constructed based on the cross-product $(z(i) - \bar{z})(z(j) - \bar{z})$ where $\bar{z}$ denotes the mean value of the $\{z(i)\}$. This quantity will tend to be positive if $z(i)$ and $z(j)$ are similar and either positive or negative otherwise. We now examine numerical descriptors of spatial dependency based on these two quantities.

For any given distance h, the quantity:

$$\hat{\gamma}(h) = (1/2N(h)) \underset{[(i,j)|d(i,j)=h]}{\Sigma_i \Sigma_j} (z(i) - z(j))^2 \tag{2.2}$$

where $N(h)$ denotes the number of pairs of sites separated by distance h, is the value of the semi-variogram at distance h. It will be small the more alike values separated by distance h are, and will be larger if values are dissimilar. Thus $\hat{\gamma}(h)$ tends to increase as h increases. The semi-variogram function is the plot $\{\hat{\gamma}(h), h\}$ and provides a graphical description of the dependency structure in the data for different distances. The semi-variogram computes half the average squared difference and this is also the basis of the Geary test for spatial autocorrelation described in chapter 7 (Geary, 1954). Figure 2.7(a) is an example of a typical semi-variogram for the case where spatial dependence is strong at short distances and then progressively weakens as h increases until beyond a certain distance (the range) spatial dependence levels off (the sill) close to 0.

We now turn to the second quantity. For any given distance h, define:

$$\hat{C}(h) = (1/N(h)) \underset{[(i,j)|d(i,j)=h]}{\Sigma_i \Sigma_j} (\bar{z}(i) - \bar{z}(i))(z(j) - \bar{z}(j)) \tag{2.3}$$

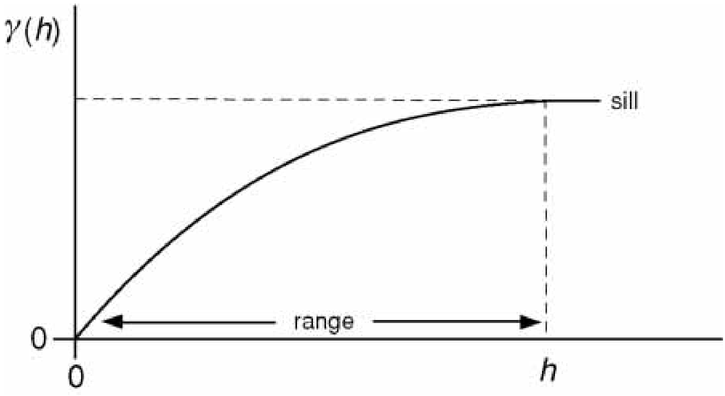


Figure 2.7(a) Model for a semi-variogram $\gamma(h)$

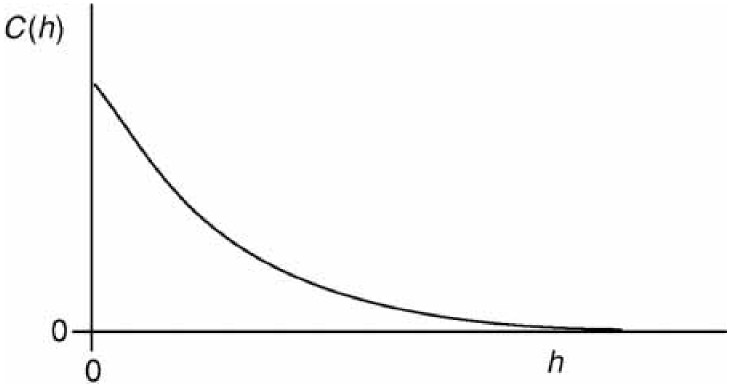


Figure 2.7(b) Model for an autocovariance function C (*h*)

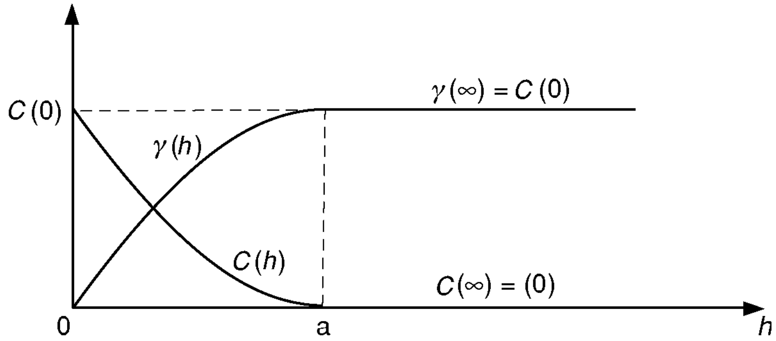


Figure 2.7(c) Relationship between $\gamma(h)$ and C(h)

