Interactive spatial data analysis

Trevor C. Bailey
Department of Mathematical Statistics
and Operational Research
University of Exeter

Anthony C. Gatrell
Department of Geography
Lancaster University

LONGMAN

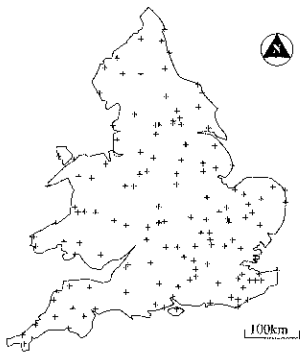

Fig. 1.3 Locations of rainfall measurement sites in England and Wales

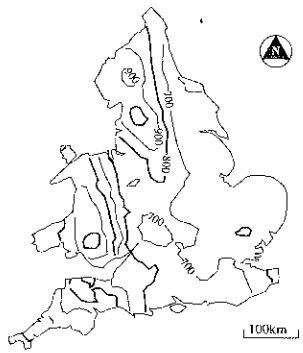

Fig. 1.4 Contoured precipitation levels (mm) in England and Wales

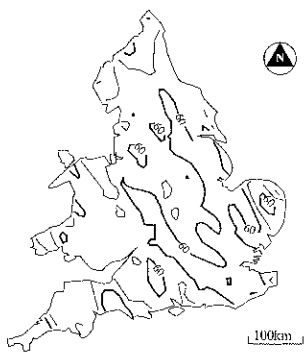

Fig. 1.5 Prediction errors (mm) of precipitation in England and Wales

have already referred to as spatial correlation; in other words, the monitored rainfall at one station is likely to be similar to that 20 kilometres away, but probably less similar to the rainfall 200 kilometres distant. This pattern of correlation can be measured and used (possibly in conjunction with other covariates) in the estimation of rainfall elsewhere on the map. The resulting estimates can be contoured (Figure 1.4), revealing that annual rainfall is generally higher in the west. However, parts of Wales and the north-west (near the Lake District) are shown as rather drier than local experience suggests. This is because the estimates there are rather more uncertain; reference back to the distribution of weather stations reveals a paucity of stations in these areas,

simply because the stations are used primarily for agricultural purposes and mountainous areas are under-sampled. The important point about the kind of analysis undertaken here is that it also generates an accompanying map of prediction errors (Figure 1.5), a map which highlights the confidence we may have in the predictions in different areas. It is these and related questions, in respect of spatially continuous data in general, rather than rainfall in particular, which form the subject of Part C of the book.

Let us turn now to a study that typifies the kind of spatial problem considered in Part D, where we deal with *area data*, spatial data that has been aggregated to areal units such as districts or census zones. We are all aware of the devastation that HIV infection and AIDS is wreaking in some parts of the world. The pool of HIV infection in parts of central Africa is huge. By the end of 1989 it was estimated that there were at least one million carriers of HIV in Uganda. There were over 12 000 reported cases, but this was thought to be a gross underestimate. Can we explore the distribution within a country such as Uganda, perhaps as a first step towards encouraging the implementation of policies to halt the spread? This is one of the tasks attempted by the geographers Andrew Cliff and Matthew Smallman-Raynor. Acknowledging the difficulties of putting together a data set to study as complex a topic as this, they mapped the incidence rates of AIDS in 34 districts of Uganda (Figure 1.6),

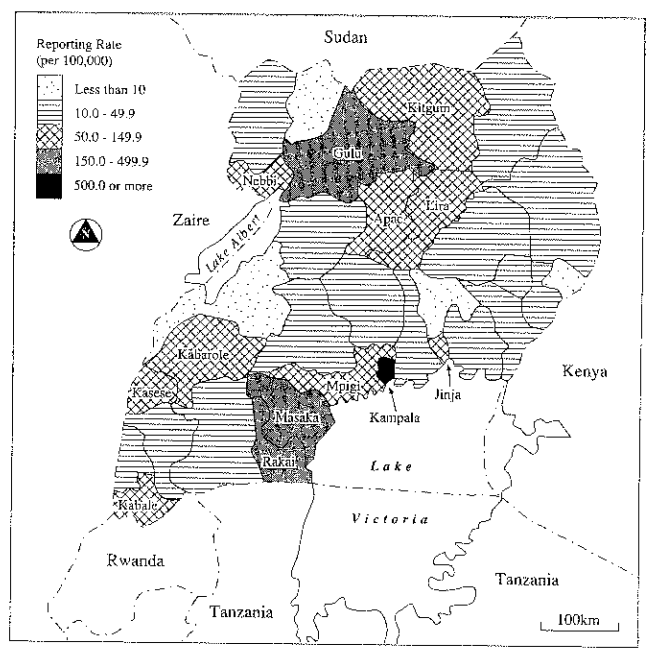

Fig. 1.6 Incidence of AIDS in Ugandan districts

a map that points to two foci of incidence. They then considered what hypotheses might explain the observed variation in this map.

