CHAPTER 4

Spatial clustering of disease and global estimates of spatial clustering

4.1 Introduction

As described in Chapter 3, the ability to visualize spatial data allows for the quick identification of any obvious patterns, and in general, spatial patterns can be classified as regular, random, or clustered. The term 'clustering' is used to describe the spatial aggregation of disease events, but as the observed spatial pattern may simply be a function of the distribution of the population at risk or of various risk factors, a more robust definition is the one proposed by Wakefield et al. (2000), that a disease is clustered if there is 'residual spatial variation in risk after known influences have been accounted for'.

Besag and Newell (1991) classified the different methods for analysing clusters as either specific or non-specific, although epidemiologists prefer to use the terms 'local' and 'global'. Global (non-specific) clustering methods are used to assess whether clustering is apparent throughout the study region but do not identify the location of clusters. They provide a single statistic that measures the degree of spatial clustering, the statistical significance of which can then be assessed. The null hypothesis for global clustering methods is simply that 'no clustering exists' (i.e. random spatial dispersion). Local (specific) methods of cluster detection define the locations and extent of clusters, and can be further divided into focused and non-focused tests. Non-focused tests identify the location of all likely clusters in the study region, while focused tests investigate whether there is an increased risk of disease around a pre-determined point, such as a nuclear power plant or chemical factory. Bear in mind that the term *clustering* applies to global methods of cluster analysis, while *cluster detection* refers to local methods of analysis.

Clustering of a disease can occur for a variety of reasons including the infectious spread of disease, the occurrence of disease vectors in specific locations, the clustering of a risk factor or combination of risk factors, or the existence of potential health hazards such as localized pollution sources scattered throughout a region, each creating an increased risk of disease in its immediate vicinity. The identification and reporting of areas with an apparent increased incidence of disease is known as a *disease cluster alarm*.

In this chapter a brief discussion of disease cluster alarms and cluster investigation protocols is followed by a review of some statistical concepts and terminology relevant to spatial clustering. The chapter concludes with an outline of some of the more commonly-used global clustering methods for point and area data, highlighting the advantages and disadvantages of each technique and exploring their application through a review of their use in the literature.

4.2 Disease cluster alarms and cluster investigation

The investigation of possible disease clustering is fundamental to epidemiology, with one of the aims being to determine whether the clustering is statistically significant and worthy of further investigation, or whether it is likely to be a chance occurrence, or is simply a reflection of the distribution of the population at risk. The statistical significance of clustering is especially important when studying the aetiology of a disease, when conducting disease surveillance programmes or when evaluating disease cluster alarms (Lawson and Kulldorff 1999). The false identification of a cluster in any of these situations may lead to wasted resources, while dismissing a genuine disease cluster can have serious consequences.

Although the reporting of suspected disease clusters is very common, only a minute proportion of these alarms are actually worthy of further investigation. Disease cluster alarms are usually based on a higher observed disease rate, a situation which can, understandably, cause concern among the public. By determining whether the observed clustering is *statistically* significant, disease cluster alarms can either be confirmed or rejected. In this way resources are not wasted on an in-depth assessment of what is frequently random spatial variation, yet at the same time scientific evidence is provided with which to allay public concern. As disease cluster alarms already define the extent of the cluster, thereby introducing pre-selection bias, it is not appropriate to determine the statistical significance of the cluster by simply comparing the standardized incidence or mortality rate within the cluster with that observed in the rest of the region. Statistical evaluation of the disease cluster alarm therefore has to take account of the pre-selection bias, and there are various ways in which this can be done. Either, the spatial scan statistic can be used to analyse data from the whole region in order to determine whether it identifies a significant cluster in the vicinity of the cluster alarm or, if the suspected cluster is the result of a potential health hazard (e.g. a landfill site), a focused cluster analysis (see Chapter 5) can be performed on the same health hazard but in a different area. A third approach is to ignore all prior data collected from the area of the alarm and instead, monitor any new cases that occur in the area. Kulldorff (1999) provides a detailed discussion of these three approaches, highlighting the strengths and weaknesses of each.

Investigating possible disease clusters requires a systematic approach and various cluster investigation protocols are available, details of which can be obtained from the EUROCAT website[39]. A widely adopted method is the one outlined by the Centers for Disease Control and Prevention (CDC 1990), although Rothenberg and Thacker (1992) conclude that this protocol is not sufficiently specific to cover all possible eventualities. The EUROCAT Working Group recommends the Dutch Triple Track approach (Drijver and Melse 1992) for investigating clusters as this protocol proposes simultaneously investigating the health effects and environmental exposures, while communicating with the public, rather than following the sequential approach frequently adopted by other protocols. For instance, the first step (orientation) of the Dutch Triple Track approach to disease cluster investigation requires the collection of general information on the expected disease frequency and potential risk factors (health track), and on local exposure possibilities (environment track), while preparing a visit to the person who contacted the health agency (communication track).

[39] http://www.eurocat.ulster.ac.uk/clusterinvprot.html

4.3 Statistical concepts relevant to cluster analysis

4.3.1 Stationarity, isotropy, and first- and second-order effects

The concepts of stationarity, isotropy, and first-order (trend) and second-order (local) spatial effects are introduced in Chapter 2, and are fundamental to cluster analysis. To summarize, a spatial process is termed stationary if the dependence between measurements of the same variable across space is the same for all locations in an area. If the dependence in a stationary process is affected by distance, but this is the same in all directions, the process is considered to be isotropic. First-order effects describe large-scale variations in the mean of the outcome of interest due to location or other explanatory variables, while second-order effects describe small-scale variation due to interactions between neighbours.

4.3.2 Monte Carlo simulation

Many tests for clustering use Monte Carlo simulation in order to determine the statistical significance of the cluster (i.e. does the observed spatial

pattern differ significantly from the null hypothesis of complete spatial randomness). This involves first calculating the test statistic using the *observed* data, and then re-calculating it using a specified number (e.g. 99, 499, 999) of *simulated* data sets (or permutations). The latter is used to generate the expected distribution of the test statistic under the null hypothesis. The likelihood of obtaining the value for the test statistic derived from the observed data is then calculated, and expressed as the p-value. A Monte Carlo estimate of the p-value for a one-sided test is given by the proportion of test statistic values, obtained from the simulated datasets, that are greater than the value of the test statistic obtained when using the observed data. As Monte Carlo methods rely on permutations of simulated datasets, slightly different p-values are obtained each time the test is run. However, using more simulations to estimate the distribution of the null hypothesis (e.g. 999 versus 99), means that smaller and more stable p-values can be calculated (e.g. p = 0.001 as opposed to p = 0.01). However, a problem with multiple testing is that the likelihood of wrongly rejecting the null hypothesis increases. To compensate the significance threshold needs to be lowered and this is usually done using either a Bonferroni or Simes adjustment.

4.3.3 Statistical power of clustering methods

This chapter discusses some of the more commonly-used cluster analysis methods, and although certain tests are more appropriately used in specific situations, when there are competing methods a commonly asked question is 'which is best'? In such instances, the statistical power of the test (i.e. the probability of detecting clustering or clusters when they actually occur) can provide an answer. Waller and Gotway (2004) discuss various issues affecting the ability of a test to correctly detect clustering, including the structure of the data and the alternative hypothesis. A comparison of the performance of different clustering methods is provided in Walter (1992; 1993), Kulldorff et al. (2003), Song and Kulldorff (2003) and Kulldorff et al. (2006a). When compared with other global clustering tests, Tango's maximized excess events test (MEET) has been found to be the most consistent for evaluating cancer maps (Kulldorff et al. 2006a) and have the highest power of the tests under evaluation (Kulldorff et al. 2003; Song and Kulldorff 2003). Cuzick and Edwards' *k*-nearest neighbour test has also been shown to perform well, especially when the clustering occurrs over small distances (Kulldorff et al. 2003). A detailed discussion of the statistical power of clustering and cluster detection methods can be found in Waller and Gotway (2004).

4.4 Methods for aggregated data

Autocorrelation statistics for aggregated data provide an estimate of the degree of spatial similarity observed among neighbouring values of an attribute over a study area. There are various autocorrelation statistics for aggregated data, four of which will be discussed in this section; Moran's *I*, Geary's *c*; Tango's excess events test and maximized excess events test. Details of other autocorrelation tests can be found in Cliff and Ord (1973; 1981).

Fundamental to all autocorrelation statistics is the weights matrix, used to define the spatial relationships of the regions so that those that are close in space are given greater weight in the calculation than those that are distant (Moran 1950). Neighbours can be defined based either on adjacency or distance. Methods based on adjacency (also known as contiguity) include rook contiguity (polygons are adjacent if they share a border), queen contiguity (polygons are adjacent if they share a border or corner), and higher-order contiguities (often called spatial lags) such as first-order (neighbours) or second-order adjacency (neighbours-of-neighbours). The concept of higher-order adjacencies is illustrated in Fig. 4.1. When using distance to define neighbours, polygons with their centroids located within a specified distance range are considered to be adjacent. More complicated neighbour definitions include row-standardization, length of common boundary, or relationships that only act one way (e.g. large polygons may influence, but not be influenced by, neighbouring, smaller polygons). Software for producing the weights matrix includes GeoDa and ArcView.

Figure 4.1 Maps illustrating (a) first-order adjacency (neighbours), and (b) second-order adjacency (neighbours-of-neighbours). In each instance the black county is the polygon of interest and the light-grey counties the defined neighbours.

4.4.1 Moran's *I*

Moran's *I* coefficient of autocorrelation is similar to Pearson's correlation coefficient, and quantifies the similarity of an outcome variable among areas that are defined as spatially related (Moran 1950). Moran's *I* statistic is given by:

$$I = \frac{n\sum_i\sum_j W_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{(\sum_i\sum_j W_{ij})\sum_k (Z_k - \bar{Z})^2} \quad (4.1)$$

where Z_i could be the residuals $(O_i - E_i)$ or standardized mortality or morbidity ratio (SMR) of an area, and W_{ij} is a measure of the closeness of areas *i* and *j*. A weights matrix is used to define the spatial relationships so that regions close in space are given greater weight when calculating the statistic than those that are distant (Moran 1950).

