Spatial point pattern analysis and its application in geographical epidemiology

Anthony C Gatrell*, Trevor C Bailey, Peter J Diggle*** and Barry S Rowlingson†**

This paper reviews a number of methods for the exploration and modelling of spatial point patterns with particular reference to geographical epidemiology (the geographical incidence of disease). Such methods go well beyond the conventional 'nearest-neighbour' and 'quadrat' analyses which have little to offer in an epidemiological context because they fail to allow for spatial variation in population density. Correction for this is essential if the aim is to assess the evidence for 'clustering' of cases of disease. We examine methods for exploring spatial variation in disease risk, spatial and space-time clustering, and we consider methods for modelling the raised incidence of disease around suspected point sources of pollution. All methods are illustrated by reference to recent case studies including child cancer incidence, Burkitt's lymphoma, cancer of the larynx and childhood asthma. An Appendix considers a range of possible software environments within which to apply these methods. The links to modern geographical information systems are discussed.

key words spatial point patterns spatial clustering epidemiology geographical information systems

*Department of Geography, Lancaster University, Lancaster LA1 4YB. a.gatrell@lancaster.ac.uk
**Department of Mathematical Statistics and Operational Research, University of Exeter, Exeter EX4 4QE. t.c.bailey@exeter.msor.ac.uk
***Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YB. p.diggle@lancaster.ac.uk
†North West Regional Research Laboratory, Lancaster University, Lancaster LA1 4YB. b.rowlingson@lancaster.ac.uk

revised manuscript received 23 May 1995

Introduction

The analysis of spatial point patterns came to prominence in geography during the late 1950s and early 1960s, when a spatial analysis paradigm began to take firm hold within the discipline. Researchers borrowed freely from the plant ecology literature, adopting techniques that had been used there in the description of spatial patterns and applying them in other contexts: for example, in studies of settlement distributions (Dacey 1962; King 1962), the spatial arrangement of stores within urban areas (Rogers 1965) and the distribution of drumlins in glaciated areas (Trenhaile 1971). The methods that were used could be classified into two broad types (Haggett *et al.* 1977). The first were *distance-based* techniques, using information on the spacing of the points to characterize pattern (typically, mean distance to the nearest neighbouring point). Other techniques were *area-based*, relying on various characteristics of the frequency distribution of the observed numbers of points in regularly defined sub-regions of the study area ('quadrats').

For many geographers, point pattern analysis will conjure up images of 'nearest-neighbour analysis' applied inappropriately to data sets of doubtful relevance. Even contemporary textbooks

ISSN 0020-2754

in quantitative methods (for example, Griffith and Amrhein 1991; McGrew and Monroe 1993) discuss quite limited distance-based and area-based methods and do not consider the substantial and systematic advances in the statistical analysis of spatial point processes that have been made in the last twenty years.[1] Given the particular area of application we consider here, the time is ripe for an assessment of the 'state of the art' in this field, though we focus on a sub-set of methods that have particular value in geographical epidemiology. Despite this emphasis, all of the methods we outline have applications in other areas of geographical inquiry.

Apart from well-understood shifts in disciplinary emphasis away from a perspective based on spatial analysis, there are perhaps two other reasons why spatial point pattern analysis has, until recently, been neglected in geography. The first (and more significant) reason is that the null hypothesis with which most of the early methods were concerned was rarely of real practical value. Typically, methods sought to establish departures from complete spatial randomness. Whilst this might prove a sensible benchmark in some cases, in others (such as examining the distribution of disease or the locations of retail outlets in urban areas) it is unlikely to prove illuminating. Although we shall make reference to the important concept of complete spatial randomness, we stress that the methods we outline go well beyond seeking solely to establish non-randomness. A second reason is simply the lack of availability of good software. While computer programs for nearest-neighbour or quadrat analysis were published (see, for example, Baker 1974), these generated purely textual output of statistical summaries and little or nothing in the way of maps or other graphical displays.

More recently, the statistical analysis of point patterns is attracting renewed interest, notably because of developments in geographical information systems (GIS). The proliferation of geo-referenced databases, many of which generate data that may be treated as spatial point patterns, coupled with the need to infuse GIS with greater analytical functionality, have been major factors motivating the kind of work reported here and elsewhere (Gatrell and Rowlingson 1994). In particular, new tools have been developed for the analysis of point data; they are reviewed in the Appendix. These are now available as libraries that may be called from existing statistical programming environments, 'macros' that may be called from proprietary GIS packages, or functions within spatial analysis packages. They provide a variety of tools for the visualization, exploration and modelling of point data. In other words, they allow us simultaneously to view the point pattern, create new views of the pattern (for instance, showing variations in point density), explore structure in the data by estimating suitable summary functions and test hypotheses relating to the process that may have given rise to the observed event distribution.

Two final introductory remarks are in order. First, we observe that the use of spatial point pattern analysis in geographical epidemiology is hardly new, though some recent accounts in the epidemiology literature (Barreto 1993) seem to have discovered simple dot mapping as a useful technique! Many accounts draw on the classic work of John Snow in Victorian London, linking the 'clustering' of cholera deaths around a pump in Soho to the probable source of infection – an example that appears in many introductory accounts of medical geography (see, for example, Cliff and Haggett 1988; Thomas 1993). A wide range of analytical methods has been devised to handle spatial point patterns in epidemiology; we do not seek to review these comprehensively here, focusing instead on those methods we have found most useful in applied work. For example, one obvious omission in what follows is a discussion of Openshaw's 'geographical analysis machine' (Openshaw *et al.* 1987). This is now quite well-known and is finding its way into texts on medical geography (Thomas 1993).

Secondly, we do not consider in detail how to obtain disease-incidence data. Suffice it to say that many epidemiological databases, particularly in Britain but also elsewhere, now contain a post-coded address that may be converted into a grid reference (Raper *et al.* 1992). For example, in Britain, the direct link between unit postcodes and Ordnance Survey grid references with a resolution of 100 m (10 m in Scotland) means that one can readily produce mapped information on disease incidence as well as performing analyses of the point-event data, instead of aggregating these to areal units such as electoral wards. The fact that one is not required to do such aggregation renders a point pattern approach attractive, since the results from any area-based analysis are dependent on the particular zoning system one uses. A priori it seems sensible to use methods that preserve the original continuous setting of the data. On the other hand, there are a number of questionable assumptions involved in accepting a unit postcode (referring, on average, to perhaps fifteen or so other households, though with some variation about this notional mean; *ibid*.) for it to be a sensible measure of location for the disease or an adequate reflection of exposure to risk factor(s). It suggests that the individuals forming the database of disease incidence are adequately represented by their address (strictly, postcode) at the time of diagnosis. This assumes, quite naively, that people are immobile and ignores any possible exposure to environmental contamination (from whatever source) in the workplace or elsewhere. It further ignores the multitude of exposures to risk factors that may well have been picked up in earlier residential and occupational environments. However, we shall later see that in raised-incidence models we can begin to incorporate more meaningful covariates into the analysis and hence strive towards a richer interpretation and explanation of disease risk.

In the remainder of the paper, we first introduce some basic properties of spatial point processes and define some useful theoretical functions which may be used to characterize their behaviour. We indicate how one would expect such functions to behave in a 'benchmark' theoretical situation and consider how to estimate such functions from an observed point pattern and how the results may be used to explore hypotheses of interest. We then look at the issue of spatial clustering in epidemiological data, followed by the extension to a space-time context. Finally, we consider a modelling framework for assessing whether there is an elevated disease risk around a possible pollution source. In the Appendix we consider some software options for implementing the ideas developed here.

Concepts and methods

Formally, a point pattern may be thought of as consisting of a set of locations ($\boldsymbol{s}_1$, $\boldsymbol{s}_2$, etc.) in a defined 'study region', R, at which 'events' of interest have been recorded. The use of the vector, $\boldsymbol{s}_i$, referring to the location of the *ith* observed event, is simply a shorthand way of identifying the 'x' coordinate, s_{i1}, and the 'y' coordinate, s_{i2}, of an event. Use of the term 'event' has become standard in spatial point process analysis as a means of distinguishing the location of an observation from any other arbitrary location within the study region (Diggle 1983). The study region R might be a rectangular or complex polygonal region. Regardless of its shape, we must be aware of possible *edge effects* in the analysis, usually coping with these by either leaving a suitable *guard area* between the perimeter of the original study region and a sub-region within which analysis is performed, or by modifying the analytical tools to take account of boundary shape.

In the simplest case, our data set comprises solely the event locations. However, in some cases we may have additional information relating to the events which might have a bearing on the nature of analysis. For example, events may be of two different types (a *bivariate* point pattern), such as a set of individuals with a disease ('cases') and those without ('controls'). Alternatively, a continuous measure might be attached to each, an important instance being the time at which disease onset occurred among cases. This gives rise to what is known as a *marked* point pattern.

The simplest theoretical model for a spatial point pattern is that of *complete spatial randomness*, in which the events are distributed independently according to a uniform probability distribution over the region R. One important question that then arises is whether the observed events display any systematic spatial pattern or departure from randomness either in the direction of *clustering* or *regularity*. However, the role of complete spatial randomness as such a benchmark is useful only in applications where departure from it is not obvious a priori. More interesting questions, especially in the human domain, include: Is observed clustering due mainly to natural background variation in the population from which events arise? Over what spatial scale does any clustering occur? Are *clusters* merely a result of some obvious a priori heterogeneity in the region studied? Are they associated with proximity to other specific features of interest, such as transport arteries or possible point sources of pollution? Are events that aggregate in space also clustered in time? All these sorts of questions serve to take us beyond the simple detection of non-randomness and all are dealt with later.