The term $\bar{z}(i)$ denotes the mean of all the values that are included in the first bracket (the $\{z(i)\}$) whilst $\bar{z}(j)$ denotes the mean of all the values that are included in the second bracket (the $\{z(j)\}$). There will, of course, be many values that contribute to both means but these two means will not generally be equal. For ease of computation (2.3) can be written:

$$\hat{C}(h) = [(1/\mathrm{N}(h))\underset{[(i,j)|d(i,j)=h]}{\Sigma_i \Sigma_j}(z(i)\,z(j))] - \bar{z}(i)\bar{z}(j) \qquad (2.4)$$

$\hat{C}(h)$ is the estimate of the autocovariance (or spatial covariance) at distance h. It is the average cross-product and it is large when values are similar (both will tend to be positive or negative in the cross-product term) and close to 0 (because positive and negative values will tend to offset one another) when values are dissimilar. Computing cross-products is also the basis of the Moran test for spatial autocorrelation described in chapter 7 (Moran, 1948). Unlike $\hat{\gamma}(h)$, the plot of the corresponding autocovariance function,$\{\hat{C}(h), h\}$, tends to decrease as h

increases as spatial dependency weakens over increasing distance (see figures 2.7(b) and (c)).

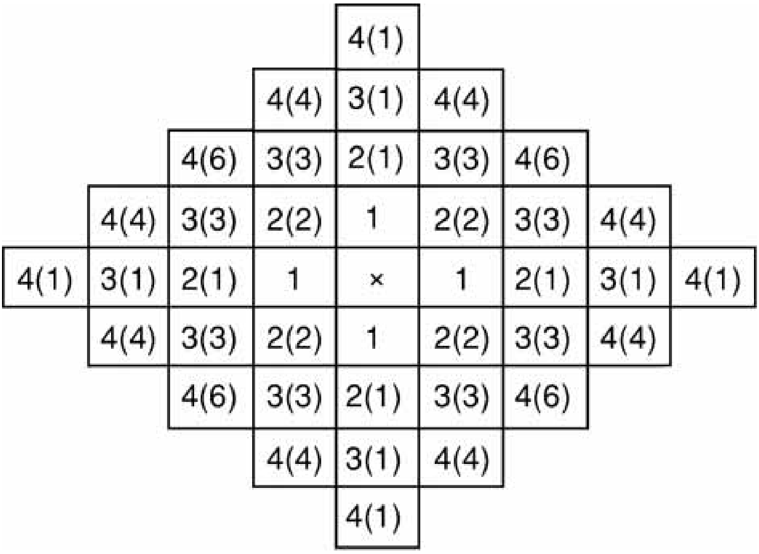
When $h = 0$, it follows from (2.3) that $\hat{C}(0)$ is the variance of the $\{z(i)\}$. If $\hat{\sigma}(i)$ and $\hat{\sigma}(j)$ are the standard deviations for the two subsets of data corresponding to $\{z(i)\}$ and $\{z(j)\}$ in (2.3) then:

$$\hat{R}(h) = \hat{C}(h)/\hat{\sigma}(i)\,\hat{\sigma}(j) \tag{2.5}$$

is the estimate of the autocorrelation (or spatial correlation) at distance h. The plot of the autocorrelation function or correlogram, $\{\hat{R}(h), h\}$, has the same behaviour as the autocovariance function but is standardized in the sense that $\hat{R}(0) = 1.0$. It can be shown that apart from boundary effects $\hat{R}(h) = \hat{R}(-h)$ and similarly $\hat{C}(h) = \hat{C}(-h)$ and $\hat{\gamma}(h) = \hat{\gamma}(-h)$ (Isaaks and Srivastava, 1989, pp. 59–60). In the case of data from a stationary process there is a close relationship between $\hat{C}(h)$, $\hat{R}(h)$ and $\hat{\gamma}(h)$:

$$\hat{\gamma}(h) = \hat{C}(0) - \hat{C}(h) = \hat{C}(0)[1.0 - \hat{R}(h)] \tag{2.6}$$