Moran's *I* is approximately normally distributed and has an expected value of $-1/(N-1)$ (where *N* equals the number of area units within a study region), when no correlation exists between neighbouring values. The expected value of the coefficient therefore approaches zero as *N* increases. Although Moran's *I* generally lies between +1 and -1, it is not bound by these limits (unlike Pearson's correlation coefficient; Waller and Gotway 2004). A Moran's *I* of zero indicates the null hypothesis of no clustering, a positive Moran's *I* indicates positive spatial autocorrelation (i.e. clustering of areas of similar attribute values), while a negative coefficient indicates negative spatial autocorrelation (i.e. that neighbouring areas tend to have dissimilar attribute values). The significance of Moran's *I* can be assessed using Monte Carlo randomization (Moran 1950). Moran's *I* can be calculated using various software packages, including ClusterSeer, R, and GeoDa.

Disadvantages of the autocorrelation test are the assumptions that the population at risk is evenly distributed within the study area (Moran 1950), and that the correlation or covariance is the same in all directions (i.e. it is isotropic). The problems posed by anisotropic data can be overcome by manipulating the weights matrix to reflect directional inequalities.

Although Moran's *I* is intended for use with continuous data it can also be used to analyse count data, even though, in such instances, any observed autocorrelation may simply be the result of variation in regional population sizes rather than any genuine spatial pattern in the disease counts (e.g. if regions with large populations are grouped together). Accounting for the population at risk by using regional incidence rates, instead of regional disease counts, increases the likelihood that any observed autocorrelation reflects a genuine spatial pattern rather than a heterogeneous population distribution.

A spatial correlogram is a series of estimates of Moran's *I* evaluated at increasing distances. Moran's *I* is plotted on the vertical axis and distance (spatial lag) is plotted on the horizontal axis.

The correlogram can therefore be used to determine where, on average, spatial autocorrelation is maximized. Correlograms can be calculated based on distance or adjacency. There are various tests to determine whether a spatial autocorrelation value in a correlogram is statistically significant, some of which are reviewed by Cliff and Ord (1981). A correlogram differs from a semivariogram (described in Chapter 6) in that the former plots correlation against distance, while the latter plots the semivariance against distance (correlation being a measure of the similarity, and semivariance a measure of the dissimilarity, between the variable under consideration and the lagged version of the same variable).

As Moran's I cannot adjust for a heterogeneous population density, Oden (1995) proposed the use of Oden's *Ipop*, a test statistic similar to Moran's I but in which rates are adjusted for population size.

In the following example, Moran's I was used to test whether there was clustering of counties in Great Britain in 1999 with respect to TB incidence rates. Using the GeoDA software, a weights matrix based on queen contiguity (i.e. polygons having a common border or corner) and a spatial lag of one (i.e. first-order adjacency) was defined and used in conjunction with Monte Carlo randomization (999 permutations), to calculate Moran's I for the data. These indicated the existence of non-significant, ($p = 0.085$) positive, spatial autocorrelation ($I = 0.0832$) in the TB incidence rates of neighbouring counties. In order to investigate spatial autocorrelation at higher-order spatial lags, weights matrices were defined and Moran's I statistics calculated for second- to fourth-order adjacencies, and used to plot a correlogram (Fig. 4.2), which illustrates that for first- and second-order adjacencies spatial autocorrelation was positive, but negative for higher-order adjacencies. In other words, neighbouring or nearly-neighbouring counties (spatial lag of one or two) had similar TB incidence rates (either high or low), whereas counties further away from the one of interest (spatial lag of three or four) tended to have dissimilar incidence rates.

There are many examples of the use of Moran's I in the literature. Kitron and Kazmierczak (1997) use Moran's I to investigate the spatial distribution of the incidence of Lyme disease by county, between 1991 and 1994 in Wisconsin State, USA, and analyse clustering in two candidate covariates of Lyme disease; the distribution of the tick vector *Ixodes scapularis* and Normalized Difference Vegetation Index (NDVI) values, derived from the National Oceanic and Atmospheric Administration's (NOAA) Advanced Very High Resolution Radiometer

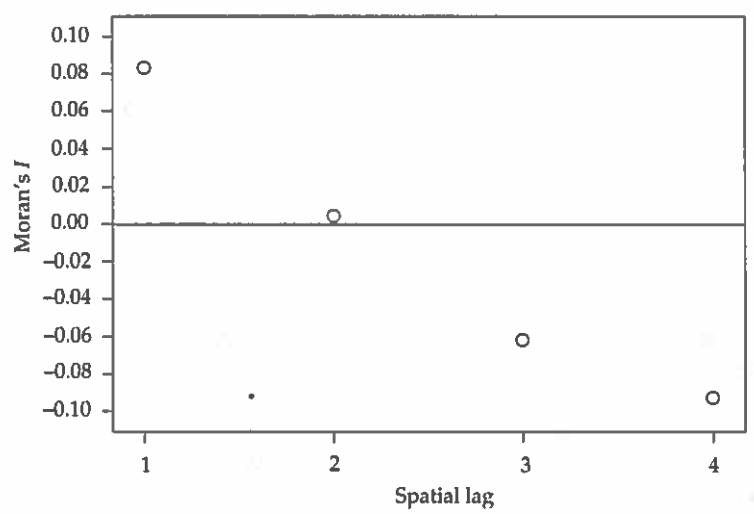


Figure 4.2 A correlogram showing Moran's I statistic computed for TB incidence rates in Great Britain in 1999 at first, second, third, and fourth-order spatial lags.

(AVHRR), during spring and autumn when the contrast between agricultural land and woodland is maximized. Assigning data for each county to its centroid, they calculate Moran's I statistic and use spatial correlograms to evaluate the distances where spatial effects are greatest. This analysis demonstrates significant clustering of counties for mean annual incidence of Lyme disease, tick endemicity, and NDVI values in spring and autumn.

Moran's I is also used by Bellec et al. (2006) to investigate spatial autocorrelation of cases of childhood acute leukaemia in France between 1990 and 2000, and by Perez et al. (2002) to investigate clustering of bovine TB in Argentina. Nødtvedt et al. (2007) use Moran's I to assess autocorrelation of incidence rates of canine atopic dermatitis in Sweden.

4.4.2 Geary's *c*

Geary's contiguity ratio, or Geary's c, is another weighted estimate of spatial autocorrelation (Geary 1954) but whereas Moran's I considers similarity between neighbouring regions, Geary's c considers similarity between *pairs* of regions. Geary's c ranges from zero to two, with zero indicating perfect positive spatial autocorrelation and two indicating perfect negative spatial autocorrelation, for any pair of regions. Geary's c is given by:

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{2\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right)} \quad (4.2)$$

where n is the number of polygons in the study area, w_{ij} the values of the spatial proximity matrix, y_i the attribute under investigation, and the mean of the attribute under investigation.

4.4.3 Tango's excess events test (EET) and maximized excess events test (MEET)

Tango (1995) developed the excess events test (EET) to measure the 'closeness' among regions based on a distance matrix. Tango's EET is a weighted sum of excess events, as the statistic considers the difference between the observed rate of cases in each region and the expected rate, and then weights these differences by a measure of the distance between the regions, with a higher weighting given when the two locations are close.

Unfortunately Tango's EET requires the parameter λ (a measure of the spatial scale of clustering) to be specified, resulting in two problems. Firstly, λ is not generally known *a priori* and therefore several values of λ are tested creating issues of multiple testing. Secondly, choosing a large λ makes the test sensitive to geographically large-scale clustering while a small λ makes it more sensitive to small-scale clustering. To overcome these problems, Tango (2000) proposed the maximized excess events test (MEET). This test statistic searches for the value of λ which gives the smallest p-value of the observed value of the test statistic.

The MEET performs well using the two distance-based exponential weight functions proposed by Tango (2000). However, Song and Kulldorff (2005) evaluate other potential weight functions, concluding that the power of the test improves when using functions that incorporate information on the spatial relationship between areas compared to functions that do not. For example, the weight may be defined purely by Euclidean distance or in terms of spatial contiguity of regions, or it could be adjusted according to population density so that the weight declines faster in urban than in rural areas (Song and Kulldorff 2005).

There are few examples of Tango's EET or MEET in the literature. Fang et al. (2004) use Tango's EET to identify significant clustering of brain cancer mortality rates among adults in the United States between 1986 and 1995, while Oliver et al. (2006) use Tango's MEET to identify significant clustering of prostate cancer incidence in Virginia between 1990 and 1999.

4.5 Methods for point data

4.5.1 Cuzick and Edwards' *k*-nearest neighbour test

Cuzick and Edwards (1990) developed a test for spatial clustering that takes into account the potentially heterogeneous distribution of the population at risk. It is based on the locations of cases and randomly selected controls from a specified region and

includes a spatial scale parameter k, determined by the user. Scale in this instance refers to the number of nearest neighbours, and not geographic distance. For each case, the test counts how many of the k-nearest neighbours are also cases, such that if there are n_1 cases, and $m_i(k)$ represents the number of cases among the k nearest neighbours of case i so that $0 \leq m_i(k) \leq k$, for $i = 1,\ldots n_1$, a test statistic T_k can be calculated as follows:

$$T_k = \sum_{i=1}^{n_1} m_i(k) \tag{4.3}$$

Thus, when cases are clustered, the nearest neighbour to a case tends to be another case and T_k will be large. However, when all cases have controls as their nearest neighbours T_k will be zero. The observed value of T_k can be compared with the distribution of values computed using Monte Carlo randomization of the dataset (Wakefield et al. 2000).

When data are available for the population at risk a modification of T_k is:

$$U_k = \sum_{j=1}^{n_1} (Y_j - E_j) \tag{4.4}$$

where circular regions are centred on each case and the radius of each circular region is chosen so that the expected number cases, E_j, is as close to the pre-defined value of k as possible, and Y_j is the number of cases within each region. Under the null hypothesis the expected value of U_k is equal to zero and the variance may be calculated (Cuzick and Edwards 1990). The Cuzick and Edwards test can be implemented using the ClusterSeer software.