From a statistical point of view, an observed spatial point pattern can be thought of as the outcome (a *realization*) of a spatial stochastic process. Mathematically, we may express this in various ways but one useful possibility is in terms of the number of events occurring in arbitrary sub-regions or areas, A, of the whole study region, R. Accordingly, the process is represented by a set of random variables: $Y(A)$, $A \in R$, where $Y(A)$ is the number of events occurring in the area A.We can never hope fully to characterize the process but we can investigate some properties that represent important aspects of the process; these we now consider.

First- and second-order properties

Useful aspects of the behaviour of a general spatial stochastic process may be characterized in terms of its so-called *first-order* and *second-order* properties. Very informally, the first-order properties describe the way in which the expected value (mean or average) of the process varies across space, while second-order properties describe the covariance (or correlation) between values of the process at different regions in space. In seeking to understand 'pattern' in observed spatial data, it is important to appreciate that this might arise either from region-wide 'trends' (first-order variation) or from correlation structures (second-order variation), or from a mixture of both.

More formally, first-order properties are described in terms of the *intensity*, $\lambda(\boldsymbol{s})$, of the process, which is the mean number of events per unit area at the point $\boldsymbol{s}$ (Diggle 1983). This is defined as the mathematical limit:

$$\lambda(s) = \lim_{ds \to 0} \left\{ \frac{E(Y(\boldsymbol{ds}))}{ds} \right\} \tag{1}$$

where $\boldsymbol{ds}$ is a small region around the point $\boldsymbol{s}$, $E()$ is the expectation operator and ds is the area of this region. $Y(\boldsymbol{ds})$ refers to the number of events in this small region.

The second-order properties, or spatial dependence, of a spatial point process involve the relationship between numbers of events in pairs of subregions within R. This is again formally defined in terms of a limit, the *second-order intensity* of the process:

$$\gamma(s_i, s_j) = \lim_{ds_i, ds_j \to 0} \left\{ \frac{E(Y(\boldsymbol{ds}_i)Y(\boldsymbol{ds}_j))}{ds_i ds_j} \right\} \tag{2}$$

with similar notation to that described above.

We say that a point process is *stationary* if the intensity is constant over R, so that $\lambda(\boldsymbol{s}) = \lambda$ and, in addition, $\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) = \gamma(\boldsymbol{d})$. The latter implies that the second-order intensity depends only on the vector difference, $\boldsymbol{d}$ (direction and distance), between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and not on their absolute locations. The process is further said to be *isotropic* if such dependence is a function only of the length, d, of this vector $\boldsymbol{d}$ and not its orientation. Henceforth, we use the term *stationary* without qualification to mean stationary and isotropic.

Kernel estimation

Having set out some basic ideas concerned with the properties of spatial point processes, we now consider some methods. We consider first an exploratory tool for examining the first-order properties of a point process, a tool that proves to be of potential value in an epidemiological context. Instead of superimposing a regular grid of quadrats on our event distribution, as is frequently done in geographical applications of point pattern analysis, we could form a count of events per unit area within a moving quadrat or 'window'. We define a window of fixed size and imagine centring this on a number of locations in turn, where these are arranged in a fine grid superimposed over R. We thus obtain estimates of the intensity at each grid point. This produces a more spatially 'smooth' estimate of variation in $\lambda(\boldsymbol{s})$ than we can obtain by using a fixed grid of quadrats. However, in each of the intensity estimates, no account is taken of the relative location of events within the window and the choice of a suitable window size is not clear.

Kernel estimation is a generalization of this idea, where the window is replaced with a moving three-dimensional function (the kernel) which weights events within its sphere of influence according to their distance from the point at which the intensity is being estimated. The method is commonly used in a more general statistical context to obtain smooth estimates of univariate (or multivariate) probability densities from an observed sample of observations (Silverman 1986). Estimating the intensity of a spatial point pattern is similar to estimating a bivariate probability density (Gatrell 1994). Formally, if $\boldsymbol{s}$ represents a vector location anywhere in R and $\boldsymbol{s}_1, \ldots, \quad \boldsymbol{s}_n$ are the vector locations of the n observed events, then the intensity, $\lambda(\boldsymbol{s})$, at $\boldsymbol{s}$ is estimated as

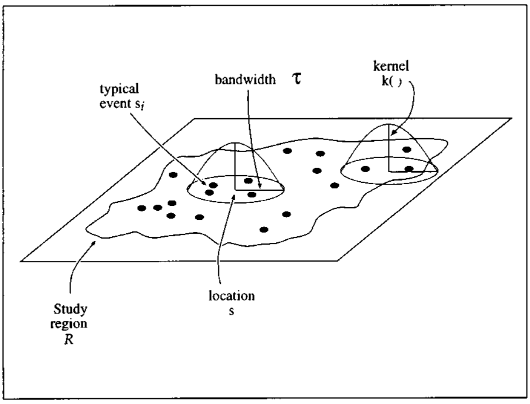

Figure 1 Kernel estimation of a point pattern

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{s-s_i}{\tau}\right) \quad (3)$$

Here, $k(\,)$ represents the kernel weighting function which, for convenience, is expressed in standardized form (that is, centred at the origin and having a total volume of 1 under the curve). This is then centred on $\boldsymbol{s}$ and 'stretched' according to the parameter $\tau > 0$, which is referred to as the *bandwidth*. The value of τ is chosen to provide the required degree of smoothing in the estimate. Graphically, we may imagine a three-dimensional function that 'visits' each point $\boldsymbol{s}$ on the fine grid (Fig. 1). Distances to each observed event $\boldsymbol{s}_i$ that lies within the region of influence (as controlled by τ), are measured and contribute to the intensity estimate at $\boldsymbol{s}$ according to how close they are to $\boldsymbol{s}$. We may then use a suitable contouring algorithm, or some form of raster display, to represent the resulting intensity estimates as a continuous surface showing how intensity varies over R.

As to the exact functional form of the kernel, $k()$ we require a decreasing radially symmetric bivariate function providing a total weight of unity over the region of influence. Different choices from amongst the range of 'reasonable' candidates have relatively little effect on the resulting intensity estimate, $\hat{\lambda}_\tau(\boldsymbol{s})$. A typical choice might be the so-called *quartic kernel*. Then, the estimate of $\hat{\lambda}_\tau(\boldsymbol{s})$ may be simply expressed as:

$$\hat{\lambda}_\tau(s) = \sum_{d_i \le \tau} \frac{3}{\pi\tau^2}\left(1-\frac{d_i^2}{\tau^2}\right)^2 \quad (4)$$

where d_i is the distance between the point $\boldsymbol{s}$ and the observed event location $\boldsymbol{s}_i$, and the summation is only over values of d_i which do not exceed τ. The region of influence within which observed events contribute to $\hat{\lambda}_\tau(\boldsymbol{s})$ is therefore a circle of radius τ centred on $\boldsymbol{s}$. At the site $\boldsymbol{s}$ (a distance of zero), the weight is simply $3/\pi\tau^2$ and drops smoothly to a value of zero at distance τ.

The kernel estimate $\hat{\lambda}_\tau(\boldsymbol{s})$ is intended to be sensitive to the choice of bandwidth, τ. As this is increased, there is more smoothing of the spatial variation in intensity; as it is reduced we obtain an increasingly 'spiky' estimate. What value, then should we choose? In practice, the value of kernel estimation is that one has the flexibility to experiment with different values of τ, exploring the surface $\hat{\lambda}_\tau(\boldsymbol{s})$ using different degrees of smoothing in order to look at the variation in $\lambda(\boldsymbol{s})$ at different scales. There are also methods which attempt automatically to choose a value of τ which optimally balances the reliability of the estimate against the degree of spatial detail that is retained, given the observed pattern of event locations (Diggle 1985). We should further note that it is possible to adjust the value of τ at different points in R in order to improve the kernel estimate. Such local adjustment of bandwidth may be achieved by *adaptive kernel estimation* (for further details and a geographical application, see Brunsdon 1991). In adaptive smoothing, sub-areas in which events are more densely packed than others (and thus where more detailed information on the variation in intensity is available) are 'visited' by a kernel whose bandwidth is smaller than elsewhere, as a means of avoiding smoothing out too much detail.

Edge effects will tend to distort the kernel estimates close to the boundary of R because an event near the boundary is denied the possibility of neighbours outside the boundary. One way to avoid the problem is to construct a guard area inside the perimeter of R, as mentioned earlier. Kernel estimates are computed only for points in R which are not in the guard area but events in the guard area are allowed to contribute to such kernel estimates. Alternatively, one can modify the kernel estimate by dividing by an explicit edge-correction term:

$$\delta_\tau(s) = \int_R \frac{1}{\tau^2} k\left(\frac{(s-\boldsymbol{u})}{\tau}\right) d\boldsymbol{u} \quad (5)$$

This is the volume under the scaled kernel centred on $\boldsymbol{s}$ which lies 'inside' R. It may result in a considerable increase in the computation required when R is an irregular polygonal region.

Extensions to kernel estimation

From an epidemiological perspective, kernel estimation is of most value in estimating the intensity of one type of event relative to another. For example, if we perform separate kernel estimates relating to cases and to controls respectively, we may then form the ratio of the two, with a view to evaluating spatial variations in disease risk. This could help identify peaks in the resulting surface corresponding to possible locations of ‘clusters’, or at least sub-regions worth further examination.