Where a representation of the field has been obtained using pixels rather than point samples, spatial relationships are defined by looking at the pixels like a set of stepping stones and defining spatial relationships by the number of *steps* required to get from any given pixel (p, q) to any other without backtracking. Define a pixel's *lag one* or *first-order neighbour* as any other pixel that can be reached by a single step that crosses their common edge ('Rook's move'). All the pixels that are *lag two* or *second-order neighbours* of any pixel (p, q) are those that can be reached by crossing two common edges without any backtracking. *Lag three* or *third-order neighbours* are those that can be reached by crossing three common edges without back tracking, and so on for fourth and higher orders of neighbours (figure 2.8). Neighbours can be differentitated by whether they can be reached by taking north/south or east/west steps and how many of each and can also be differentiated by the number of paths that can be followed. So whilst pixels $(p + 1, q + 1)$ and $(p, q + 2)$ are both two steps from pixel (p, q), pixel $(p + 1, q + 1)$ can be reached by two paths both involving one northward and one eastward step, whilst pixel $(p, q + 2)$ can only be reached by taking one path involving two eastward steps. The numbers in brackets in the cells of figure 2.8 denote the number of pathways from pixel (p, q) to the given pixel in those cases where there is more than one. These steps are not distances but because of the regular nature of the partition and since pixels are generally small they approximate distance bands and allow spatial relationships to be classified, as in the case of point samples, not only in terms of distance but direction as well. The same set of graphical and numerical methods as were described above can be adapted to this situation. However it is now necessary to distinguish between



				4(1)				
			4(4)	3(1)	4(4)			
		4(6)	3(3)	2(1)	3(3)	4(6)		
	4(4)	3(3)	2(2)	1	2(2)	3(3)	4(4)	
4(1)	3(1)	2(1)	1	×	1	2(1)	3(1)	4(1)
	4(4)	3(3)	2(2)	1	2(2)	3(3)	4(4)	
		4(6)	3(3)	2(1)	3(3)	4(6)		
			4(4)	3(1)	4(4)			
				4(1)				

Figure 2.8 Steps and numbers of paths from x to any other cell

pairs of the same lag order where there are different numbers of pathways. This means, in the case of a non-isotropic process, computing separate estimates not only for $\hat{\gamma}(0, 2)$ and $\hat{\gamma}(2, 0)$ but also for $\hat{\gamma}(1, 1)$. The full set of semi-variogram estimates are given by:

$$\{\hat{\gamma}(i, j)\}_{i=0,1,\ldots;\ j=\ldots-1,0,1,\ldots}.$$

Pixel data values may be classified into two or more nominal level categories. For example, an image may be classified into vegetation types or land-use classes. It is then possible to count the numbers of adjacent pairs whose categories are the same or the number whose categories are different. Figure 2.9 shows how these counts vary for three different types of map pattern for the case of two categories. In the case of positive dependence the number of joins of the same category will tend to be ‘large’ and there will tend to be relatively few joins where the categories are different. In the case of negative dependence the opposite will apply. The case of independence (or randomness) lies in between these two cases. This approach to quantifying spatial dependence underlies the join-count test discussed in chapter 7 (see Moran, 1948; Krishna Iyer, 1949; Cliff and Ord, 1981, pp. 11–13).

(b) Objects: data from two-dimensional discrete space

Scatterplots using measures based on average squared differences and average cross products and join-counts can again be used for quantifying spatial dependency between pairs of measurements taken on spatial objects in