Information on the exact locations of cases and controls is not always available, with locations instead being frequently assigned to the centre of administrative areas such as counties or parishes. As a result of assigning cases and controls to the same area 'ties' arise, precluding the calculation of Cuzick and Edwards' test statistic. Jacquez (1994) proposed an extension to Cuzick and Edwards' method that allows the test to be used in such situations.

A distinct advantage of Cuzick and Edwards' method is that it takes account of the heterogeneous distribution of the population at risk, as cases and controls are selected from the same population. In this way, the existence of any clustering of the population at risk, such as in built-up areas is accounted for. Furthermore, through the careful selection of controls, this method allows confounders to be accounted for (Jacquez 1994). Disadvantages of the test include the fact that the user is required to select a value for the parameter k (Kulldorff et al. 2006a), and that interval data must be categorized as 'case' and 'control' locations resulting in a possible loss of information, although at the same time, this aspect of the test allows for flexibility in defining case and control locations (Ward and Carpenter 2000). For instance, cases may be locations where disease has been detected or alternately, locations at which disease is above a specified threshold, such as the mean disease prevalence (Ward and Carpenter 1995). Although Cuzick and Edwards' test was originally developed for use with point data it can easily be adapted for aggregated data (Song and Kulldorff 2003). When the unit of analysis is aggregated, defining cases and controls can require careful thought in order to avoid the introduction of bias.

Cuzick and Edwards (1990) apply this cluster detection method to explore the distribution of childhood leukaemia and lymphoma diagnosed in North Humberside between 1974 and 1986, an example that highlights the importance of selecting a suitable value of k, as the authors point out that 'formal significance testing requires either foreknowledge of the best k or some adjustment for multiple testing', such as a Bonferroni or Simes p-value adjustment.

There are many examples of Cuzick and Edwards test in the literature. Carpenter et al. (2006), using the Cuzick and Edwards method with multiple definitions for cases, find no evidence of spatial clustering of abortions among Danish dairy herds. Doherr et al. (2002) use the test to identify geographical clustering of BSE in Switzerland in animals born after imposition of the feed ban of December 1990. Perez et al. (2002) investigate clustering of bovine TB in Argentina, for which data are aggregated by county. Counties with a prevalence greater than the median are defined as cases, and those with a prevalence lower than or equal to the median as controls. Overall significance of clustering is assessed using values of k ranging from one to 10. Although no significant overall clustering is detected, significant clustering of case counties is detected at k=1 and k=2 suggesting that clustering of TB in Argentina is present at a relatively small spatial scale, with an elevated likelihood of disease occurrence only in neighbouring, or nearly neighbouring case counties.

4.5.2 Ripley's *K*-function

Second-order analysis describes the spatial dependence between events of the same type. The *K*-function is the most commonly-used method and identifies the distance at which clustering occurs. For an isotropic process with an intensity of λ points per unit area, the *K*-function at distance s may be defined as $K(s)$ such that $\lambda K(s)$ gives the expected number of events within a distance s of an arbitrarily-chosen event. Formally, $K(s)$ is defined as:

$$K(s) = \frac{1}{\lambda^2 R} \sum_{i \neq j} \sum I_s(d_{ij}) \tag{4.5}$$

where R equals the area of a region of interest, d_{ij} is the distance between the ith and jth events in R, and $I_s(d_{ij})$ is an indicator function which equals 1 if $d_{ij} \leq s$ and 0 otherwise. Where spatial autocorrelation is present, each event is likely to be in close proximity to other members of the same event type and, for small values of s, $K(s)$ will be large.

An important assumption of the *K*-function is that there are no first-order effects in the spatial pattern, as any evidence of spatial trend may influence the computed *K*-function. In addition, the variance of $K(s)$ increases with increasing distance, and therefore the *K*-function is suitable for estimating general tendencies toward clustering over distances that are small compared with the size of the region (Diggle 2003). As a rule of thumb it is recommended to restrict the range of s to no greater than 0.5 times the length of the shorter side of a rectangular study area.

Due to variations in the spatial distribution of the population at risk, a *K*-function computed only for cases may not be very informative. Instead, the *K*-function calculated for cases ($K_{case}(s)$) can be compared with one calculated for non-cases (or controls) ($K_{control}(s)$), with the difference between the two functions,

$$D(S) = K_{case}(s) - K_{control}(s) \tag{4.6}$$

representing a measure of the extra aggregation of cases over and above that observed for the non-cases (Diggle and Chetwynd 1991). Monte Carlo randomization can then be used to randomly permute the locations of cases and non-cases/controls, and values of the difference function $D(s)$ computed for each permutation (Chetwynd and Diggle 1998). The upper and lower bounds of these permutations are then plotted together with the observed difference function $D(s)$. Any deviation of $D(s)$ above the envelope formed by the upper and lower bounds indicates significant clustering of cases, relative to non-cases/controls. The software for calculating the *K*-function includes ClusterSeer, or the splancs package (Rowlingson and Diggle 1993) adapted for use in R.

Advantages of using the *K*-function to investigate clustering include the fact that it does not depend on the shape of the study region, and precise spatial locations of events are used in its estimation (Cressie 1993). It also takes into account the density of events in the region of interest, enabling spatial dependence to be compared among groups regardless of event prevalence (Diggle 2003; Broman et al. 2006). Furthermore, the test can accommodate an edge-correction weighting factor, the most commonly used of which was proposed by Ripley (1976), although other edge-correction strategies can be found in Ripley (1988), Stoyan et al. (1995), Diggle (2003), and Waller and Gotway (2004).

In the following example of spatial clustering, the British cattle TB data were used to determine the distance over which clustering of TB-positive holdings was significant in a 60 × 60 km² area in the north-east of Cornwall in 1999. The analysis was performed using the splancs package in R. The TB-status of all holdings in this area is shown in Fig. 4.3a. *K*-functions were plotted for the TB-positive holdings (Fig. 4.3b) and for all the holdings (Fig. 4.3c) in the area of interest. Monte Carlo randomization was then used to randomly permute the locations of cases and non-cases, and values of the difference function ($D(s)$) computed for each permutation. The upper and lower bounds of these

permutations were then plotted together with the observed difference function $D(s)$ (Fig. 4.3d). As described above, any deviation of $D(s)$ above the envelope formed by the upper and lower bounds indicates significant clustering of cases, relative to non-cases. In other words, compared with the spatial distribution of the holding population at risk, TB-positive holdings showed a greater tendency to be aggregated at distances of between 2 and 30 km, with maximum clustering occurring at a distance of 17 km.

O' Brien et al. (2000) use the *K*-function to investigate clustering of different types of neoplasms in humans and dogs in Michigan, USA between 1964 and 1994, concluding that the processes determining spatial aggregation of cases in the two species are not independent of one another. Abernethy et al. (2000) use the *K*-function to identify spatial clustering of poultry flocks affected with Newcastle disease in Northern Ireland between 1996 and 1997.

Broman et al. (2006) apply the *K*-function to investigate the effect of treatment on the clustering of ocular chlamydial infection in children among households in a Tanzanian village. Initially, clustering of households with high levels of infection occurs at distances of less than 2 km, suggesting

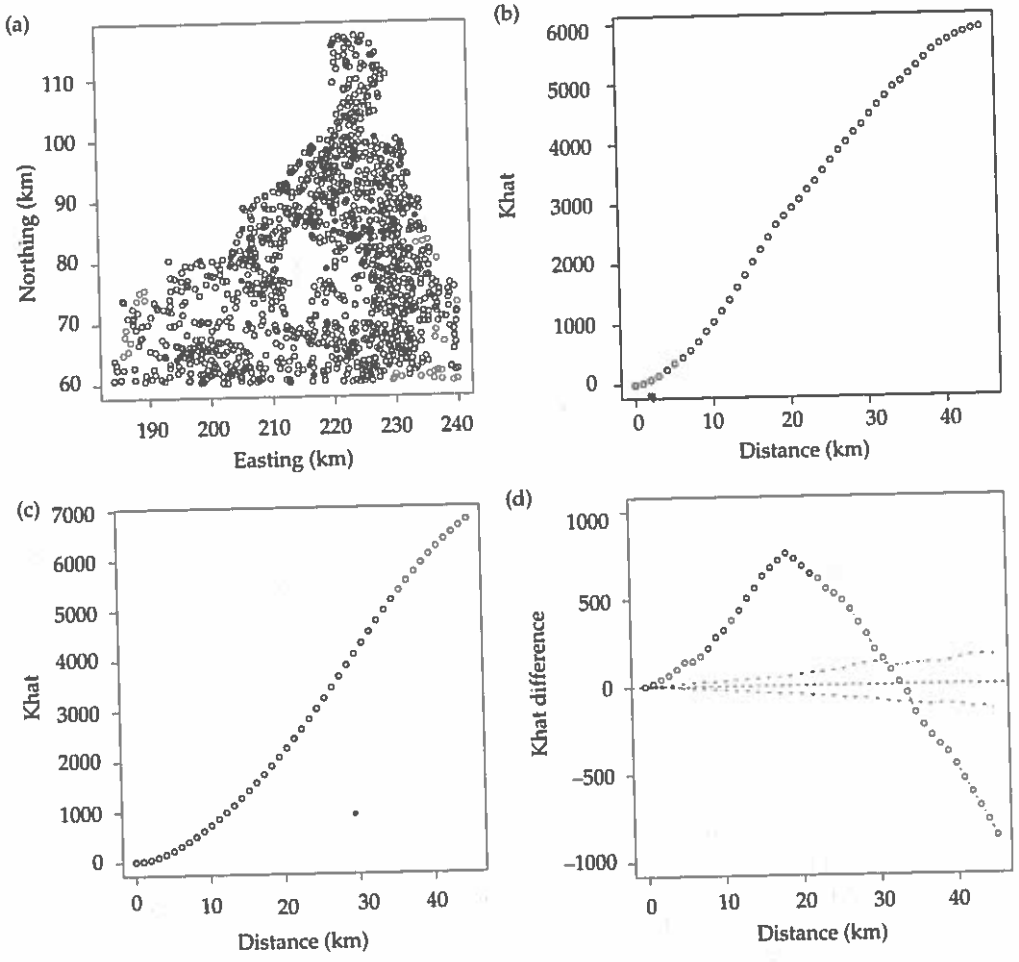


Figure 4.3 (a) Easting and northing coordinates of holdings that tested positive (•) and negative (▫) for TB in a 60 × 60 km² area of Cornwall, (b) *K*-function for TB-positive holdings, (c) *K*-function for all holdings, and (d) the difference between the two *K*-functions.

that either the spread of infection occurs on a relatively small spatial scale or that nearby households share common risk factors for infection. Two months after treatment clustering is no longer evident, but by 12 months post-treatment clustering of households with high infection levels is again apparent, at distances of less than 1.3 km, indicating that treatment was effective at disrupting the spread of infection but not at maintaining low levels of the disease.