In the simplest case, the same bandwidth, τ, and kernel, $k(\,)$, are used in the estimates of both case intensity and control intensity, so allowing some cancellation in the ratio. Clearly, this is not necessary and it would be a simple modification to use different bandwidths in each. In fact, the use of different bandwidths may well be sensible since we are concerned here with the ratio of two kernel estimates of intensity. It does not necessarily follow that ‘good’ estimates of the numerator and the denominator will automatically lead to a ‘good’ estimate of their ratio. For instance, relatively small changes in the denominator (the estimate of control intensity) in regions where its value is small may produce dramatic and unacceptable variations in the ratio of the two kernel estimates. For such reasons, it may be preferable deliberately to ‘over-smooth’ the kernel estimate of control intensity when estimating the ratio by selecting a larger bandwidth than would be appropriate if we were interested only in an estimate of the population density.

These ideas were initially exploited by Bithell (1990) in a study of clustering of childhood leukaemia in Cumbria. The resulting density or probability surface was contoured, allowing peaks on the surface that correspond to an excess of case (leukaemia) intensity over that of background (child) population to be readily visualized. As a further exploratory device, this approach has much to commend itself, though, as with related methods, the controls must be selected with care and the choice of bandwidth in the kernel estimation is critical (Bithell experiments with adaptive smoothing and displays a variety of possible maps of relative risk). In a further examination of larynx and lung cancer data for Lancashire, Kelsall and Diggle (1994) have exploited ideas of kernel estimation and are currently refining the methodology for estimating two-dimensional variation in relative risk.

The K function

Having considered a method for characterizing the first-order behaviour of a point pattern, we now examine a very useful function for estimating the second-order properties of the process that gave rise to the data.

Stationarity is the minimal assumption under which inference is possible from a single observed pattern. If a point process is stationary (and isotropic), there is a close mathematical relationship between the second-order intensity and an alternative characterization of second-order properties known as the *K function* (Ripley 1981). This is defined by the relationship

$$\lambda K(d) = E(\#(\text{events} \leqslant \text{distance } d \text{ of an arbitrary event})) \qquad (6)$$

where $E(\,)$ denotes expectation, # means ‘the number of’ and λ is the intensity or mean number of events per unit area. Essentially, the K function describes the extent to which there is spatial dependence in the arrangement of events. We see shortly how this function can be estimated from an observed event distribution but, first, we establish how we would expect it to behave in a particular theoretical situation.

We have already made reference to the idea of a random arrangement of events. Formally, the point process that gives rise to such an arrangement is called a *homogeneous Poisson process*. We say that an arrangement of events shows *complete spatial randomness* (CSR) if it is a realization of such a process. As far as the K function for a CSR process is concerned, the important point is that the probability of the occurrence of an event at any point in R is independent of what other events have occurred and is equally likely over the whole of R. Thus, for a homogeneous process with no spatial dependence, the expected number of events within a distance d of a randomly chosen event is simply $\lambda\pi d^2$. In other words,

$$K(d) = \pi d^2 \qquad (7)$$

(Boots and Getis 1988; Diggle 1983; Getis 1983; Upton and Fingleton 1985). If there is clustering of point events, we would expect to see an excess of events at short distances. Thus, for small values of d, the observed value of $K(d)$ will be greater than πd^2. We say more about this later when we consider the estimation of the K function.

Extensions to the K function: bivariate and space-time patterns

The homogeneous Poisson process is a convenient benchmark against which to evaluate certain classes of phenomena. However, in many applications, especially those in the human domain, it makes little sense to compare observed spatial distributions against an homogeneous Poisson model. As we have already noted, a priori we may expect to observe a certain amount of clustering due to natural background variation in the population from which events arise. For example, cases of cancer will always cluster because of the distribution of population at risk. In such instances, we are more interested in detecting evidence of clustering over and above this underlying environmental heterogeneity; in other words, in discovering whether the distribution of one type of event clusters relative to that of another. In addition, events may have occurred at different points in time and our interest may lie in detecting space-time clustering. Accordingly, we need to consider how we might expect the kind of functions we have introduced to behave in theoretical situations where we have more than one type of event, or where ancillary information is attached to each event in the form of time of occurrence.

Consider first the detection of clustering over and above that of natural variation in background population. Given n_1 type 1 events of primary concern (cases) and n_2 type 2 events that purport to represent environmental heterogeneity (controls) then, in the absence of clustering among the cases relative to the controls, if we pool the two sets of events we would expect the n_1 case 'labels' to be attached at random to the combined set of events; this is called a *random labelling* of events. This stipulates that the type of an event is independent of its location. Under such random labelling, Diggle (1993) shows that the K functions for the cases ($K_{11}(d)$) and for the controls ($K_{22}(d)$), are identical. We shall make use of this result when looking at ways of investigating spatial clustering.

Turning now to processes in both time and space, where we have a time 'label' attached to each event, we define the 'space-time' K function by

$$\lambda_D\lambda_T K(d,t) = E(\#(\text{events} \leqslant \text{distance } d \text{ and time } t \text{ of an arbitrary event})) \tag{8}$$

where λ_D is the spatial intensity of events and λ_T their temporal intensity. If the processes operating in time and space are independent (that is, there is an absence of space-time interaction), $K(d,t)$ should be the product of separate space and time K functions. That is, we might theoretically expect

$$K(d,t) = K_D(d)K_T(t) \tag{9}$$

to apply in the case where there is no space-time interaction in the process (Diggle *et al.* 1995). Again, we can make use of this result when conducting empirical tests for space-time interaction.

Estimation of the K function

An estimate of the K function is given by

$$\hat{K}(d) = \frac{1}{\lambda^2 \mathrm{R}} \sum_{i \neq j}\sum I_d(d_{ij}) \tag{10}$$

(Boots and Getis 1988; Diggle 1983) where R is the area of region R and $I_d(d_{ij})$ is an indicator function that takes the value 1 when d_{ij} is less than d. However, this does not allow for edge effects close to the boundary of R which will distort the estimate. Consider a circle centred on event i, passing through the point j, and let w_{ij} be the proportion of the circumference of this circle which lies within R (Boots and Getis 1988). Then w_{ij} is the conditional probability that an event is observed in R, given that it is a distance d_{ij} from the *ith* event. A suitable edge-corrected estimator for $K(d)$ is then

$$\hat{K}(d) = \frac{1}{\lambda^2 \mathrm{R}} \sum_{i \neq j}\sum \frac{I_d(d_{ij})}{w_{ij}} \tag{11}$$

To complete our estimate we need to replace the unknown intensity λ with an estimate, say $\hat{\lambda} = n/\mathrm{R}$, where n is the observed number of events. The final estimate of $K(d)$ is therefore

$$\hat{K}(d) = \frac{\mathrm{R}}{n^2} \sum_{i \neq j}\sum \frac{I_d(d_{ij})}{w_{ij}} \tag{12}$$

Ignoring the edge correction, we can visualize the estimation of a K function as shown in Figure 2. We

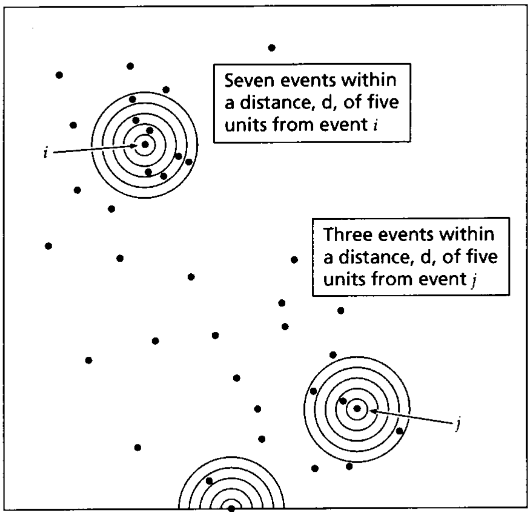
Seven events within a distance, d, of five units from event *i*
i
Three events within a distance, d, of five units from event *j*
j

Figure 2 Estimation of a *K* function

may imagine that an event is 'visited' and that around this event a set of concentric circles at a fine spacing is constructed. The cumulative number of events within each of these distance 'bands' is counted. Every other event is similarly 'visited' and the cumulative number of events within distance bands up to a radius d around all events becomes the estimate of $K(d)$ when scaled by R/n^2.

In practice, the calculation of $\hat{K}(d)$ is not easy since, for arbitrary-shaped regions, the weights w_{ij} are hard to derive. Explicit formulae for w_{ij} can be written down for simple shapes such as rectangular or circular R. In other cases the derivation of w_{ij} will require more intensive computation, although computer solutions are available (see Appendix).

Once calculated, $\hat{K}(d)$ can be compared with its expected form according to particular theoretical situations. For example, as we noted, we expect $K(d)=\pi d^2$ for a homogeneous process with no spatial dependence. Under regularity, $K(d)$ would be less than πd^2, whereas, under clustering, $K(d)$ would be greater than πd^2. So we can compare $\hat{K}(d)$, estimated from the observed data, with πd^2. This may be done through a plot of $\hat{K}(d)-\pi d^2$ against d. Peaks in positive values tend to indicate spatial clustering and troughs of negative values indicate regularity, at corresponding scales of distance, d. How do we assess whether the observed peaks or troughs in this plot are significant?

This may be done using simulation techniques. Under the assumption of CSR, we may perform m independent simulations of n events in the study region (where m might be, say, 99). For each simulated point pattern, we can estimate $K(d)$ and use the maximum and minimum of these functions for the simulated patterns to define an upper and lower simulation envelope. If the estimated $K(d)$ lies above the upper envelope, we can speak of aggregation. If it lies below the lower envelope, this is evidence of spatial 'inhibition' or regularity in the arrangement of events.