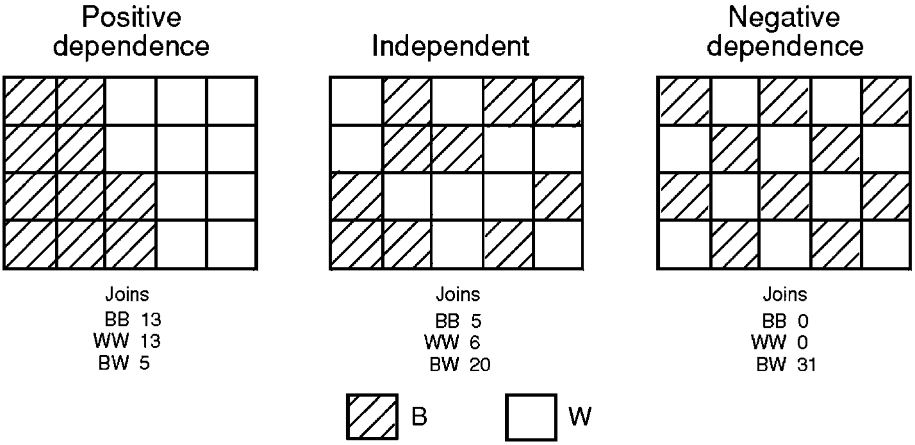


Figure 2.9 Join-counts for different map patterns

discrete space. However in this case the information provided by $\{\mathbf{s}(i)\}$ has to be supplemented with neighbourhood information. Neighbourhood information defines not only which object pairs are adjacent to each other but may also quantify the 'closeness' of that adjacency. The information provided by $\{\mathbf{s}(i)\}$ may be used in this process of defining neighbourhood information but other data may be employed and assumptions (usually untestable) made. This is needed because in discrete space and for many of the types of processes defined in discrete space there is no single or natural definition of spatial relationships (Gatrell, 1983).

The criteria used for defining neighbourhoods include:

Straight line distance: each point is linked to all other points that are within a specified distance.

Nearest neighbours: each point is linked to its k ($k = 1, 2, 3, \ldots$) nearest neighbours. (Note that if point A is one of the k nearest neighbours of B, this does not imply that B is one of the k nearest neighbours of A.)

Gabriel graphs: any two points A and B are linked if and only if all other points are outside the circle on whose circumference A and B lie at opposite points (Matula and Sokal, 1980).

Delaunay triangulation: all points with a shared edge in a Dirichlet partitioning of the area are linked. A Dirichlet partition, constructed on the points, ensures that the area surrounding any point A contains all the locations which are closer to point A than to any other point on the map (Ripley, 1981; Griffith, 1982). Figure 2.10 shows a set of points from which a Dirichlet partition has been constructed and points joined on the basis of whether they share a common border in this partition.

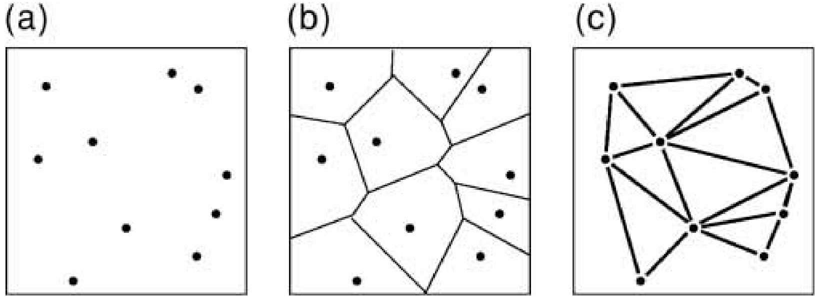


Figure 2.10 Neighbours defined using a Dirichlet partition

In the case of a set of pre-defined areas like administrative units, if they partition the study area they will share common borders so it may be appropriate to define linkages directly in terms of whether the areas share common borders or not. If the areas do not partition the study area one option is to define their location by a point (such as the area or population weighted centroid) and then apply one of the three methods for point objects described above.

Rather than defining linkages between objects in purely geometrical or spatial terms, ancillary data may be used. For example Haining (1987) modelled variation in income levels across a group of urban places in Pennsylvania whose spatial relationships were specified by drawing on central place theory. Cities were classified into orders in the urban hierarchy on the basis of population thresholds. This introduces a strong directionality into the way spatial relationhips are defined in the system of cities. Directionality was also used by Pace et al. (1998) in the analysis of house price data in Fairfax County, Virginia. 'It seems eminently reasonable to assume that the sales price of a neighbouring property will influence the subject property only if the neighbouring sale is earlier in time' (p. 17). Information on which plants are parents and which are offspring, which factories are the main sites and which are branch sites may also allow an ordering to be introduced which goes beyond purely spatial criteria.

Where the analyst wishes linkage to reflect the level of social or economic interaction between two areas then flow data on numbers of journeys or trade data may provide a sounder basis than distance or geometric properties of the set of areas. Linkage may be allowed providing interaction exceeds a certain threshold level (Holmes and Haggett, 1977). Linkages can be constructed to reflect different assumptions about the routes by which effects might be relayed across space as in the case of different pathways for the spread of an infectious disease (Cliff et al., 1985, pp. 182–5).