4.5.3 Rogerson's cumulative sum (CUSUM) method

Rogerson (1997) developed a cumulative sum (CUSUM) statistic for detecting changes in spatial pattern using a modified version of Tango's statistic (Tango 1995). Whereas Tango's test is used retrospectively to identify clustering, Rogerson's modified version of the statistic aims to detect emerging clusters shortly after they occur, and can therefore be used for spatial surveillance. Owing to the problem of multiple testing, it would be inappropriate simply to re-calculate Tango's statistic after each new observation. Instead, once the test statistic has been determined for a particular set of observations, the expected value and variance of Tango's statistic after the next observation is estimated, based on the current value of the statistic. The expected value and variance is then used to convert the Tango's statistic that is observed after the next observation into a z-score, with all z-scores being incorporated into a CUSUM framework (Rogerson 2006). As Rogerson's CUSUM method is based on Tango's statistic, the test includes a measure of the spatial scale of clustering (λ) and thus, choosing a small value of λ makes the test more sensitive to small clusters and *vice versa*. Although Rogerson (1997) chose to base his CUSUM method on Tango's statistic he suggests that similar CUSUM statistics can be developed for other measures of spatial pattern. For example, Moran's I could be used for aggregated data, or Cuzick and Edwards' method for point data. Rogerson's CUSUM method can be implemented using the ClusterSeer software.

Rogerson (1997) uses his CUSUM version of Tango's test to investigate clustering of cases of Burkitt's lymphoma in Uganda between 1961 and 1975, and finds that spatial clustering exists in specific time-intervals, thereby corroborating the findings of Williams et al. (1978) who had previously obtained similar results analysing the data using the chi-squared test and the Knox test for space–time interaction.

4.6 Investigating space–time clustering

While spatial patterns of disease are of great interest, space–time interactions are also important, particularly so when trying to determine whether a disease is infectious. In such instances it is necessary to evaluate whether cases that are close in space are also close in time and *vice versa*, adjusting for any purely spatial or temporal clustering. Various tests have been developed for this purpose; the tests for global space-time clustering are briefly reviewed in this chapter, while the local space–time cluster detection tests are dealt with in Chapter 5. It is worth noting that some tests look for clusters that relate to a fixed point in space, whilst others allow the spatial focus to move with time. In the first instance, the idea of a two-dimensional circular scan window is extended to that of a cylinder passing through time. In the second case, the geographical focus of a cluster may migrate with time, as long as it relates back to previous events. There is also an important distinction between tests that require knowledge of the population at risk, and those that do not. Not requiring population or control data has obvious advantages in terms of ease of implementation, but has a serious drawback in that it assumes that any change in the population at risk occurs evenly across the distribution under study (Kulldorff 1988). This would obviously have serious analytical implications, where interventions such as culling or vaccination may cause highly inhomogeneous changes in the population at risk.

4.6.1 The Knox test

Knox and Bartlett (1964) developed the first technique to identify spatio-temporal clustering of disease events and, although it has been the subject of much criticism, Knox's test has formed the platform from which subsequent tests have been developed. In this method, pairs of cases separated

by less than a user-defined critical space-distance are considered to be near in space, and pairs of cases separated by less than a user-defined critical time-distance are said to be near in time. This classification allows pairs of points to be assigned to one of four cells in a 2 × 2 contingency table (near space – near time, near space – far time, far space – near time, far space – far time), and a test statistic, T_K is calculated as the number of pairs of cases that are near to one another in both space and time. The test statistic is compared against simulated results under a Poisson model, which Knox argues is the sampling distribution of the statistic in the absence of space-time clustering. Knox and Bartlett (1964) apply this method to data on cases of childhood leukaemia in northeast England, finding significant evidence of space–time clustering.

Jacquez (1996) highlights two principal limitations with Knox's test. Firstly, that the choice of critical distances is subjective and secondly, that the critical distance in space does not vary with changing population density. This is unrealistic since the distance from case to case would decrease with increasing population density. Baker (1996) discusses the problems associated with specifying thresholds of proximity in space and time in the Knox test, and develops an adaptation that does not require unknown critical parameters to be specified, but instead allows for a range to be given for each. In its most flexible form the range can be specified from zero to the maximum space and time differences between any two pairs of cases. The test becomes more powerful as the ranges for thresholds can be specified with increasing accuracy, and reduces to the Knox test itself when the range for each parameter is reduced to zero. He compares this test with a number of examples to which the Knox test had been applied. In the first instance, an exploration of cardiac defects among newborns identifies non-significant space–time clustering, compared with a significant result from the standard Knox test. In the second instance, involving Kaposi's sarcoma in the West Nile district of Uganda, the standard Knox test produces non-significant results whilst re-analysis with Baker's adaptation (Baker 1996) suggests the existence of space–time clustering.

Kulldorff and Hjalmars (1999) review the Knox method, and discuss in some detail its statistical limitations, such as the problems of population-shift bias and the subjective choice of critical thresholds. They propose a modification of the test that overcomes these problems which they demonstrate using cases of lung cancer in New Mexico (1973–1991). Using the standard Knox test, significant space–time clustering is indicated at a range of critical distances whilst with their adaptation it is not. They show that changes in the distribution of the population at risk over time can have a strong influence on the standard test. These effects are particularly interesting with animal diseases where populations may be culled.

More recent applications of the Knox method include Norström et al. (2000), who investigate outbreaks of acute respiratory disease in Norwegian cattle herds, and Tinline et al. (2002), who use the method to estimate the incubation period of racoon rabies.

4.6.2 The space–time *k*-function

Diggle et al. (1995) extend existing second-order analysis methods for spatial data (Ripley 1976; 1977) in order to investigate space–time interactions in point process data. They find second-order properties to be closely related to Knox's statistic (Knox and Bartlett 1964), of which they refer to their test as an extension. If $K_S(s)$ defines the *K*-function in space and $K_T(t)$ defines the *K*-function in time, the *K*-function difference $D(s,t)$ is:

$$D(s,t) = K(s,t) - K_S(s)K_T(t) \tag{4.7}$$

$D(s,t)$ estimates the cumulative number of cases expected within distance s and time-interval t of an arbitrarily-selected case attributable to the interaction between space and time. An alternative expression is:

$$D_0(s,t) = \frac{D(s,t)}{K_S(s)K_T(t)} \tag{4.8}$$

which estimates, for given distance and time separations, the proportional increase in cases attributable to space–time interaction (Diggle et al. 1995).

French et al. (1999) apply the space–time *K*-function to investigate sheep scab outbreaks in Great Britain (1973–1992), revealing strong evidence

that disease occurrence is clustered in both space and time. Wilesmith et al. (2003) and Picardo et al. (2007) both use the space–time *K*-function to investigate different aspects of the 2001 UK FMD epidemic. French et al. (2005) find strong evidence for space–time clustering of equine grass sickness cases using the space–time *K*-function, and suggest that this may be attributable either to a contagious process or to other spatially and temporally localized processes, such as pasture management practices or local climatic effects. Porphyre et al. (2007) use the space–time *K*-function to investigate whether the persistence of TB in a region of New Zealand is the result of contact between infected farms and conclude that there is no evidence of an increased TB risk for those farms close in time and space to TB-positive farms. Software for implementing the space–time *K*-function includes the splancs package in R.

4.6.3 The Ederer–Myers–Mantel (EMM) test

Ederer et al. (1964) developed a cell occupancy approach for exploring space–time clustering whereby the study region is divided into a series of space–time sub-regions within which unusual distributions of cases are sought. They demonstrate the Ederer–Myers–Mantel (EMM) test, using data on cases of childhood leukaemia in Connecticut (1945–1959) and find no evidence of space–time clustering, in contrast to data on poliomyelitis (1940–1954) and infectious hepatitis (1953–1962), both of which did show strong evidence for clustering in space and time. Fosgate et al. (2002) apply the EMM test to evaluate clustering of human brucellosis in California (1973–1977), finding significant space–time clustering that they attribute to clustering of work or food-related disease risk factors.

4.6.4 Mantel's test

Mantel (1967) reviews the methods of Knox and Bartlett (1964) and Ederer et al. (1964), and proposes a new test that compares inter-event distances in space and time against a null hypothesis that time and space distances are independent. The test statistic T_M is the sum, across all pairs of cases, of the spatial distances multiplied by the time distances. A transformation is used to reduce the effects of large space and time distances, which would not be expected to be correlated for contagious diseases. Significance levels are then tested using a standard Monte Carlo randomization process.

Jacquez (1996) highlights a number of limitations of Mantel's test, in particular the linear form of the statistic. For most disease processes a non-linear relationship would be expected between space and time distances. Whilst this can be accounted for by transforming the data, the choice of transformation is subjective.

Wartenberg and Greenberg (1990) compare the statistical power of the tests developed by Ederer et al. (1964) and by Mantel (1967). They suggest that both tests have low statistical power for the typically small numbers of cases involved in such studies, and conclude that true clinical disease excesses, as might result from proximity to a single pollution source, are more likely to be detected by the EMM method whilst the Mantel method is more likely to detect hotspots such as those due to a more general exposure to a putative source.

