Spatial Data Analysis

Theory and Practice

Robert Haining
University of Cambridge

Cambridge University Press

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Lexicon No. 2 10/14 pt and Lexicon No. 1 *System* $\LaTeX 2_\varepsilon$ [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Haining, Robert P.
Spatial data analysis: theory and practice / Robert Haining.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-77319-9 – ISBN 0-521-77437-3 (pbk.)
1. Spatial analysis (Statistics) 2. Geology – Statistical methods – Data processing. I. Title.
QA278.2 .H345 2003
001.4′22–dc 2002031242

ISBN 0 521 77437 3 paperback

Contents

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

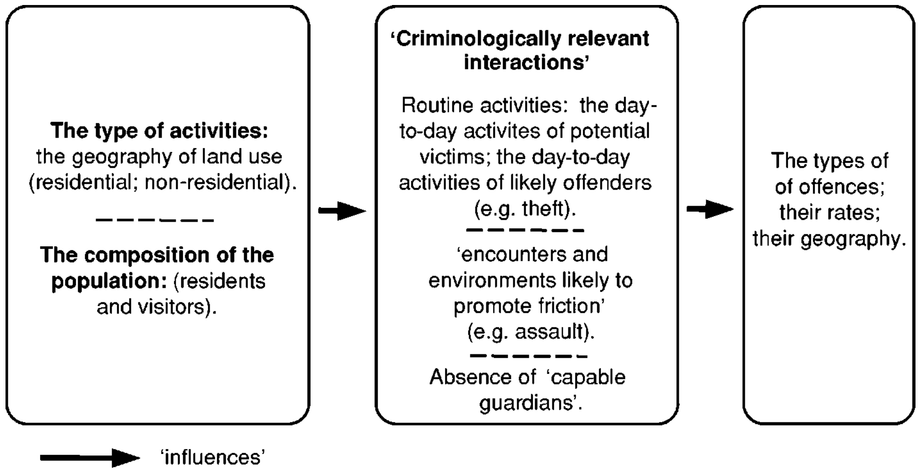

Figure 1.1 Wikström' s (1990) model for the geography of offences (after Bottoms and Wiles, 1997)

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). Wikströn 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 socio-economic 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).						