Geometric or spatial criteria are appropriate for defining relationships between objects if the analyst has no external criteria on which to make a judgement or where physical proximity is the main determinant of similarity. In the case of a continuous surface, distance is a natural criterion. In some circumstances there may be no strong reason to prefer any one of the geometric criteria over another but the analyst should be aware of the implications of different choices. If the purpose of analysis is spatial interpolation (see section 4.4) the analyst might want to define relationships to reduce the effects of clustering of data values on the interpolator (such as a nearest neighbour criterion by segment). However if the aim is to fit a model which is consistent with what is understood about underlying process, a pure distance-based criterion might be appropriate in the case of environmental processes; an interaction criterion in the case of social processes.

Before introducing the versions of equations (2.2) and (2.3) used to describe spatial dependency in discrete space we detour slightly to show how spatial relationships can be described using matrix methods.

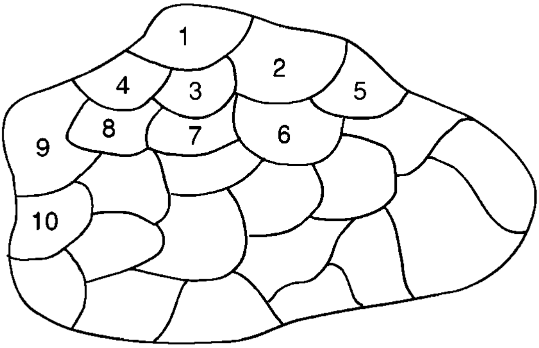
Spatial relationships can be represented in the form of a binary contiguity or *connectivity matrix* (**C**). If there are n objects (points or areas), define a matrix with as many rows and columns as there are objects ($n \times n$). Each area is assigned a unique row and column. If two objects i and j are to be defined as mutually linked then:

$$c(i, j) = c(j, i) = 1$$

where $c(i, j)$ denotes the entry on row i, column j of **C**. Otherwise any cell has the value 0. Any point or area j where $c(i, j) = 1.0$ will be called a 'neighbour' of i and be denoted $N(i)$. An object cannot be connected to itself (cannot be a neighbour of itself) so $c(i, i) = 0$ for all i. However sometimes a matrix is needed where spatial operations are performed that accumulate all values for a group of areas that include i and other areas connected to i. In this case we use $\mathbf{C}^+ = \mathbf{C} + \mathbf{I}$ (where **I** is the identity matrix with ones down the diagonal and zeros elsewhere). So it is understood that $c^+(i, i) = 1.0$.

The matrix shown in figure 2.11 is the **C** matrix corresponding to the definition of adjacency based on two areas sharing a common border. It is a matrix that is symmetric about its diagonal. In the case where object j is a neighbour of i but i is not a neighbour of j (as can arise for example with the nearest neighbour proximity criterion) then whilst $c(i, j) = 1$, $c(j, i) = 0$ and the matrix will not be symmetric about its diagonal.

If **C** is multiplied with itself, $\mathbf{C}^2 = \mathbf{C} \times \mathbf{C}$, then the non-zero cell entries identify all pairs of areas that can reach each other in two steps – second-order adjacencies. The values in the cells in the resultant matrix $\{c^2(i, j)\}_{i,j}$ identify the number of pathways. This count includes backtracking routes. If i is adjacent to four other areas then $c^2(i, i) = 4$ since there will be four ways of



$$C = \begin{array}{c|ccccccc} & 1 & 2 & 3 & 4 & 5 & \cdots & n \\ \hline 1 & 0 & 1 & 1 & 1 & 0 & \cdots & 0 \\ 2 & 1 & 0 & 1 & 0 & 1 & \cdots & 0 \\ 3 & 1 & 1 & 0 & 1 & 0 & \cdots & 0 \\ 4 & 1 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 5 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ n & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{array}$$

Figure 2.11 Binary connectivity matrix based on area adjacency

travelling from *i* back to itself in two steps. The matrix $\mathbf{C}^3$ provides the same information for third-order adjacencies and so on. These relationships, together with methods for sweeping out the paths that involve backtracking which are deemed redundant, are important for the purpose of writing software for spatial analysis of regional data where the areas are irregular (Anselin and Smirnov, 1996). In the case of analysing pixel data, this matrix representation is unnecessary and, since the matrices are sparse, wasteful of computer storage and possibly computation time. With the exception of pixels on the boundary of the study region all pixels have four neighbours that share a common edge.