4.6.5 Barton's test

Barton et al. (1965) designed a test to detect changes in spatial patterns associated with the passage of time, based on analysis of variance and which tests the null hypothesis that these patterns do not change with time. They illustrate their test using three datasets and, although unable to demonstrate significant space–time clustering of cases of childhood leukaemia in northeast England, for which Knox and Bartlett (1964) had previously found evidence of space–time clustering, they did identify space–time associations for measles in Southall in 1954, and poliomyelitis in Eccles in 1974. Ekstrand and Carpenter (1998) use Barton's test to demonstrate significant space–time clustering of flocks with very high prevalence of foot-pad dermatitis in Swedish broilers.

4.6.6 Jacquez's *k* nearest neighbours test

Jacquez (1996) developed a *k* nearest neighbours test for space–time interaction. The null hypothesis is of no association between time and space adjacencies

(i.e. that the probability of two events being nearest neighbours in space is independent of the probability of their being nearest neighbours in time). This approach is based on the argument that geographic distance is not a good measure of spatial proximity in an epidemiological context. Jacquez (1996) compares the statistical power of his test against those of Knox and Mantel using a computer simulation of a viral epidemic at a range of cluster sizes. By plotting statistical power against cluster size he shows that the *k* nearest neighbours test performs considerably better than the other two, of which Mantel's test was superior. He also proposes an adaptation of the *k* nearest neighbours test using fuzzy set theory to accommodate imprecision in the spatial and temporal nearest neighbours.

Norström et al. (2000) compare Jacquez's *k* nearest neighbours to the Knox test in an investigation of outbreaks of acute respiratory disease in Norwegian cattle herds. They find strong evidence of clustering in time as well as space, and express a preference for the Jacquez test since it overcomes the need to specify critical space and time distances. Ward and Carpenter (2000) evaluate a number of space-time cluster tests to investigate possible clustering of blowfly catches on a commercial sheep property in Australia between 1997 and 1998. The tests compared include Jacquez's *k* nearest neighbour test (Jacquez 1996), the Knox test (Knox and Bartlett 1964), Mantel's test (Mantel 1967) and Barton's test (Barton et al. 1965). The Knox test indicates significant spatial (within 3 km) and temporal (within 1 month) clustering, which the other three tests do not identify. Ward and Carpenter (2000) use this to argue for the application of a number of different tests in cluster investigations. Jacquez's *k* nearest neighbour test can be implemented using the ClusterSeer software.

4.7 Conclusion

Clustering of a disease can occur for a variety of reasons, and the investigation of possible disease clustering is fundamental to epidemiology. The concepts and analytical methods discussed in this chapter take the epidemiologist beyond the mere visualization of spatial patterns as the use of statistical methods to assess whether observed patterns differ significantly from spatial randomness moves into the second stage of spatial analysis: that of exploration. Although the techniques outlined in this chapter are all estimates of global clustering, a sound understanding of their different methodologies, data requirements, and associated assumptions and limitations is essential if the most appropriate technique is to be chosen.

CHAPTER 5

Local estimates of spatial clustering

5.1 Introduction

Global measures of spatial association assume that the spatial process under investigation is stationary. Under this assumption, which is rarely met, global tests of association run the risk of obscuring local effects. As a result, significant local clustering may not be detected and conversely, large non-clustered areas within a study area may be ignored. The likelihood of including regions with inherently different local relationships increases as the size of the study area increases. With very large spatial datasets, and particularly with large raster datasets such as remotely sensed images, global statistics such as Moran's *I* (Moran 1948) run the risk of losing information on spatial autocorrelation since they summarize an enormous number of possibly dissimilar spatial relationships. Local statistics overcome this problem by scanning the entire dataset, but only measuring dependence in limited portions of the study area, the bounds of which have to be specified. Thus, for clustering to be detected, it need not occur over the entire dataset, nor need it have the same characteristics throughout the study area.

In studying local area clustering we are concerned with defining the characteristics of clusters, such as their location, size, and intensity. Knox (1989) proposes some definitions of clustering, one of which defines a cluster as 'a geographically and/or temporally bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance'. This purely statistical definition lends itself to this chapter. Whether these occurrences are due to some environmental, biological, or social variable (i.e. they are the direct result of the distribution of risk factors) is the subject of Chapter 7.

The detection of disease clusters, particularly around putative point sources, has been the subject of much debate. In 1989 the Royal Statistical Society reported on a meeting held specifically to discuss the occurrence of cancer near nuclear installations (Muirhead and Darby 1989). The meeting was triggered by the highly popularized reports of excesses of childhood cancer, in particular leukaemia, around the reprocessing plants at Sellafield in Cumbria and Dounreay in northern Scotland. Its objective was to bring together epidemiologists (Doll 1989) and statisticians (Hills and Alexander 1989) to try and resolve whether excesses of cancer really were occurring in the vicinity of nuclear installations.

This triggered a spate of meetings, reported in Rothenberg et al. (1990), Lawson et al. (1995), Cliff (1995b), Jacquez et al. (1996), and Smith (1996). These addressed the legal aspects of disease cluster detection (Henderson 1990; Jacquez et al. 1992), and provided some useful reviews of statistical tests including those by Walter (1993), Waller and Turnbull (1993), Cliff (1995a), and Elliott et al. (1995).

In 1999 the Royal Statistical Society hosted an international conference on the 'analysis and interpretation of disease clusters and ecological studies' (Wakefield et al. 2001), during which the debate on whose method performs best was fuelled, old methodologies were adapted, and new methods introduced. Also addressed at this meeting were more philosophical questions such as when and how to investigate disease clusters, and indeed whether we should bother at all (Elliott and Wakefield 2001; Wartenberg 2001).

The complexity of the issues involving the public, media, local health authorities, epidemiologists, and those responsible for putative sources of increased incidence of disease, is exemplified in a case where the media reacted to documents released by the Ministry of Defence in 1977, disclosing simulation trials of germ warfare along the south coast of England between 1961 and 1977 (Stein 2001). Families in East Lulworth, a coastal village in Weymouth Bay, believed that exposure to the agents released during these trials had caused high rates of miscarriages, stillbirths, birth defects, and learning disabilities in the village. Under pressure from the media, campaigning families, environmental activists, and Members of Parliament the local health authority was forced to conduct a full scale investigation. No evidence of clustering could be found and the 'cluster that never was' was eventually refuted (Spratt 1999).

So 'cluster busting' has a long history and cuts into some difficult statistical, epidemiological, environmental, legal, and social territories. Cautious words on the interpretation of disease clustering studies are given by Besag and Newell (1991), Gardner (1992), Urquhart (1992), and Rothman (1990) who discuss some of the potential pitfalls surrounding cluster detection. Rothman (1990), for example, warns of the difficulties in defining the population base from which incidence rates can be calculated, and points out that due to high levels of publicity often surrounding such cases, unbiased data are difficult to collect. Rothman also warns that investigations into disease clusters are often made for political, rather than for scientific reasons and, quoting from Oakes (1986), that they are often based on significance tests that lack meaningful descriptive statistics and are prone to abuse and faulty inference. Urquhart (1992) highlights the risks of drawing conclusions about clustering (or its absence) in a particular area without knowledge of the underlying distribution of the disease. Besag and Newell (1991) point to the assumptions behind cluster detection tests, emphasizing that the null hypothesis must be based on equal and independent risk to each individual, and that the definition of the population at risk in terms of sex or age-group, and of geographical and temporal boundaries can have a profound impact on the outcome and significance levels. Gardner (1992) further emphasizes the risk of distorting estimates of disease excess, and of their significance levels, by *post hoc* selection of controls. He aptly describes the problem as 'moving the goalposts'.

Many texts have been written on the analysis of spatial point patterns. Examples include Cliff and Ord (1981), Upton and Fingleton (1985), Upton and Fingleton (1989), and Diggle (2003). Reviews of some of the available methods are provided by Marshall (1991a), Elliott et al. (1995), Morris and Wakefield (2000), Robinson (2000), Wakefield et al. (2000), Ward and Carpenter (2000), and Song and Kulldorff (2003). In this chapter some of the more commonly-used methods are reviewed. They are divided into those primarily designed for aggregated data, and those for point data, although this distinction is not rigid and most can be used with either type of data. Moreover, point data can easily be aggregated and area data can be represented as points, for example by using the centroids of the areas. The methods listed for point data are based on variously defined circles that 'scan' the data for areas of elevated (or reduced) disease frequency. Such statistics are better suited to point, than to area data, since points fall clearly inside or outside a scan circle. Scan circles can be defined in terms of geographic distance (e.g. Openshaw's method), number of cases (e.g. Besag and Newell's method) or population size (e.g. Turnbull's method and Kulldorff's scan statistic). Cluster detection is demonstrated by applying Kulldorff's scan statistic to the British cattle TB data. Methods for investigating clusters around point sources are then reviewed, and finally methods that look for clusters in both space and time. The space–time variant of Kulldorff's scan statistic is demonstrated using the British cattle TB data.

5.2 Methods for aggregated data

5.2.1 Getis and Ord's local G*i*(*d*) statistic

Unlike Moran's *I* statistic (see Chapter 4), which measures the correlation between attribute values in adjacent areas, the G*i*(*d*) local statistic (Getis and Ord 1992; 1996) is an indicator of local clustering that measures the 'concentration' of a spatially

distributed attribute variable. Computational details of the of the G*i*(*d*) statistic are provided by Ding and Fotheringham (1992), Getis and Ord (1996), and Kitron et al. (1997). The test statistic is calculated as:

$$Gi(d) = \frac{\sum_{j}^{n} w_{ij}(d)(x_j - \bar{x}_i)}{S_i \sqrt{\frac{w_i(n-1-w_i)}{n-2}}}, \quad j \neq i \tag{5.1}$$

where n is the number of areas within the region of interest and x_i is the observed value for area i

$$\bar{x}_i = \frac{1}{n-1} \sum_{j,\, j \neq i}^{n} x_j \tag{5.2}$$

and w_{ij} is a symmetric binary spatial weights matrix:

$$w_i = \sum_{j,\, j \neq i}^{n} w_{ij} \tag{5.3}$$

$$S_i^2 = \frac{1}{n-1} \sum_{j,\, j \neq i}^{n} (x_j - x_i)^2 \tag{5.4}$$

It has been shown that E(*Gi*) = 0 and Var(*Gi*) = 1, and that the distribution of *Gi* under the null hypothesis of no spatial association among x_i is approximately normal (Ord and Getis 1993). By comparing local estimates of spatial autocorrelation with global averages, the G*i*(*d*) statistic identifies 'hotspots' in spatial data. The G*i*(*d*) statistic and a variety of other similar local statistics are discussed by Getis and Ord (1996), who also highlight some of their shortcomings, particularly those arising as a result of small sample sizes and the existence of high levels of global autocorrelation. Software to implement this statistic includes the spdep package in R, and ClusterSeer.