As noted earlier, estimation of a K function for a single set of events will not usually be informative in geographical epidemiology. More interesting is a test of the hypothesis of 'random labelling', which suggests that

$$K_{11}(d)=K_{22}(d)=K_{12}(d) \tag{13}$$

We might, therefore, use a plot of the difference

$$\hat{D}(d)=\hat{K}_{11}(d)-\hat{K}_{22}(d) \tag{14}$$

against d to explore such a hypothesis. When $\hat{D}(d)$ is plotted against d, peaks in this plot will show clustering over and above that of environmental heterogeneity.

A more formal assessment of the significance of peaks in this plot again employs the idea of a simulation test. Suppose, as before, that there are n_1 'cases' and n_2 'controls', then upper and lower simulation envelopes for assessing the peaks in the $\hat{D}(d)$ plot may be developed by repeated simulations in which case labels are randomly assigned to n_1 of the events (Bailey and Gatrell 1995; Diggle 1993; Diggle and Chetwynd 1991). Such a method provides a useful complement to alternative approaches (see, for example, Cuzick and Edwards 1990) which are essentially based on extensions to the distribution function of inter-event distances.

Turning to the space-time context, where we have a time 'label' attached to each event, we may develop an estimate of the 'space-time' K function defined earlier. As noted before, this definition involved the expected number of events within a distance d and time interval t of an arbitrary event, scaled by the expected number of events per unit area and per unit time. By analogy with our previous K function estimates, an appropriate edge-corrected estimate of $K(d,t)$, from an observed space-time event distribution is therefore

$$\hat{K}(d,t)=\frac{RT}{n^2}\sum\sum i\neq j\,\frac{I_d(d_{ij})I_t(t_{ij})}{w_{ij}v_{ij}} \tag{15}$$

where R, d_{ij}, $I_d(d_{ij})$ and w_{ij} are as used previously, T is the overall timespan observed, t_{ij} is the time interval between the *ith* and *jth* observed events, $I_t(t_{ij})$ is an indicator function which is 1 if $t_{ij} \leq t$ and 0 otherwise, and v_{ij} is the temporal equivalent of the spatial-edge correction, based upon whether a time interval centred on i of length t_{ij}, lies wholly within the (0, T) timespan observed (Diggle 1993; Diggle *et al.* 1995).

As mentioned earlier, if the processes operating in time and space are independent (that is, there is an absence of space-time interaction), $K(d,t)$ should be the product of separate space and time functions, $K_D(d)$ and $K_T(t)$ respectively. Thus, one possible exploratory tool for space-time interaction is the function

$$\hat{D}(d,t) = \hat{K}(d,t) - \hat{K}_D(d)\hat{K}_T(t) \qquad (16)$$

Evidence of space-time interaction will be observed as peaks on the surface of $\hat{D}(d,t)$ plotted against space and time. We see later how this may be more formally used to detect space-time interaction.

Before considering some applications, we need to establish a final point that should be appreciated throughout the rest of our discussion, namely that first- and second-order effects may be inherently confounded in many real data sets. Although part of the art of spatial point process analysis is trying to disentangle these two effects in order better to understand the process generating the observed events, inferring process from pattern is ultimately a question of judgement and may not always be clear cut. In certain cases it may not be possible to distinguish on a purely statistical basis between competing explanations (Bartlett 1964; Cliff and Ord 1981). For example, if we detect clustering in the distribution of events, is this clustering the outcome of a process of environmental heterogeneity, in which the location of an event is functionally independent from that of another, or is it more a reflection of ‘true contagion’, where the environment is homogeneous but there is direct spatial dependence between events, the location of an event being ‘influenced’ in some way by that of others? For example, we introduced the K function as characterizing the second-order properties of a stationary point process. However, if we estimate a K function in a situation where there are large-scale first-order effects – in other words, where intensity varies greatly across the study region – then any spatial dependence indicated by the estimated K function could be due more to these first-order effects rather than to interaction between the events themselves. We may have to adopt an explanation which acknowledges some overall first-order heterogeneity and then proceed to examine smaller sub-regions of R for possible additional second-order effects. In some cases, the nature of the data and the strength of trends in the observed pattern may make such judgements relatively straightforward. In other cases, this may be difficult and open to debate and interpretation. Existence of environmental heterogeneity does not necessarily invalidate the assumption of stationarity, since we may be prepared to assume that a non-constant spatial intensity may itself be a realization of an underlying stationary process. This leads to a class of models known as *stationary Cox processes* (Diggle 1983).

Applications in environmental epidemiology

Having set out some fundamental concepts and methods, we now illustrate the usefulness of these ideas. We examine first the important issue of spatial clustering, then consider the space-time extension before examining a spatial point process model of considerable value in geographical epidemiology.

Spatial clustering

The question of whether the geographical incidence of disease shows any tendency towards clustering in geographical space has a long and rich history. Do cases of disease tend to occur in proximity to other cases? The problem has become more urgent in recent years in the light of concerns raised about possible links between disease incidence and potential sources of environmental contamination, such as nuclear installations. Evidence of clustering might also lend support to other theories of disease incidence, such as a viral aetiology. For example, exposure to a common, persistent viral infection, either during gestation or as a young child with an immune system that had been protected at a very early age, might provide clues to explaining possible leukaemia clustering (see Alexander 1993 for a clear and up-to-date overview). We do not review here all the various theories or methodological developments

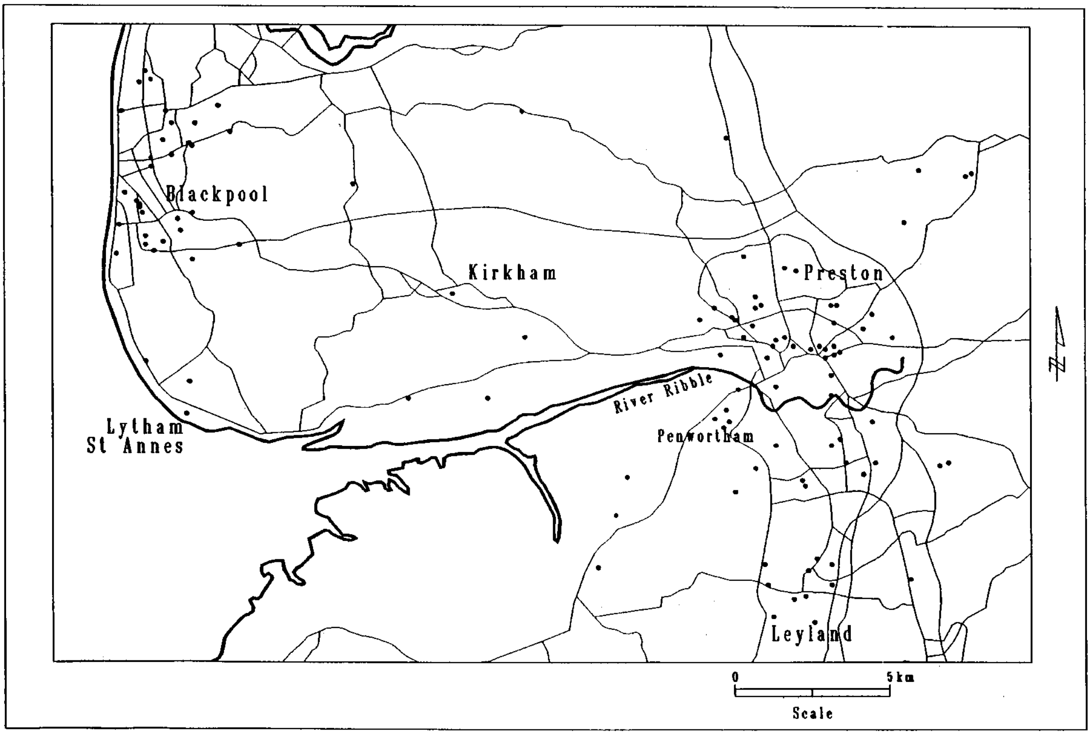

Figure 3 Locations of cases of childhood leukaemia in west-central Lancashire, 1954–92

concerned with disease clustering. Good overviews are provided in Draper (1991) and Beral *et al.* (1993) We focus instead on appropriate use of the *K* function to evaluate evidence of overall spatial clustering in an observed event distribution. In a later section we discuss the issue of whether there are specific clusters around point or linear sources of environmental contamination.