There is no requirement to describe relationships between objects as simply present (1) or absent (0). Spatial relationships or those defined using interaction criteria can be defined using a more general *weights matrix*, **W**. Below are examples of different types of weights matrices (if these are applied to areas, distances may be defined by reference to area centroids):

(a) *Distance*: $w(i, j) = d_{i,j}{}^{-\delta}$ where $d_{i,j}$ denotes the distance between *i* and *j* and the parameter $\delta \geq 0$. Distance can be defined in many different metrics (see Gatrell, 1983, pp. 23–34).

(b) *Exponential function of distance*: $w(i, j) = \exp(d_{i,j}{}^{-\delta})$ where exp( ) denotes the exponential function.

(c) *Common border*: $w(i, j) = (l_{i,j}/l_i)^{\tau}$ where $l_{i,j}$ is the length of the common border between i and j, and l_i is the length of the border of i (excluding any segment which is on the boundary of the study area). The parameter $\tau \geq 0$.

(d) *Combined border and distance weighting*: $w(i, j) = (l_{i,j}/l_i)^{\tau} d_{i,j}{}^{-\delta}$.

(e) *Interaction weights* (Bavaud, 1998). Export weight: $w(i, j) = n(i, j)/n(i,.)$. Import weight: $w(j, i) = n(j, i)/n(., i)$. $n(i, j)$ is the spatial interaction from i to j; $n(i,.)$ is the total interaction leaving i; $n(., i)$ is the total interaction entering i.

In cases (a) and (b) the weighting is non-zero for all pairs of points or areas but gets smaller as distance increases. The larger the parameter δ the steeper the fall with distance. In case (c) the weighting is only non-zero in the case of areas sharing a common border and decreases as j's share of the border of i decreases, the decrease greater for larger values of τ. In case (d) the weighting is only non-zero in the case of areas sharing a common border and gets smaller as j's share of the border gets smaller and the further away it is. Case (e) is based on shares of export and import totals.

The use of the **W** notation will signify a general weights matrix so it will include the possibility of a binary connectivity matrix **C** as a special case. Note that for all areas i and j, the elements $w(i, j) \geq 0$ and usually $w(i, i) = 0.0$. If the **W** matrix has been row standardized, so that row sums equal 1 as a result of dividing each entry on a row by the sum of the row values, then this will be identified as $\mathbf{W}^*$. So:

$$w^*(i, j) = (w(i, j)/\Sigma_{j=1,\ldots,n} w(i, j))$$

Bavaud (1998, pp. 154–7) describes properties of general weights matrices and shows how they imply properties of the spatial system such as the prominence of any region or place within the total area. Several authors have examined the eigenvalues and eigenvectors of connectivity and weights matrices and identified how they characterize spatial structure (see for example Tinkler, 1972; Boots, 1982, 1984; Griffith, 1996). The principal eigenvalue of matrix **C** provides an index of the connectivity of the set of areas whilst the individual elements of the corresponding eigenvector indicates the centrality of each site within the overall configuration. This work illustrates the assumptions latent in any choice of **C** or **W**. There are further implications of the choice of neighbour that only become apparent in the context of the particular model chosen for representing spatial variation. These implications apply for example to the mean and variance properties of the model (Haining, 1990, pp. 110–13). This will be considered in section 9.1.2.

We are now in a position to define a new group of variables that will be referred to generically as spatially averaged variables and which are obtained as functions of the original set $Z_1, Z_2,\ldots, Z_k$ by performing spatial operations on the data. The general notation for these derived variables will be $\mathbf{WZ}_1,\ldots, \mathbf{WZ}_k$. The **W** prefix is simply intended to signal some spatial operation on the original variable. They are obtained as matrix products, so for example:

$$\mathbf{WZ}_1(i) = \Sigma_{j=1,\ldots,n} w(i, j) z_1(j) \quad i = 1,\ldots, n \tag{2.7}$$

If **W** is a row standardized binary connectivity matrix then (2.7) is just the mean of the values in the adjacent regions. This sort of operation is useful for representing neighbourhood conditions around an area i. If **W** is a row standardized weights matrix based on a distance function and with $w(i, i) \neq 0$, then this operation is useful for some forms of data smoothing (see chapter 7). If **W** is the unstandardized binary connectivity matrix and $w(i, i) = 1$, then (2.7) is the sum of values in region i and its neighbours. This operation is used in some cluster detection methods (see chapter 7).