Getis and Ord (1992) use the G*i*(*d*) statistic to explore the spatial pattern of sudden infant death syndrome (SIDS) by county in North Carolina between 1979 and 1984. In addition to a number of small clusters, they identify a major hotspot in the mid-south of the state which they attribute to the distribution of health care facilities (Getis and Ord 1996). The general pattern of clustering is the same as that found by Grimson et al. (1981) using join count methods.

Kitron et al. (1997) use the G*i*(*d*) statistic to identify significant clustering of cases of La Crosse encephalitis over distances of 5 and 10 km, around particular towns in the Peoria area of Illinois. This analysis reveals a number of hotspots which they associate with homes on the edge of gullies and ravines where breeding and biting activities of the vector (*Aedes triseriatus*) are likely to occur.

Ding and Fotheringham (1992) developed the statistical analysis module (SAM) for ARC/INFO, based on the G*i*(*d*) local statistic. They use SAM to explore clustering of population growth in China between 1982 and 1990 and show that high rates of population growth are concentrated in the south. They attribute this to less severe restrictions on family size and greater proportions of ethnic minorities with large families in this region.

Moving towards exploring spatio-temporal clustering, Getis and Ord (1996) propose the use of local statistics to quantify the pattern and intensity of spread of a disease away from the core of a hotspot by estimating a series of local statistics at different time periods. Local statistics can be used to estimate the intensity of a disease at various distances from a core location, and the time dimension can then be used to estimate the rate of spread. Getis and Ord (1998) use the G*i*(*d*) statistic in this way to trace the spread of AIDS away from San Francisco.

5.2.2 Local Moran test

The local Moran test (Anselin 1995) detects local spatial autocorrelation in aggregated data by decomposing Moran's *I* statistic into contributions for each area within a study region. Termed Local Indicators of Spatial Association (LISA), the LISA statistic for each area is calculated as:

$$I_i = Z_i \sum_{j,\, j \neq i}^{n} w_{ij} Z_j \tag{5.5}$$

where Z_i and Z_j are the observed values in standardized form, and w_{ij} is a spatial weights matrix in row-standardized form.

These indicators detect clusters of either similar or dissimilar disease frequency values around a given observation. The sum of the LISAs for all

observations is proportional to the global Moran's I statistic. There are two uses of LISA statistics: either as indicators of local autocorrelation or as tests for outliers in global spatial patterns in the form of a Moran scatterplot (Anselin 1995; 1996). In a Moran scatterplot the horizontal axis represents the vector of observed values and the vertical axis specifies the weighted average of neighbouring values. The extent of the 'mix' of pairs among the four types of association in the quadrants of the plot defined by the axes (low–high, high–high, high–low and low–low) provides an indication of the stability of the spatial association throughout the data. It may also suggest the existence of different types of association in different subsets of the data; for example, positive association in one area and negative association in another. Software for implementing a local Moran test includes ClusterSeer, GeoDa, SpaceStat[40] and the spdep package in R.

Anselin (1995) demonstrates the local Moran test using patterns of conflict in Africa and compares it with Getis and Ord's local $Gi(d)$ statistic (Getis and Ord 1992). Anselin (1995) also developed LISAs for Mantel's test (Mantel 1967) and Geary's c (Geary 1954). The ClusterSeer and GeoDa software can be used to implement the LISA for Moran's I.

Burra et al. (2002) compare the ability of the local Moran test and Getis and Ord's local $Gi(d)$ statistic to detect clustering in mortality data from Hamilton, Ontario. These authors also evaluate the effect of geocoding errors on pattern detection, concluding that small geocoding errors can significantly influence the results of, and therefore the conclusions drawn from, these statistical tests. Jacquez and Greiling (2003) use the local Moran statistic to analyse spatial clustering in diagnoses of breast, lung, and colorectal cancers on Long Island, USA, identifying significant spatial patterns for all three diseases. Their analysis confirms the clustering of breast cancer mortality previously identified by Kulldorff et al. (1997) (see Section 5.3.4 for details of Kulldorff's spatial scan statistic), but they find that the two methods identify slightly different cluster locations. As a result of these differences Jacquez and Greiling (2003) recommend that a combination of statistics be used when studying local clustering to ensure that different aspects of spatial patterns are identified and to check whether the results from different analyses are consistent. In a similar vein, Hanson and Wieczorek (2002) compare the local Moran statistic with Kulldorff's spatial scan statistic by exploring alcohol-related mortality in New York, USA. The clusters identified differ somewhat between the two methods and they conclude that the scan statistic is the more sensitive method. They advocate a multi-method approach to cluster detection since different methods tend to identify different characteristics of the clusters; the LISA identifying the core of a cluster and the scan statistic identifying its extent. A similar comparison is made by Nødtvedt et al. (2007) in an analysis of the spatial distribution of cases of atopic dermatitis in dogs in Sweden (1995–2000). They conclude that the tests produce similar results but that the LISA statistic identifies more localized clusters, compared to the larger clusters identified by the spatial scan statistic.

The following example continues on from the global Moran's I example presented in Section 4.4.1, which identified the existence of positive spatial autocorrelation among county-level TB-incidence rates in Britain in 1999 ($I = 0.0832$).

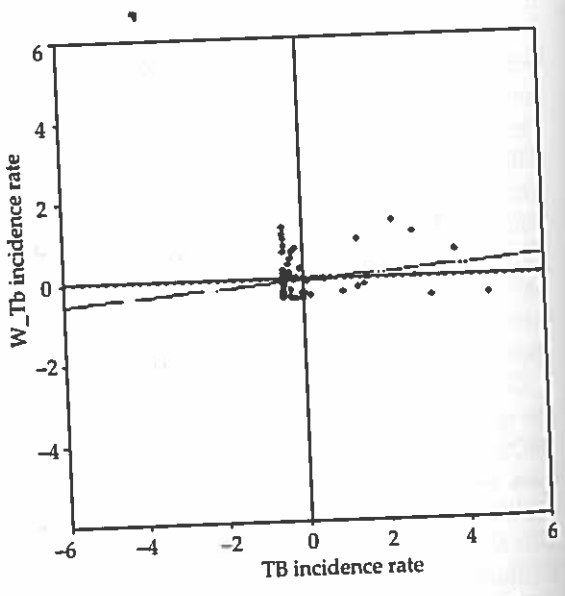


Figure 5.1 Moran scatterplot of county-level TB incidence rates in Britain in 1999.

[40] http://www.terraseer.com/products/spacestat.html

Visual representation of this spatial autocorrelation can be obtained by applying the local Moran test to the data. Examination of the resulting Moran scatterplot (Fig 5.1) showed that most of the points were in the lower left (low-low) and upper left (low-high) quadrants indicating the existence of both positive and negative spatial autocorrelation among county-level TB incidence rates in Britain in 1999; the negative spatial autocorrelation arising as a result of outlier counties with a low TB incidence rate being surrounded by neighbours with high TB incidence rates, while the positive spatial autocorrelation results from counties with a low TB incidence rate having neighbours with similar low incidence rates. The LISA cluster map and LISA significance map are additional, useful methods of visualising the spatial autocorrelation (and can be generated using the GeoDa software (Anselin et al. (2006)).

5.3 Methods for point data

5.3.1 Openshaw's Geographical Analysis Machine (GAM)

Openshaw's Geographical Analysis Machine (GAM) was the first in a series of methods developed to explore disease data for evidence of spatial pattern (Openshaw et al. 1987). The GAM involves applying a fine grid across a study area and generating a series of circles of varying radii with their centres based at each intersection of the grid. The observed number of cases of disease within each circle is then compared with the expected number of cases, assuming the process under investigation follows a Poisson distribution. Circles that have a higher than expected occurrence of disease are retained, resulting in a large number of overlapping circles concentrated around 'disease centres'. Visual inspection is then relied upon to decide where clusters occur. Further details and a quantitative treatment are provided in Openshaw et al. (1990). Openshaw's GAM can be implemented using the Dcluster package in R. Openshaw et al. (1988) and Turnbull et al. (1990) apply the GAM to investigate clustering of acute lymphoblastic leukaemia in northern England and leukaemia incidence in northern New York, respectively.

Statistical and computational limitations of the GAM methodology are described by Cuzick and Edwards (1990), Turnbull et al. (1990), Marshall (1991a), and Besag and Newell (1991). The main criticism of the technique is that it does not account for multiple testing, and therefore the change in radius and shifts in location of candidate clusters are not taken into account in the calculation of significance levels. Despite having being widely criticized, the GAM seems to have inspired the development of a series of tests for cluster detection based on scan circles, which have addressed and overcome many of the deficiencies of the original GAM, gradually adapting and improving upon its basic methodology.

5.3.2 Turnbull's Cluster Evaluation Permutation Procedure (CEPP)

Turnbull et al. (1990) developed the first test that was able to both locate and test the significance of disease clusters. The test, called the Cluster Evaluation Permutation Procedure (CEPP), creates a circular window for each area that contains a pre-determined number of individuals at risk, R. The number of individuals at risk in an area can be defined as p_i and the number of disease events as O_i. If the number of individuals at risk (p_i), is less than R, then area i is included in the window and the area whose centroid is nearest to that of area i (say cell j) is included if $O_i + O_j < R$. If $O_i + O_j > R$ a fraction of the population of area j is added so that the total population at risk equals R. If $O_i > R$ then the window contains only a fraction of area i. For the fractional area included in a window, cases are allocated in the same proportion to the window as that of the population of the cell. A series of i overlapping windows is created with a population at risk of constant size, R. Turnbull's test statistic, as a function of R, equals the maximum number of cases in each of the windows. Monte Carlo simulation is used to evaluate the significance of the observed test statistic. Software for implementing the CEPP includes the Dcluster package in R, and ClusterSeer.