As noted above, we may estimate *K* functions both for disease cases and for controls. The difference between the two functions, along with an appropriate simulation envelope, is used to assess evidence for or against clustering. The particular application we discuss here relates to a study of the incidence of childhood cancer in the Penwortham area of central Lancashire, England (Gatrell and Whitelegg 1993). We wished to examine whether the distribution of child cancer mirrored that of the child population as a whole or whether there was evidence, as implied by concerned local residents, of clustering. Data were provided for the study by the Manchester Children's Tumour Registry on a variety of cancers (leukaemias, lymphomas, central nervous system, renal, hepatic, bone and soft-tissue sarcoma), covering the period 1954–92. All 325 cases were postcoded, the link to Ordnance Survey grid references providing the means for subsequent spatial analysis. Figure 3 shows the incidence of all leukaemias in the study region. We defined a wide study area in order to embrace not merely the area of immediate concern which covers part of the Fylde region – as far north as Blackpool – and stretches south of Preston. Defining a suitable set of controls is clearly an important and non-trivial research problem. With cases taken from as far back as 1954, it is out of the question to use school records. Instead, we extracted all unit postcodes in the study region, together with all centroids of 1981 Census enumeration districts (EDs). The latter were used to construct Thiessen polygons, a set of pseudo-ED boundaries. Next, the childhood population (ages 0–4, 5–9, 10–14) in each ED were obtained and multiplied by the fraction of all cancers in each group (as reported by Draper 1991). This defined a 'weighted' population for each ED, from which control cases were then sampled. First, an ED was selected in proportion to its weighted population (so that one with a

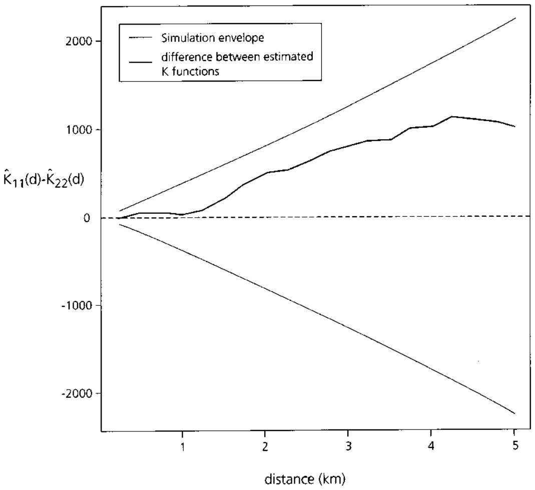

Figure 4 Difference between *K* functions (bold line) and simulation envelope (lighter lines) for childhood leukaemia and 'population at risk'

weighted population of 40, for example, had twice the chance of being drawn as one with a value of 20). A unit postcode within that ED was then selected at random, its grid reference becoming the location of a 'control'. A total of 975 controls were selected in this way, corresponding to three times the number of cases of childhood cancer. While this is far from an ideal way of establishing a set of controls and is, of course, a simulated distribution, it was felt reasonably to reflect the distribution of the population at risk.

K functions were then estimated separately for cases and controls, as described above, and subtracted to obtain the difference function, $\hat{D}(d)$. Under a random labelling of the combined set of cases and controls, the expected value of this difference is zero for all distances, *d*. Upper and lower simulation envelopes were developed by performing 99 random labellings of cases and controls. If $\hat{D}(d)$ lies above the upper simulation envelope we can speak of significant spatial clustering.

We took each of the diagnostic categories in turn and compared their spatial distribution with the control population. Results for the leukaemias are shown in Figure 4. While there is some weak tendency to cluster, there is no statistically significant evidence for clustering. The same basic pattern is repeated for other childhood cancers. We have not analysed sub-regions of the study area, though clearly this would be straightforward using the interactive software environments outlined in the Appendix. Analysis of a temporal sub-set of the data – cases diagnosed since 1974 and since 1980 – also failed to find any evidence of clustering. Results from analyses conducted in Draper (1991) for Britain as a whole, did suggest some evidence for the aggregation of childhood leukaemia between 1966 and 1983.

Space-time interaction

We now consider the problem of testing for space-time clustering. Classical tests for space-time interaction, such as those of Knox (1964) (see Thomas 1993 for a discussion), require that the user specify, in advance, distance and time thresholds within which events are considered 'close' in a spatial and temporal setting respectively. A count is then made of the number of pairs of events that are close both in time *and* in space. But, because of the simple discretization of space and time, most empirical applications of Knox's method usually perform multiple tests using different time and space intervals. Extensions are available where continuous measures of spatial and temporal separation are used (Mantel 1967) but these are also sensitive to somewhat arbitrary judgements concerned with choice of appropriate parameter values used in the transformations involved.

An alternative approach is to use the space-time *K* function described earlier (Diggle *et al.* 1995). Recall that we suggested an exploratory tool for space-time interaction based upon

$$\hat{D}(d,t)=\hat{K}(d,t)-\hat{K}_D(d)\hat{K}_T(t)$$

Evidence of space-time interaction will be observed as peaks on the surface of $\hat{D}(d,t)$ plotted against space and time. One way to devise a more formal assessment of the significance of the observed values of $\hat{D}(d,t)$ is to perform *m* simulations, in each of which the *n* events are randomly labelled with the observed *n* time 'markers'. We can thus obtain *m* estimates $\hat{D}_i(d,t)$, $i=1,..,\quad m$. The observed standard error $\hat{\varsigma}^2(d,t)$ of these *m* estimates in turn gives us an estimate of the standard error of $\hat{D}(d,t)$. We may then define a set of standardized residuals as

$$\hat{R}(d,t)=\frac{\hat{D}(d,t)}{\hat{\sigma}^2(d,t)} \qquad (17)$$

These residuals have the property that, in the absence of space-time interaction, they have expectation zero and variance of one. A more precise interpretation is difficult because residuals at different values of d and/or t are not independent.

An overall test of space-time clustering may be obtained by comparing the actual observed sum of $\hat{D}(d,t)$ over all d and t, with the empirical frequency distribution of m such sums, each of which is obtained from one of the corresponding $\hat{D}_i(d,t)$. When compared with this distribution, an 'extreme' value of the observed sum would indicate evidence of overall space-time interaction. For example, if the observed sum exceeds, say, 95 per cent of the simulated values, we can infer that the observed space-time interaction occurs by chance with a probability of less than 5 per cent. More specific details of these methods may be found in Diggle *et al.* (1995)

This may be illustrated with reference to a classic research problem, relating to the incidence of Burkitt's lymphoma in east Africa. The disease has been the subject of several previous space-time analyses, notably in the west Nile district of Uganda (Siemiatycki *et al.* 1980; Williams *et al.* 1978). Burkitt's lymphoma is a tumour occurring mostly in childhood, with a peak in the 5–8 year age range. Geographically it is largely, though not exclusively, restricted to parts of central Africa, notably in areas of low altitude and where temperatures in the coolest month are greater than 15°C and annual precipitation exceeds 50 cm. It was these geographical and environmental restrictions that led Burkitt originally to suggest a role for malaria in the aetiology (see Lenoir 1985 for this background). More recent work has suggested that intense malaria suppresses the immune system and promotes the multiplication of lymphocytes that had previously been infected with the common Epstein-Barr virus, thereby increasing the likelihood of the development of abnormal, cancerous cells (*ibid.*). Although by no means conclusive, any suggestion of space-time clustering in observed data could imply person-person transmission of the Epstein-Barr virus, which might be expected to be spatially and temporally constrained (Siemiatycki *et al.* 1980).

We used data collected in Malawi between 1977 and 1987, comprising a total of 174 patients (Gatrell *et al.* 1994). A plot of the data (Fig. 5a) shows the preponderance of cases in the south of the country but, of course, gives no temporal sequencing. Recall that the test of space-time interaction involves subtracting the product of separate spatial and temporal K function estimates from that of the combined space-time K function, giving a plot of these differences as a function of both space and time. Results for these data (Fig. 5b) show peaks in this function at relatively large spatial and temporal scales but no evidence of very localized space-time clustering. To evaluate this more formally, a set of 999 simulations was performed, randomly permuting the time 'markers' attached to the cases. A plot of standardized residuals was then constructed (Fig. 5c) which shows relatively large numbers of values in excess of two standard errors. We may also compute the sum of the observed differences between the space-time K function and the product of the separate space and time K functions and compare this with the frequency distribution of the same summary for the simulations. In this case the observed sum of these differences ranked 975 out of 1000 'possible' values (Fig. 5d). This suggests some evidence for space-time clustering – the probability that the observed space-time configuration arose by chance being less than 0·025. This result is similar to that obtained from different space-time analyses of data from the west Nile district of Uganda (presented in Williams *et al.* (1978) and analysed both there and in Siemiatycki *et al.* (1980); see Bailey and Gatrell (1995) for a copy of the data). The Uganda data show some evidence of space-time clustering between 1961–5, though weaker evidence in later years.

These ideas have also been applied in other epidemiological contexts, for example to the study of Legionnaires' disease in Glasgow and Edinburgh (Bhopal *et al.* 1992). Some cases of this disease are known to be associated with particular point sources of contaminated water supplies or malfunctioning air-conditioning or cooling systems. These cases are recognized 'outbreaks'; others are referred to as 'sporadic'. Analysis of so-called *sporadic* cases demonstrated quite significant space-time interaction, however, suggesting that at least some of these might also be related to locationally specific sources of contamination.