In summary, the data used in the spatial analysis of discrete space comprise the original data matrix with data values and an identifier for the location of the spatial object:

$$\{z_1(i), z_2(i), \ldots, z_k(i) \mid \mathbf{s}(i)\}_{i=1,\ldots,n}$$

However, in addition at least one weights matrix (**W**) is needed to capture spatial relationships. These spatial relationships define for each $\mathbf{s}(i)$ all the neighbours, $N(\mathbf{s}(i))$. The collection of pairs $\{\mathbf{s}(i), N(\mathbf{s}(i))\}$, or $\{i, N(i)\}$ defines a graph. From this matrix new variables $\{\mathbf{WZ}_i\}$ may be constructed.

The general expressions that correspond to (2.2) and (2.3) and which can be used to quantify spatial dependence at different scales in discrete space are:

$$\hat{\gamma}(\mathbf{C}^1) = (1/2 \mid N(\mathbf{C}^1)|)\Sigma_i \Sigma_j c(i, j)\,(z(i) - z(j))^2 \tag{2.8}$$

and:

$$\hat{C}(\mathbf{C}^1) = (1/ \mid N(\mathbf{C}^1)|)\Sigma_i \Sigma_j c(i, j)\,(z(i) - \bar{z}(i))(z(j) - \bar{z}(j)) \tag{2.9}$$

where the $\mathbf{C}^1$ simply denotes the use of the connectivity matrix, **C**. $|N(\mathbf{C}^1)|$ denotes the number of pairs used in the computation. The summation terms used in (2.2) and (2.3) have been simplified by specifying the restriction on