One of the conditions of this procedure is that cluster size (in terms of number of individuals at risk) must be defined *a priori* for the procedure to

be valid. Turnbull et al. (1990) apply the CEPP to leukaemia incidence data obtained from the New York State Cancer Registry, for an upstate New York region comprising eight contiguous counties, and compare its performance with Openshaw's GAM, concluding that the main advantage of the CEPP over the GAM is that it provides a quantitative assessment of the statistical significance of identified clusters.

5.3.3 Besag and Newell's method

One disadvantage of the GAM is that, since distance is the method used to define the scan circles, circles of the same size can refer to different-sized populations and are therefore not directly comparable. Besag and Newell (1991) developed a method to rectify this problem whereby the user specifies k, the expected cluster size. Typical values for k range between 2 and 10 for rare diseases. Each area with non-zero cases is considered in turn as the centre of a possible cluster. When evaluating an area it is labelled as 0 and the remaining areas are ordered according to their distance from area 0 and labelled 1, 2,..., $i-1$. Using the notation defined in the previous section (O_i is the number of disease events) the statistic D_i is calculated such that $D_i = \sum_{j=0}^{i} O_j$, O_j and $D_0 \leq D_1 \leq \ldots$ are the accumulated number of cases in cells 0, 1, ... and $u_0 \leq u_1 \leq \ldots$ are the corresponding accumulated number of individuals at risk. $M = \min\{i: D_i \geq k\}$ so that the nearest M areas contain the closest k cases. A small observed value of M indicates a cluster centred at cell 0. If m is the observed value of M, then the significance level of each potential cluster is:

$$\Pr\{M \leq m\} = 1 - \sum_{s=0}^{k-1} \exp(-u_m Q) (u_m Q)^s / s! \qquad (5.6)$$

where Q equals the total number of cases in the study area divided by the total population at risk. The test statistic of clustering in the study area, T_{BN}, equals the total number of individually significant clusters, for example, at $p < 0.05$. The significance of the observed T_{BN} can be determined by Monte Carlo simulation. Further details of the methodology are provided by Besag and Newell (1991) and Alexander and Cuzick (1992). Software for implementing Besag and Newell's test includes the Dcluster package in R, and ClusterSeer.

Two limitations of this method, common to many other cluster detection tests are firstly, the *a priori* choice of cluster size and secondly, the problem of multiple testing arising from the large number of potential clusters. Since the calculations to determine significance are based on the specified number of nearest-neighbour cases, selecting its value is an important issue. If it is too small then larger clusters cannot be detected, and if it is too large spurious clusters may be identified (Le et al. 1996). Besag and Newell (1991) demonstrate their methodology using data for acute lymphoblastic leukaemias in children under the age of 14 years in part of the Mersey Regional Health Authority Districts of England (1975–1985). They identify 23 clusters significant at the 5% level, which fall into seven discrete groups. Interestingly, and in agreement with the Black Report (Black 1984), they find no evidence of clustering in the vicinity of the Sellafield nuclear reprocessing plant during this period (see Section 5.4 for more details on this issue).

Alexander et al. (1991) adapted Besag and Newell's method into the nearest neighbour areas (NNA) test for local clustering (formulae are provided in Alexander et al. (1991)). They apply the NNA test to data on the incidence of selected childhood cancers and to adult haematopoietic malignancies in Yorkshire, UK, confirming an unusual distribution of lymphoma in Yorkshire associated with the North Yorkshire moors. Alexander et al. (1989) apply the NNA test to data on the incidence of Hodgkin's disease in England and Wales (1984–1986), finding weak but significant evidence of localized spatial clustering, particularly in young adults.

Fotheringham and Zhan (1996) review and compare the tests developed by Openshaw and Besag and Newell. These authors use a database on housing quality in the city of Amherst, New York, containing 277 very low quality residences among the total population of 28,832. By using the two different methods to identify clusters of low quality residences they conclude that, although both tests perform reasonably well in identifying genuine

clusters, the specificity of Besag and Newell's test is superior to that of the GAM.

More recently Huillard d'Aignaux et al. (2002) use Besag and Newell's method to explore the possibility of spatial clustering in sporadic Creutzfeldt-Jakob disease (CJD) in France between 1992 and 1998, identifying five clusters that were persistent over a range of values of k. Only one of these clusters is identified by Kulldorff's scan statistic (see Section 5.3.4). Using a simulated benchmark dataset Marcelo and Renato (2005) compare Besag and Newell's test with Kulldorff's scan statistic, finding that they give similar results but that the scan statistic is more likely to identify clusters in sparsely populated areas.

5.3.4 Kulldorff's spatial scan statistic

Inspired by Openshaw's GAM (Openshaw et al. 1987) and a generalization of Turnbull's CEPP (Turnbull et al. 1990), Kulldorff developed the spatial scan statistic (Kulldorff and Nagarwalla 1995), which brings together the advantages of each technique. For each specified location a series of circles of varying radii is constructed. Each circle absorbs the nearest neighbouring locations that fall inside it and the radius of each circle is set to increase continuously from zero until some fixed percentage of the total population is included. For each circle the alternative hypothesis is that there is an elevated risk of disease within the circle compared to that outside (Kulldorff et al. 1998b). The test statistic T_{KN} is calculated as:

$$T_{KN} = \sup_z \left(\frac{O(Z)}{p(Z)}\right)^{n(Z)} \left(\frac{O(Z^c)}{p(Z^c)}\right)^{n(Z^c)} I\left(\frac{O(Z)}{p(Z)} > \frac{O(Z^c)}{p(Z^c)}\right) \qquad (5.7)$$

where Z^c indicates all circles except for Z, $O()$ and $p()$ are the observed number of cases and the population size in each area respectively, and $I(.)$ is the indicator function. Monte Carlo simulation is conducted to compare T_{KN}, with the distribution of values generated under the null hypothesis.

Kulldorff (1997) implemented the spatial scan statistic in a cluster detection programme called SaTScan[41], which searches for clusters in datasets using two different probabilistic models; a Bernoulli model where cases and controls are compared as Boolean variables, and a Poisson model where the number of cases is compared to the background population data and the expected number of cases in each unit is proportional to the size of the population at risk. Circle centres are defined either by the case and control/population data or by specifying an array of grid coordinates. Secondary clusters are computed, based on the degree of overlap allowed in the cluster circles, and includes the options no geographical overlap, and no cluster centres in other clusters. Software for implementing the spatial scan statistic includes SaTScan, the Dcluster package in R, and ClusterSeer.

Until recently a major limitation of the spatial scan statistic was the use of a circular scanning window which decreased the likelihood of the statistic detecting non-circular clusters. However, the most recent version of SaTScan can implement an elliptic version of the spatial scan statistic, which uses a scanning window of variable location, shape, angle and size, thereby greatly increasing the ability of the statistic to detect non-circular clusters (Kulldorff et al. 2006b). Choosing the shape of the ellipsoid will depend on the nature of the data and the questions being asked, and the authors stress that the type of ellipsoid to be used must be chosen before looking at the data in order to prevent any pre-selection bias. For example, a long, narrow ellipsoid might be used to investigate potential clusters along a riverbank. The circular and elliptic scan statistics have similar power, with the circular scan statistic able to detect elliptic clusters and *vice versa*, although the elliptic scan statistic may provide a better estimate of the true area of the cluster. Although the elliptic scan statistic is more flexible than the circular scan statistic it still imposes a shape on potential clusters and therefore, for irregularly-shaped clusters (e.g. along a winding river) it may be more appropriate to use one of the non-parametric spatial scan statistics described in Section 5.3.5.

Kulldorff and Nagarwalla (1995) use the spatial scan statistic, based on a Bernoulli model, to detect clusters of leukaemia cases in northern New York,

[41] http://www.satscan.org

USA. The results of this analysis are compared with a number of other cluster-detection algorithms, including those of Turnbull et al. (1990) and Openshaw et al. (1987). One primary cluster and four non-overlapping secondary clusters are identified in similar locations to those identified visually using Openshaw's GAM in an analysis of the same data (Turnbull et al. 1990).

Hanson and Wieczorek (2002) use the spatial scan statistic (which they compare to a LISA, see Section 5.2.2) in a study of alcohol-related mortality in New York, USA. Huillard d'Aignaux et al. (2002) explore the possibility of spatial clustering of sporadic CJD in France between 1992 and 1998. Using the spatial scan statistic they identify only one significant cluster of sporadic CJD yet identify five disease clusters using Besag and Newell's method. Kulldorff et al. (2003) find the spatial scan statistic performs well when compared against two global tests, Tango's maximized excess events test (Tango 1995; 2000) and the nonparametric M statistic (Bonetti and Pagano 2004), on a collection of simulated datasets generated under different cluster models. They encourage other investigators to contribute to a common bank of cluster-simulation datasets, and to use this collection as a benchmark against which to evaluate both existing and new cluster-detection tests.

Early applications of the scan statistic include an investigation of childhood leukaemia in Sweden (Hjalmars et al. 1996) in which no evidence was found to support previous assertions of disease clustering. Kulldorff et al. (1997) demonstrate clustering of breast cancer mortality in the northeastern part of the USA (1988–1992) after adjusting for confounding variables such as age, race, urbanicity, and parity.

Recuenco et al. (2007) compare clustering of enzootic racoon rabies in New York State (1997–2003) under different regimes of covariate adjustment: (i) no adjustment, (ii) adjustment for landscape covariates, and (iii) adjustment for landscape covariates and large-scale geographic variation. Clusters identified under the unadjusted test regime tend to also be identified in the covariate-adjusted analyses, though some drop from significance. They use these differences in clusters identified under the three regimes to make inferences regarding the possible reasons for high-risk areas.