Modelling the raised incidence of disease

Testing for spatial clustering – the aggregation of events over and above that due to environmental heterogeneity – needs to be distinguished from the detection of specific 'clusters'. By the latter, we

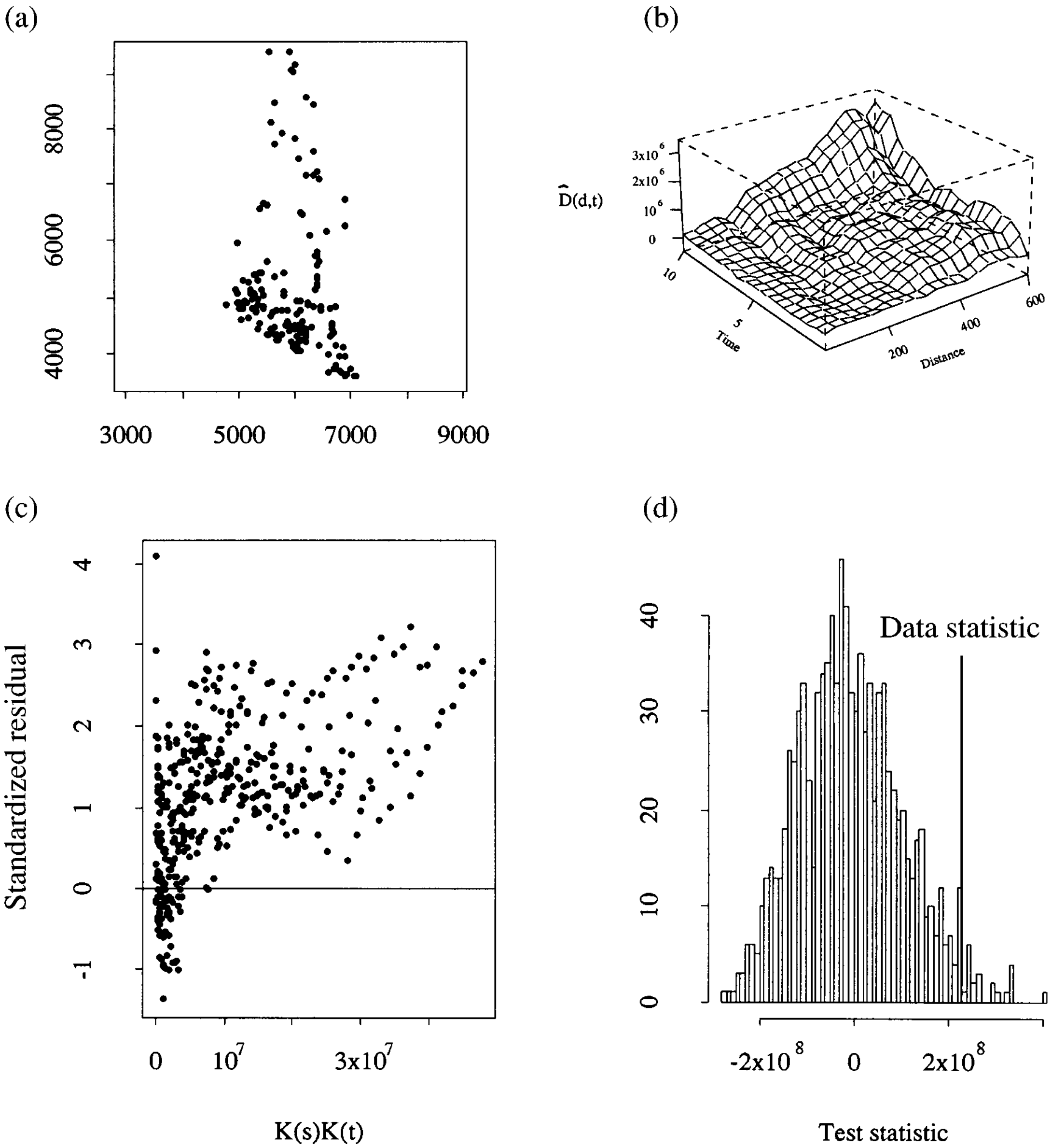

Figure 5 Testing for space-time interaction in Burkitt's lymphoma data, Malawi: (a) data map; (b) D plot; (c) residual plot; and (d) simulation results

mean the assessment of whether there are significant, unusual local aggregations of events. A distinction needs also to be made between the search for clusters in an exploratory data analysis context (as in Openshaw *et al.* 1987 for example) and the testing of a priori hypotheses about possible clusters in the vicinity of fixed locations (see, for instance, Bithell and Stone 1989). We consider only the latter problem here, using a model from Diggle (Diggle 1990; Diggle *et al.* 1990; Diggle and Rowlingson 1994).

An earlier version of the model considered a single possible point source, around which it sought to show whether type 1 events (call them

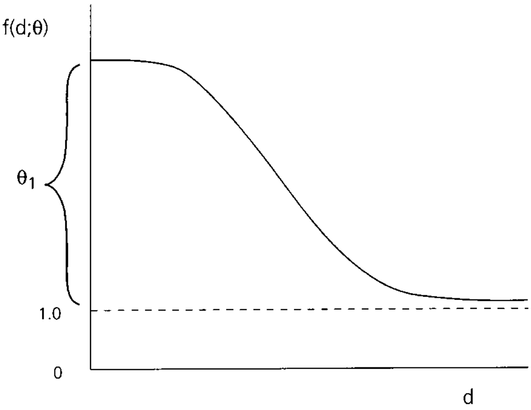

Figure 6 Functional form for Diggle's raised-incidence model

'cases' again) cluster. Diggle (1990) formulated a multiplicative model of the intensity of cases, $\lambda_1(\boldsymbol{s})$, expressing this as a function of environmental heterogeneity (the intensity of controls) and distance from the point source. Formally,

$$\lambda_1(\boldsymbol{s})=\rho\lambda_2(\boldsymbol{s})f(d;\boldsymbol{\theta}) \tag{18}$$

where ρ is a scaling parameter (representing the ratio of the number of cases to controls), $\lambda_2(\boldsymbol{s})$ represents the background intensity and $f(\,)$ is a distance-decay function, involving a vector of parameters, $\boldsymbol{\theta}$, describing how the incidence of cases varies with the distance, d, of the location, $\boldsymbol{s}$, from the point source. In the absence of an elevated risk, we would expect that $f(d;\boldsymbol{\theta})=1$ for all distances, d. As a functional form for $f(\,)$, Diggle postulated

$$f(d;\boldsymbol{\theta})=1+\theta_1 e^{-\theta_2 d^2} \tag{19}$$

where φ_1 and φ_2 are parameters to be estimated, ϕ_1 is an intercept term and φ_2 represents a distance-decay effect (Fig. 6). Diggle suggested estimating background intensity, $\lambda_2(\boldsymbol{s})$, using the kernel estimation method discussed earlier. The parameters, φ, were estimated using maximum likelihood methods. The null model implies an absence of any distance-decay effect. If this is true, then the intensity of cases is simply equal to the background intensity, scaled by the constant ρ.

A more recent version of the model (Diggle and Rowlingson 1994) avoids the need to perform kernel estimation on the controls. The authors derive the probability that an event at $\boldsymbol{s}$ is a case (rather than a control) and express this as

$$p(s)=\frac{\rho f(d;\boldsymbol{\theta})}{(1+\rho f(d;\boldsymbol{\theta}))} \tag{20}$$

with notation as in the previous version of the model.

An advantage of this formulation is that multiple point sources (or, indeed, linear sources) can be added to the model in a multiplicative form. Another extension is to include other covariates that might be available for the cases. For example, we might have age and gender information on the hypothetical children with leukaemia, together with information on parental occupation. If so, the function $f(\,)$ can be replaced with a general formulation:

$$f(d_1,\ldots,\mathrm{d}_q;z_1,\ldots,z_r,\boldsymbol{\theta};\phi)$$
$$=\prod_{i=1}^{q}(g(d_i;\boldsymbol{\theta}))\exp\left(\sum_{j=1}^{r}\phi_j z_j(s)\right) \tag{21}$$

where there are q possible sources at distances d_i from $\boldsymbol{s}$, and r possible covariates z_j at $\boldsymbol{s}$. $g(\,)$ is a suitable distance-decay function involving a vector of parameters $\boldsymbol{\theta}$ and $\varphi=(\varphi 1,..,\ \varphi_r)$ associated with each of the covariates. Maximum likelihood can then be used to generate estimates of the spatial effects, represented by the $\boldsymbol{\theta}$ parameters, along with the aspatial effects, represented by φ.

The motivation for the development of these approaches was the wish to establish whether there was an elevated risk of respiratory cancer in the vicinity of a former industrial waste incinerator in south Lancashire (see Diggle *et al.* 1990 for the background to the research problem). Simple exploratory geographical analysis had suggested that larynx cancer might be elevated, with five cases (from a total of 58 observed in the period 1974–83) observed within 2 km of the site. The first approach expresses the intensity of the disease as a function of two elements: the intensity of background population and distance from the potential source of pollution. Interest centres on whether the intensity of larynx cancer declines with distance from the incinerator and this hypothesis is evaluated against a null model (which states that the intensity of larynx cancer is simply equal to the intensity of background population, scaled by the ratio of cases to controls). The model is fitted using

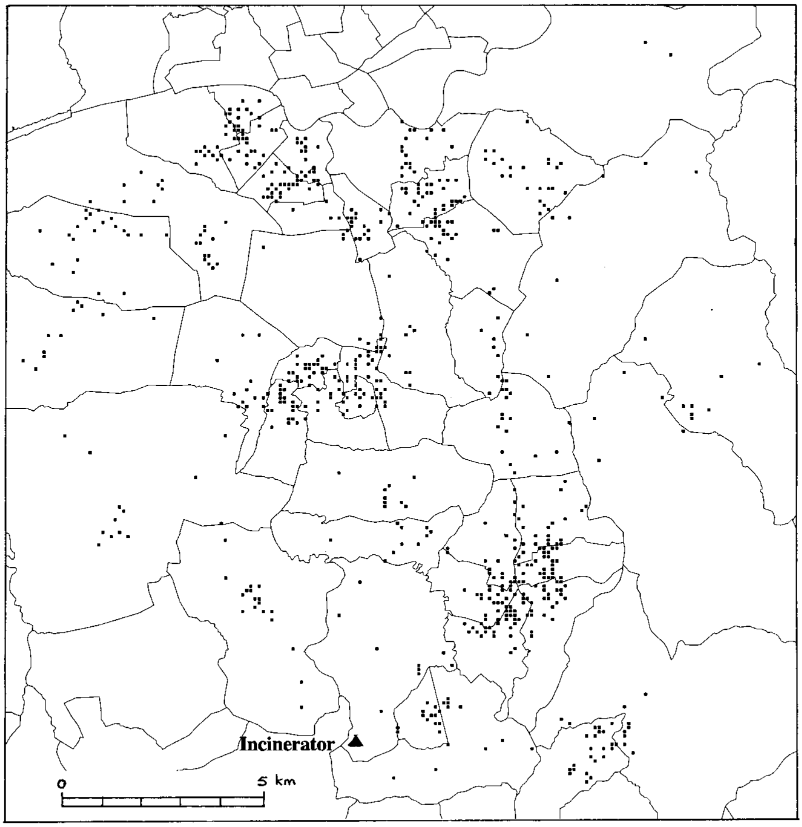

Figure 7 Locations of lung cancers, Chorley and South Ribble, Lancashire, 1974–83

maximum likelihood methods. Initially, we took cases of a more commonly occurring cancer – that of the lung (see Fig. 7) – to represent the controls or background population. Fitting the model yielded estimated parameters $\hat{\theta}_1$=23·67 and $\hat{\theta}_2$=0·91, the former representing the effect at the site, the latter the distance-decay effect. A likelihood ratio test gave a probability of less than 0·01 that the null model was correct, suggesting that there was indeed an elevated risk of larynx cancer within about 2 km of the incinerator.