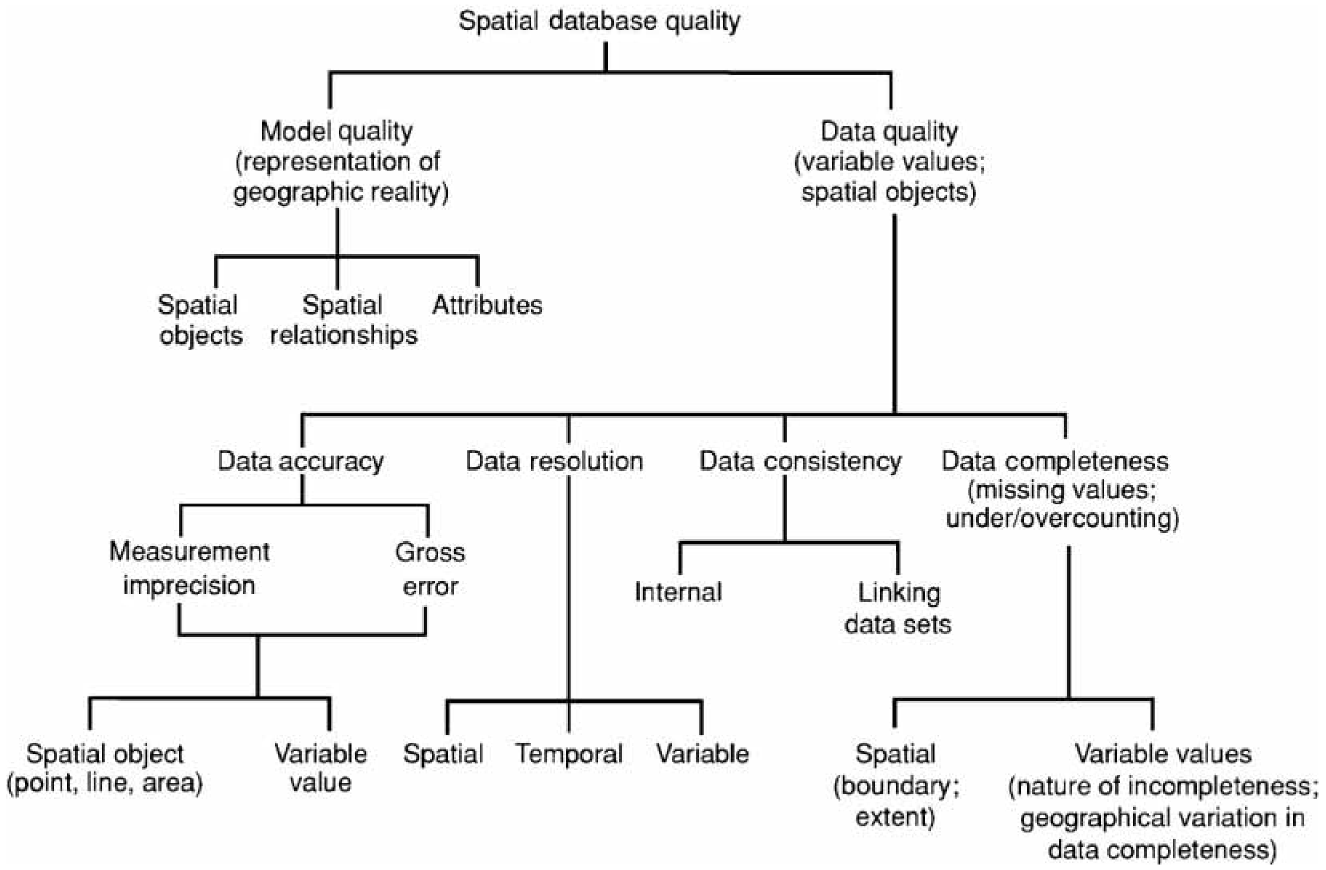

Spatial database quality
Model quality (representation of geographic reality)
Data quality (variable values; spatial objects)
Spatial objects
Spatial relationships
Attributes
Data accuracy
Data resolution
Data consistency
Data completeness (missing values; under/overcounting)
Measurement imprecision
Gross error
Internal
Linking data sets
Spatial object (point, line, area)
Variable value
Spatial
Temporal
Variable
Spatial (boundary; extent)
Variable values (nature of incompleteness; geographical variation in data completeness)

Figure 2.12 Dimensions of spatial database quality

which pairs of data values to use through the elements of the **C** matrix. Equivalently (2.8) and (2.9) can be interpreted using the pixel stepping stone analogy in section 2.4(a). Orders of neighbours or lags are specified by steps that are defined by the connectivity matrix (**C**) and powers of that matrix $\mathbf{C}^2$ ($\mathbf{C} \times \mathbf{C}$), $\mathbf{C}^3$ ($\mathbf{C} \times \mathbf{C} \times \mathbf{C}$), . . . suitably swept to remove the redundant steps. In this way semi-variogram function plots and autocovariance function plots can be obtained for irregularly distributed spatial objects. More general versions of (2.8) and (2.9) can be computed using **W** rather than **C**. These will be encountered in chapter 7 as generalized autocorrelation tests for data measured at the ordinal level and above as well as generalized join-count tests for nominal data (Cliff and Ord, 1981; Hubert et al., 1981).

The usual expectation is that values at adjacent locations tend to be similar. Spatial dependence is positive and it is common to refer to the presence of *positive* spatial dependence or autocorrelation in a data set. In the case of continuous space, in the limit (as the distance between any two points goes to zero), it is difficult to visualize any other form of spatial dependence. In the case of continuous data, where point samples are separated or pixels are of sufficient size, it is at least possible that if $z(i)$ is large (small) then $z(j)$ could be small (large). This is called *negative* spatial dependence or autocorrelation. For continuous space the presence of negative autocorrelation can only occur at a distance – between hill top and valley say. In the case of discrete space, a competition process might induce negative spatial autocorrelation between adjacent, but discrete objects. For example, adjacent plants might compete for soil nutrients which might induce negative autocorrelation in plant size.

2.5 Concluding remarks

This chapter has described the framework for spatial data analysis to be used in this book and has brought together a number of model and data quality issues that arise in working with spatial data (see figure 2.12). Any spatial data set provides an abstraction of a complex reality. This chapter has outlined the generic characteristics of data quality in terms of accuracy, precision, consistency and completeness with respect to variables, spatial objects and time. The principal message is that the analyst needs to be alert not only to the problems but how they may impact differentially across a study area or in making comparisons through time or between different study areas. It is for the user to make an assessment of the quality of the data set and to establish fitness for purpose. It may not be feasible to examine the entire data set since this will greatly increase the costs of data capture but a representative sample of the data should be examined in order to evaluate its quality. It will also be necessary to decide which errors are critical (and must be addressed) and which errors are unlikely to lead to serious consequences.

Chapter 4 will examine some of the implications of these data quality issues for spatial data analysis. Different aspects of data quality influence different stages of spatial data analysis. Some concerns, like the presence of data errors, need to be considered at the stage of collecting, preparing and finalizing the data to be analysed. Others, including the resolution or precision of the data, need to be considered in relation to the form and conduct of analysis and in relation to the interpretation of findings. However the next chapter considers sources of spatial data and the special considerations that may arise in obtaining data through spatial sampling.