One hypothesis is that the infection is acquired by those working in major urban areas, who then return home to infect partners in rural areas. Another hypothesis is that the infection is passed along major routes or 'corridors'. A third hypothesis is that involving contacts between soldiers and prostitutes; given that Uganda has been wracked by civil war during much of its recent history we surely cannot discount such factors? Cliff and Smallman-Raynor set out to test these various hypotheses, assembling an admittedly imperfect data set, but one which is probably as good as can be constructed. This involves explanatory variables such as accessibility to roads, in and out-migration rates, and army recruitment. Only the last of these variables can explain a significant proportion of variation in AIDS incidence, offering confirmation of the third hypothesis. However, the level of explanation is not high; there is considerable residual variation. This can be mapped (Figure 1.7) in order to see if there are clusters of districts with more AIDS cases than the model predicts. The residuals seem to be spatially quite random and this suggests the existence of non-spatial local factors that account for higher than expected rates in these areas. At this stage, spatial data analysis hands over to field workers, such as public health specialists, who can begin to interpret such residual variation. No-one, least of all the present authors, would be so naive as to suggest that

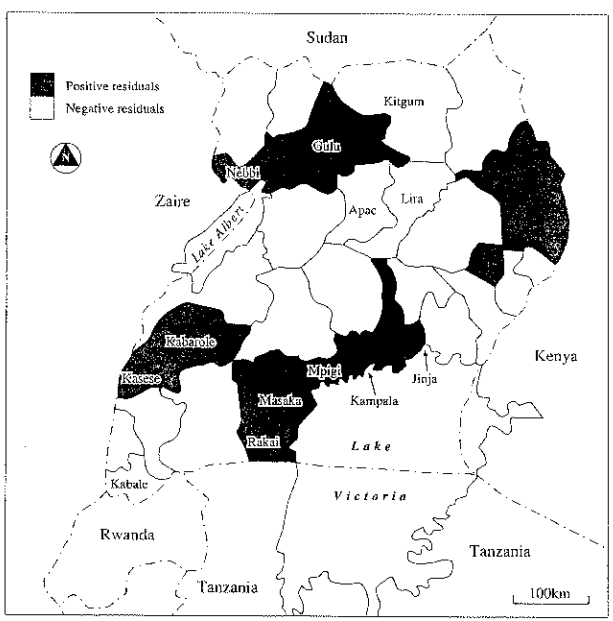

Fig. 1.7 AIDS residuals in Uganda

this kind of spatial data analysis can provide any definitive answers to the AIDS problem in Uganda; but it does offer some prospect of suggesting policies for controlling the pandemic. It is problems such as these and many others related to the analysis of area data, which form the subject of Part D.

Ex 1.7

Finally, we outline the kind of problem considered in Part E of the book. Planners, both in the private and public sectors, are often faced with problems of finding suitable locations for new investment. Whether dealing with sites for new supermarkets, or sites for such public facilities as hospitals, libraries or sports centres, planners need information about usage of existing facilities in order to make sensible investment decisions. One way, though rather crude, of seeing whether existing coverage is adequate is simply to draw circles of fixed radius around the sites and then to inspect the map for 'gaps'. But this takes no account of current human spatial behaviour or trip-making patterns. In other words, an important piece of information is the patterns of flows from residential locations to facilities. We can envisage this as a table, or matrix, of flows, from residential zones (rows of the matrix) to locations of existing facilities (columns). Can we model the observed pattern of flows?

This kind of problem goes under the general name of the modelling of *spatial interaction data* and has spawned a huge literature over the past thirty years. An interesting application is to the problem of modelling the demand for sports facilities. A specific example is work done by the geographers Goodchild and Booth on the demand for swimming pools in the city of London, Ontario. A set of 460 Census enumeration districts were used as residential zones, and a 10 per cent sample of visitors to 11 swimming pools was used to draw up a (460×11) matrix of flows. The problem is to construct a simple model of these flows; if this is possible we can then assess the effect of opening a new pool somewhere else, or the impact on existing flows if new housing is built in one or more zones. Goodchild and Booth show what the impact on flows might be if two new pools were constructed.

Such models are typically predicated on the assumption that the flow from a residential zone to a facility is directly related to the demand from that zone (such as population size) and the 'attractiveness' of the facility. In the current example, demand was weighted by the age distribution of the population, so that zones with the same total population generated different demands if their age structures differed. Spatial interaction models also assume that such interaction declines with distance, so that people are more likely to patronise a pool which is closer than one which is further away. This 'distance decay effect' can be built into the model in a variety of ways. When the model is fitted to observed data the importance of the distance decay effect can be estimated, as can the differential attractiveness of the swimming pools. Precisely how such models are constructed, and estimated statistically, we shall explore in Part E of the book.

Ex 1.8

These, then, are outlines of four case studies that deal with problems involving spatial data analysis. They each represent one of the classes of problem with which we deal in the book. To repeat, the first class of problem (Part B) deals with data for a set of point events, or a *point pattern*; sometimes