An important research question is how critical is the choice of lung cancer as a measure of background population? This may be answered by adopting other common cancers, such as stomach cancer, as well as postcode density as alternatives. Regardless of what was chosen, results remained much the same (Gatrell 1990). However, it was recognized that the number of cases was small: deletion of a case near the site reduced the significance of the fit, while removal of two cases failed to give a significant result. Of course, this works both ways: adding a case to the vicinity, to reflect possible out-migration of an individual who presents with larynx cancer in another part of the country, increases the significance of the fit.

Initially, the model was fitted by running FORTRAN programs to both generate the kernel estimate of background population and subsequently to estimate the model parameters. More recently, we have linked the model to the proprietary GIS ARC/INFO (Gatrell and Rowlingson 1994) so that both types of events – cases and controls – may be displayed within the ARCPLOT module and the model fitted by running a macro that simply requires the user to specify the location of the point source of interest, together with starting values for the parameters. The location of the point source is specified by entering a map reference or interactively with a mouse. The latter gives the model sufficient interactive power such that *any* location of interest could be evaluated in order to see whether there is an elevated risk there. Moreover, one can run the model over a regular lattice of such ‘point sources’, resulting in a map of relative risk for the entire study region. When this is done for south Lancashire, only the location of the former incinerator gives a significantly elevated risk. This procedure is similar in spirit to Openshaw’s geographical analysis machine (Openshaw *et al.* 1987).

Following on from this tentative suggestion of raised incidence of laryngeal cancer around one site, the Department of the Environment asked the Small Area Health Statistics Unit (based at the London School of Hygiene and Tropical Medicine) to conduct a study of larynx cancer in the vicinity of other, broadly similar, incinerators. The results (Elliott *et al.* 1992) were based on an alternative methodology (essentially area-based rather than using spatial point process methods) and so are not directly comparable. Nonetheless, they failed to find any evidence of an excess risk in the vicinity of other sites. Whether this is genuinely so, a function of method, or a function of the fact that the point sources examined were not directly comparable has not been established.

One of the limitations of the larynx cancer study was the absence of any useful covariates relating to cases or controls. For example, we had no information about smoking behaviour or alcohol consumption, the two major risk factors for that cancer (Diggle *et al.* 1990). Research on another problem – the incidence of asthma in north Derbyshire (Singleton *et al.* 1994) – indicates how the alternative non-linear binary regression approach (Diggle and Rowlingson 1994) may profitably be used and such covariates incorporated.

There is some evidence to suggest that the incidence of asthma has been increasing in Britain in recent years. Certainly, it represents a public health problem to which much recent research effort has been devoted (see, for example, Alderson 1989). Some writers (for example, Perry *et al.* 1983) have attributed variations in incidence to air pollution in the external environment. In order to see whether there is indeed any association between asthma incidence among children and proximity to potential sources of pollution, a survey was conducted of over 2000 children aged 4–11 years in ten north Derbyshire schools. The incidence of asthma in this population was adjudged by self (parental) reporting and yielded 216 cases, together with 1076 children without asthma, taken to be ‘controls’. A comprehensive questionnaire collected ancillary information on a range of covariates, including exposure to tobacco smoke, presence of pets in the home and problems of dust and mould in the home environment. Three potential point sources of pollution were evaluated: a chemical plant, a coking works and a plant for treating hazardous waste. A non-spatial logistic regression model indicated that the presence of a closed ‘Parkray’ fire in the home had a positive effect on asthma incidence, while an open fire was inversely associated with such incidence (Singleton *et al.* 1994). The (spatial) non-linear binary regression model referred to earlier suggested some elevation of risk around the coking works, though other point sources failed to show such raised relative risk. Inclusion of the obvious covariates (such as exposure to cigarette smoke and to household dust) failed to improve on goodness of fit.

There are alternative software environments available for both the approaches to raised-incidence modelling. One is via SPLANCS (see Appendix) within the framework of S-Plus. This is an attractive option, since maps of the events, together with point sources of pollution being assessed, may be displayed in one window with graphical and statistical output in another. An alternative option is to run a macro from within ARC/INFO. This implements a very general menu-driven system, prompting the user for ‘coverages’ of cases and controls, any covariates that one wishes to include in the model and a set of one or more ‘source locations’ to be evaluated. The latter may be contained within a separate coverage, or entered via the keyboard, or by pointing on the screen. As with the first approach, results comprise parameter estimates – those relating both to the distance effects of the sources as well as any covariates being considered.

Conclusions

In this paper we have described how spatial point patterns can be represented statistically and how their first- and second-order properties may be characterized through the concepts of intensity and second-order intensity. Kernel estimation and, in particular, the use of ratios of kernel estimates was suggested as a means of assessing spatial variation in disease risk. We suggested that second-order properties may be characterized by the *K* function and that this could be extended to consider different types of event. This led us to consider the important topic of how we could detect spatial clustering in the presence of environmental heterogeneity. We also examined the use of *K* functions in a test for space-time clustering. Next, we presented two variants of a model in order to assess whether there is evidence for raised incidence of one type of event in the vicinity of a fixed site, such as a potential source of environmental contamination. Throughout, we have been keen to stress the empirical usefulness of these techniques in an important area of applied geography.

While we have stressed the use of point process methods in an epidemiological context, it is worth drawing attention to other applications of modern point process analysis. For example, Okabe and Sadahira (1994) have examined the spatial association between a set of retail outlets and population distribution in Osaka, Japan. This approach has much in common with the raised-incidence models reviewed earlier: for example, we could take a non-homogeneous distribution of demand, together with the locations of one or more retail outlets attempting to meet that demand. A raised-incidence model would allow us to estimate both the distance-decay effects around the outlet(s) and to estimate their attractiveness. In another area of application, Odland and Ellis (1992) have drawn attention to similarities between raised-incidence models and the spatial form of proportional hazards models used more often in a temporal context. They apply such a model to the spacing of settlements in Nebraska. Hopefully, applications such as these and those considered in this paper will serve to reintroduce geographers to the value of point process methods in their research.

Acknowledgements

Anthony Gatrell and Peter Diggle would like to acknowledge receipt of grants from ESRC (R000232547 and R000234229) which permitted them to explore some of the ideas discussed in this paper, particularly in terms of software developments. Anthony Gatrell is grateful to Lancaster University for granting the study leave that permitted completion of this paper and he and Trevor Bailey would like to thank the Department of Mathematical Statistics and Operational Research, University of Exeter, for funding of research visits. The data on asthma in north Derbyshire were collected by Jo Briggs. The data on Burkitt's lymphoma in Malawi were made available by Dr C van den Bosch.

Note

[1] See, for example, Bailey and Gatrell (1995); Boots and Getis (1988); Cressie (1991); Diggle (1983); Ripley (1981); and Upton and Fingleton (1985).

Appendix: software environments

As noted in the Introduction, applications of modern spatial point process techniques have been hampered, in part, by the lack of suitable software. As will be clear from this paper, in order to apply the range of exploratory methods and to fit possible models, we need good, interactive analytical software that will permit us to visualize the mapped pattern of events, to estimate variation in intensity, to see the empirical functions arising from any second-order analyses, to display results from model-fitting and so on. What software is available to aid such analysis?

One approach is to use S-Plus, a statistical programming language to which may be attached user-written libraries (Becker *et al.* 1988). One such library, 'SPLANCS' (Rowlingson and Diggle 1993), implements all the above methods (and more). Embedded into S-Plus is a set of high-level statistical functions. This feature, coupled with the excellent graphics facilities, makes it an excellent vehicle for interactive spatial analysis. When operating in a windows environment on a workstation, it is possible, for example, to have a map of the events in one window, graphical plots (such as the *K* function) in another and the commands entered in a third. Other windows might be opened if one wished to display the results of kernel estimation or if one wished to display a histogram of the temporal distribution of events in a space-time context. The empirical work has made extensive use of S-Plus and SPLANCS. (SPLANCS is available for UNIX workstations at a small charge from the Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YB. However, the user must, of course, have a licence to run S-Plus. Details of obtaining such a licence may be obtained from Statistical Sciences Ltd, Oxford.)

For those interested in conducting such analyses within a GIS environment, some progress has been made in linking computer code for spatial point process analysis to the proprietary system, ARC/

INFO, running in a workstation environment (Gatrell *et al.* 1994; Gatrell and Rowlingson 1994). Having invoked the GIS, one can call a 'macro' (known as an 'AML' within the particular system under consideration) that itself calls compiled code. Various tools for point process modelling, notably *K* function estimation, kernel estimation and the fitting of raised-incidence models, are part of a so-called 'spatial analysis toolkit' (SAT / 1) that seeks to extend the spatial analytical functionality offered by current releases of ARC / INFO. Details of this are available from Anthony Gatrell.

Finally, as an educational aid, Trevor Bailey has written a package for interactive spatial data analysis called INFO-MAP (Bailey 1990; Bailey and Gatrell 1995). This is designed to perform a range of statistical analyses on small spatial data sets using a minimal PC hardware configuration. Although not a fully fledged GIS, nor containing any facility for linking plots in different windows, it does offer some functionality for much of the analysis we report above. For example, kernel estimation is available, as is the estimation of *K* functions (including those for analysis of bivariate point patterns), together with some simulation functionality. The command language also permits the user to explore some of the ideas concerned with raised incidence. Multiple maps may be plotted on the same screen by saving currently displayed plots to a 'clipboard' and retrieving them subsequently. As a result, we could, for example, compare the effects of bandwidth selection on kernel estimation. Selection of sub-sets of the data is also possible, so we could examine spatial patterns in particular sub-regions of interest. The software accompanies the text produced by Bailey and Gatrell (1995).

References

Alderson M 1989 Trends in morbidity and mortality from asthma *Population Trends* 49 18–23

Alexander F 1993 Viruses, clusters and clustering of childhood leukaemia *European Journal of Cancer* 29 1424–43

Bailey T C 1990 GIS and simple systems for visual, interactive, spatial analysis *The Cartographic Journal* 27 79–84

Bailey T C and Gatrell A C 1995 *Interactive spatial data analysis* Longman Higher Education, Harlow

Baker L 1974 A selection of geographical computer programs Geographical Papers No. 6, Department of Geography, London School of Economics

Barreto M L 1993 The dot map as an epidemiological tool: a case study of Schistosoma mansoni infection in an urban setting *International Journal of Epidemiology* 22 731–41

Bartlett M 1964 The spectral analysis of two-dimensional point processes *Biometrika* 51 299–311

Becker R A Chambers J M and Wilks A R 1988 *The new S language* Wadsworth, Pacific Grove

Beral V Roman E and Gardner M 1993 *Childhood cancer and nuclear installations* British Medical Association Press, London

Bhopal R S Diggle P J and Rowlingson B S 1992 Pinpointing clusters of apparently sporadic Legionnaires' disease *British Medical Journal* 304 1022–7

Bithell J F 1990 An application of density estimation to geographical epidemiology *Statistics in Medicine* 9 691–701

Bithell J F and Stone R A 1989 On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations *Journal of Epidemiology and Community Health* 43 79–85

Boots B N and Getis A 1988 *Point pattern analysis* Sage Scientific Geography Series, Vol. 8, Sage Publications, London

Brunsdon C 1991 Estimating probability surfaces in GIS: an adaptive technique *Proceedings, European Conference on Geographical Information Systems* EGIS Foundation, Utrecht

Cliff A D and Haggett P 1988 *Atlas of disease distributions* Blackwell, Oxford

Cliff A D and Ord J K 1981 *Spatial processes: models and applications* Pion, London

Cressie N A C 1991 *Statistics for spatial data* John Wiley, Chichester

Cuzick J and Edwards R 1990 Spatial clustering for inhomogeneous populations *Journal of the Royal Statistical Society Series B* 52 73–104

Dacey M F 1962 Analysis of central place and point patterns by a nearest neighbour method *Lund Studies in Geography, Series B, Human Geography* 24 55–75

Diggle P J 1983 *Statistical analysis of spatial point patterns* Academic Press, London

Diggle P J 1985 A kernel method for smoothing point process data *Applied Statistics* 34 138–47

Diggle P J 1990 A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point *Journal of the Royal Statistical Society, Series A* 153 349–62

Diggle P J 1993 Point process modelling in environmental epidemiology in **Barnett V and Turkman K F** eds *Statistics for the environment* John Wiley, Chichester

Diggle P J and Chetwynd A G 1991 Second-order analysis of spatial clustering *Biometrics* 47 1155–63

Diggle P J Chetwynd A G Haggkvist R and Morris S 1995 Second-order analysis of space-time clustering *Statistical Methods in Medical Research* 4 124–36

Diggle P J Gatrell A C and Lovett A A 1990 Modelling the prevalence of cancer of the larynx in part of Lancashire: a new methodology for spatial epidemiology in **Thomas R** ed. *Spatial epidemiology* Pion, London

Diggle P J and Rowlingson B S 1994 A conditional approach to point process modelling of elevated risk *Journal of the Royal Statistical Society, Series A* 157 433–40

Draper G J ed. 1991 The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966–83 *OPCS Studies on Medical and Population Subjects* No. 53 HMSO, London

Elliott P Hills M Beresford J Kleinschmidt I Jolley D Pattenden S Rodrigues L Westlake A and Rose G 1992 Incidence of cancer of the larynx and lung near incinerators of waste solvents in Great Britain *The Lancet* 339 854–8

Gatrell A C 1990 On modelling spatial point patterns in epidemiology: cancer of the larynx in Lancashire Research Report No. 9 North West Regional Research Laboratory, Lancaster University

Gatrell A C 1994 Density estimation and the visualisation of point patterns in **Hearnshaw H J and Unwin D J** eds *Visualization in geographical information systems* John Wiley, Chichester 65–75

Gatrell A C Diggle P J van den Bosch C and Rowlingson BS 1994 Space–time clustering of Burkitt's lymphoma in east Africa: methodology and application Research Report North West Regional Research Laboratory, Lancaster University

Gatrell A C Openshaw S Brunsdon C Charlton M Rowlingson B and Rao L 1994 A spatial analysis toolkit for ARC/INFO Paper presented at the Annual Conference of the Institute of British Geographers, Nottingham, 3–6 January

Gatrell A C and Rowlingson B S 1994 Spatial point process modelling in a geographical information systems environment in **Fotheringham A S and Rogerson P** eds *Spatial analysis and GIS* Taylor and Francis, London

Gatrell A C and Whitelegg J 1993 Incidence of childhood cancer in Preston and South Ribble Research Report Environmental Epidemiology Research Unit, Lancaster University

Getis A 1983 Second order analysis of point patterns: the case of Chicago as a multicenter region *The Professional Geographer* 35 73–80

Griffith D A and Amrhein C G 1991 *Statistical analysis for geographers* Prentice-Hall, Englewood Cliffs, NJ

Haggett P Cliff A D and Frey A E 1977 *Locational methods in human geography* Edward Arnold, London

Kelsall J E and Diggle P J 1994 Kernel estimation of relative risk *Technical Report MA94/96* Department of Mathematics, Lancaster University

King L J 1962 A quantitative expression of the pattern of urban settlements in selected areas of the United States *Tijdschrift voor Economische en Sociale Geografie* 53 1–7

Knox E G 1964 Epidemiology of childhood leukaemia in Northumberland and Durham *British Journal of Preventative and Social Medicine* 18 17–24

Lenoir G M 1985 *Burkitt's lymphoma: a human cancer model* IARC, Lyon

Mantel N 1967 The detection of disease clustering and a generalised regression approach *Cancer Research* 27 209–20

McGrew J C and Monroe C B 1993 *An introduction to statistical problem-solving in geography* Wm C Brooks, New York

Odland J and Ellis M 1992 Variations in the spatial pattern of settlement locations: an analysis based on proportional hazards models *Geographical Analysis* 24 97–109

Okabe A and Sadahira Y 1994 A statistical method for analysing the spatial relationship between the distribution of activity points and the distribution of activity continuously distributed over a region *Geographical Analysis* 26 152–67

Openshaw S Charlton M Wymer C and Craft A 1987 A mark 1 geographical analysis machine for the automated analysis of point data sets *International Journal of Geographical Information Systems* 1 335–58

Perry G B Chai H and Dickey D W 1983 Effect of particulate air pollution on asthmatics *American Journal of Public Health* 73 50–6

Raper J Shepherd J and Rhind D W 1992 *Postcodes: the new geography* Longman, Harlow

Ripley B D 1981 *Spatial statistics* John Wiley, Chichester

Rogers A 1965 A stochastic analysis of the spatial clustering of retail establishments *Journal of the American Statistical Association* 60 1094–1102

Rowlingson B S and Diggle P J 1993 SPLANCS: spatial point pattern analysis code in S-Plus *Computers and Geosciences* 19 627–55

Siemiatycki J Brubaker G and Geser A 1980 Space-time clustering of Burkitt's lymphoma in east Africa: analysis of recent data and a new look at old data *International Journal of Cancer* 25 197–203

Silverman B W 1986 *Density estimation for statistics and data analysis* Chapman and Hall, London

Singleton C Gatrell A C and Briggs J 1995 Prevalence of asthma and related factors in primary schoolchildren in an industrial part of England *Journal of Epidemiology and Community Health* 49 326–7

Thomas R W 1993 *Geomedical systems* Routledge, London

Trenhaile A S 1971 Drumlins: their distribution, orientation and morphology *Canadian Geographer* 15 113–26

Upton G and Fingleton B 1985 *Spatial data analysis by example* Vol. 1 John Wiley, Chichester

Williams E H Smith P G Day N E Geser A Ellice J and Tukei P 1978 Space-time clustering of Burkitt's lymphoma in the West Nile district of Uganda: 1961–1975 *British Journal of Cancer* 37 109–21