The spatial scan statistic has now been applied to an impressive range of health-related problems[42]. This is due in part to its addressing many of the problems of earlier scan statistics but also because SaTScan, the freely available primary software for implementing the spatial scan statistic, makes it accessible to a wide, non-specialized audience. There is insufficient space here to review these works in detail (although salient points are raised where appropriate) but the method has been applied to identify clustering in a variety of cancers (Viel et al. 2000; Jemal et al. 2002; Roche et al. 2002; Boscoe et al. 2003; Buntinx et al. 2003; Gregorio and Samociuk 2003; Klassen et al. 2005; Jung et al. 2006), sexually transmitted diseases (Jennings et al. 2005; Wylie et al. 2005); variant CJD (Cousens et al. 2001), systemic lupus erythematosus (SLE) (Walsh and DeChello 2001), human granulocytic ehrlichiosis (HGE) (Chaput et al. 2002), insulin-dependent diabetes mellitus (Green et al. 2003), childhood mortality (Sankoh et al. 2001), low birth rate (Ozdenerol et al. 2005), congenital malformations (Forand et al. 2002), SIDS (George et al. 2001), highland malaria (Brooker et al. 2004), and systemic sclerosis (Walsh and Fenster 1997). In the veterinary literature the method has been applied to psoroptic sheep scab (Falconi et al. 2002), bovine TB (Perez et al. 2002; Olea-Popelka et al. 2003; Olea-Popelka et al. 2005), BSE (Stevenson et al. 2000; Abrial et al. 2003), sylvatic plague agents in coyotes (Hoar et al. 2003), toxoplasma infections in sea otters (Miller et al. 2002; 2004a), and atopic dermatitis in dogs (Nødtvedt et al. 2007), to name but a few.

5.3.5 Non-parametric spatial scan statistics

A significant restriction imposed by all of the methods reviewed here is that they assume disease clusters are circular. As these conditions seldom occur in reality, attempts have recently been made to overcome this problem by developing a variety of tests that can locate irregularly-shaped disease clusters, including those described by Duczmal and Assunção (2004), Patil and Taillie (2004), Tango and Takahashi (2005), and Duczmal et al. (2006).

Spatial scan statistics use circles for the scanning window, that have a low power for detecting

[42] http://www.satscan.org/references.html

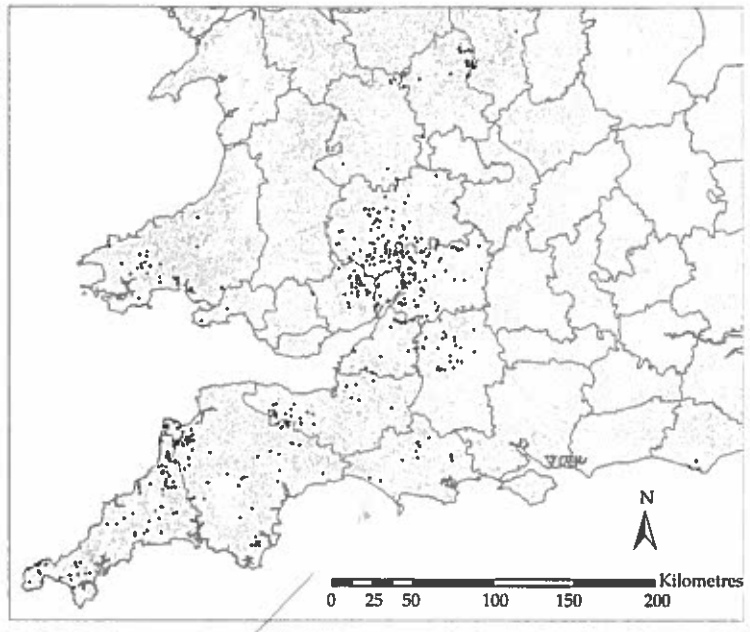


Figure 5.2 Distribution of herds (small grey spots) and tuberculosis breakdowns (large black spots) in the southwest of Britain in 1997.

irregularly-shaped clusters, and may in fact identify an irregularly-shaped cluster as a series of small circular clusters. In an attempt to overcome this and other limitations of spatial scan statistics Patil and Taillie (2004) developed the upper level set scan statistic which they apply to three ecological situations in order to illustrate its uses: the early detection of biological invasions, mapping vegetative disturbance, and the characterization of biological impairment of rivers and streams in Pennsylvania.

Duczmal and Assunção (2004) use a simulated annealing approach to identify 'connected clusters with arbitrary shape'. However, Tango and Takahashi (2005) suggest that although Duczmal and Assunção's (2004) test can detect irregularly-shaped clusters, they are much larger than the true clusters and they therefore developed a flexibly shaped spatial scan statistic for detecting irregular-shaped clusters within small areas of a region. This statistic works well at detecting small circular and non-circular clusters (Tango and Takahashi, 2005) and can be implemented using the FlexScan software.[43]

[43] http://www.niph.go.jp/soshiki/gijutsu/download/flexscan/index.html

5.3.6 Example of local cluster detection

To demonstrate local cluster detection Kulldorff's spatial scan statistic was applied to the British cattle TB breakdown herd data (1986–1997). Due to the nature of the TB data (see Chapter 1), incidence rates cannot be accurately derived, and therefore the analysis was restricted to the use of the Bernoulli model. All TB breakdown herds were included as cases, and compared against a sample of control herds (due to computer memory limitations) and the combined data were used to define the circle centres for cluster detection (rather than specifying a special set of grid coordinates). Based on some preliminary analyses a 30% sample of control herds was chosen, above which no real change in the pattern of clustering occurred. The most restrictive option was selected for dealing with overlapping clusters, in which secondary clusters were reported only if they did not overlap with a previously reported cluster.

Fig. 5.2 shows the distribution of herds (small grey dots) and of breakdown herds (larger black dots) in the southwest of Britain in 1997. In order to compare various options for implementing the spatial scan statistic, the 1997 dataset was used as

this was the year with most recorded breakdowns (amongst those years within the dataset).

The single most important subjective choice that has to be made when using the spatial scan statistic is specification of the maximum percentage of the population at risk (between 1 and 50%) that can be included in any one cluster. The SaTScan manual (Kulldorff 2003) recommends specifying a high upper limit (i.e. 50%) of the population at risk, since SaTScan will then look for clusters of both small and large sizes without any pre-selection bias in terms of the cluster size. When looking for clusters of high rates a cluster of larger size indicates areas of exceptionally low rates outside the circle rather than an area of exceptionally high rates within the circle. A review of the literature reveals that most authors do not indicate the upper limit specified in their analyses, presumably using the default value of 50%. Studies stating explicitly that an upper limit of 50% was selected include Walsh and Fenster (1997), George et al. (2001), Huillard d'Aignaux et al. (2002), Roche et al. (2002), Brooker et al. (2004), and Schwermer et al. (2007). A variety of values below this are also reported including 10% by Norström et al. (2000) (to avoid scanning outside the geographic region of the study), 10% by Sheehan et al. (2000) (no reason given), 3.4% by Walsh and DeChello (2001) (based on the population size of the largest county in their study), 5% by Falconi et al. (2002) (no reason given), 2.5% by Forand et al. (2002) (in order to provide better focus on geographical areas for further evaluation and follow up), 5% by Sauders et al. (2003) (no reason given), and 25% by Recuenco et al. (2007) (no reason given).

The results of specifying different upper limits for cluster size when using the British cattle TB data for 1997 (with a 30% sample of control herds) are shown in Fig. 5.3 in which the selected upper limits were (a) 1%, (b) 5%, (c) 10%, (d) 20%, (e) 30% and (f) 50%.

The 1% upper limit produced ten highly significant ($p < 0.001$) and ten less significant ($1.0 > p > 0.001$) clusters. The 50% upper limit produced one enormous, highly significant cluster, one tiny highly significant cluster and one very small, less significant cluster. Moving from an upper limit of 1% towards one of 50% produced smaller numbers of increasingly large and increasingly significant clusters. Which of these patterns of clustering best reflects the epidemiology of TB in the southwest of Britain cannot be determined from this analysis.

To explore the outcomes of the cluster analysis in greater detail consider Table 5.1 and the associated Fig. 5.4. Table 5.1 shows some statistics for the (arbitrarily chosen) 5% maximum cluster size of the 1997 TB data (Fig. 5.3b).

It is quite clear that, as suggested by Boscoe et al. (2003), some of the most significant clusters are of relatively low risk and *vice versa*. Cluster 8 for example, has by far the highest prevalence but is barely significant, while the similarly sized Cluster 3 also has a high prevalence but is highly significant. The very large, highly significant and so called primary cluster (Cluster 1), only has a relatively low prevalence.

Fig. 5.4 illustrates these points through two contrasting representations of the analysis: the log likelihood ratio, which indicates the probability of a cluster being real (Fig. 5.4a), and the prevalence (Fig. 5.4b), which is directly proportional to the relative risk within a cluster but corrected for sampling of the population.

Cluster analysis was conducted for each year from 1986 to 1997. Non-overlapping clusters were identified using the Bernoulli model with all TB breakdown herds as cases compared against 30% of control herds, and a maximum cluster size set to 5% of all points. Fig. 5.5 shows the evolution of TB clusters during this period. Annual clustering of TB breakdowns is represented by the log likelihood ratios, using the same scale for each year. For most of the period two large and highly significant clusters were fairly constant. These were Cluster 1 (as labelled in Fig. 5.4) on the Gloucestershire, Hereford, and Worcester boundaries, and Cluster 2 on the Devon/Cornwall boundary. Cluster 7 in Dyfed in western Wales is much smaller and of lower significance but is persistent, falling from significance only during 1987 and 1989. Some clusters tended to appear, disappear, and then re-appear in much the same place. For example, Cluster 8 in East Sussex was apparent between 1989 and 1991, in 1994 and again in 1997, while Cluster 3 on the Staffordshire/Derbyshire boundary was first evident briefly in 1993 and then re-emerged between 1996 and 1997. Perhaps these are the main centres of endemicity that act as sources of infection via, for

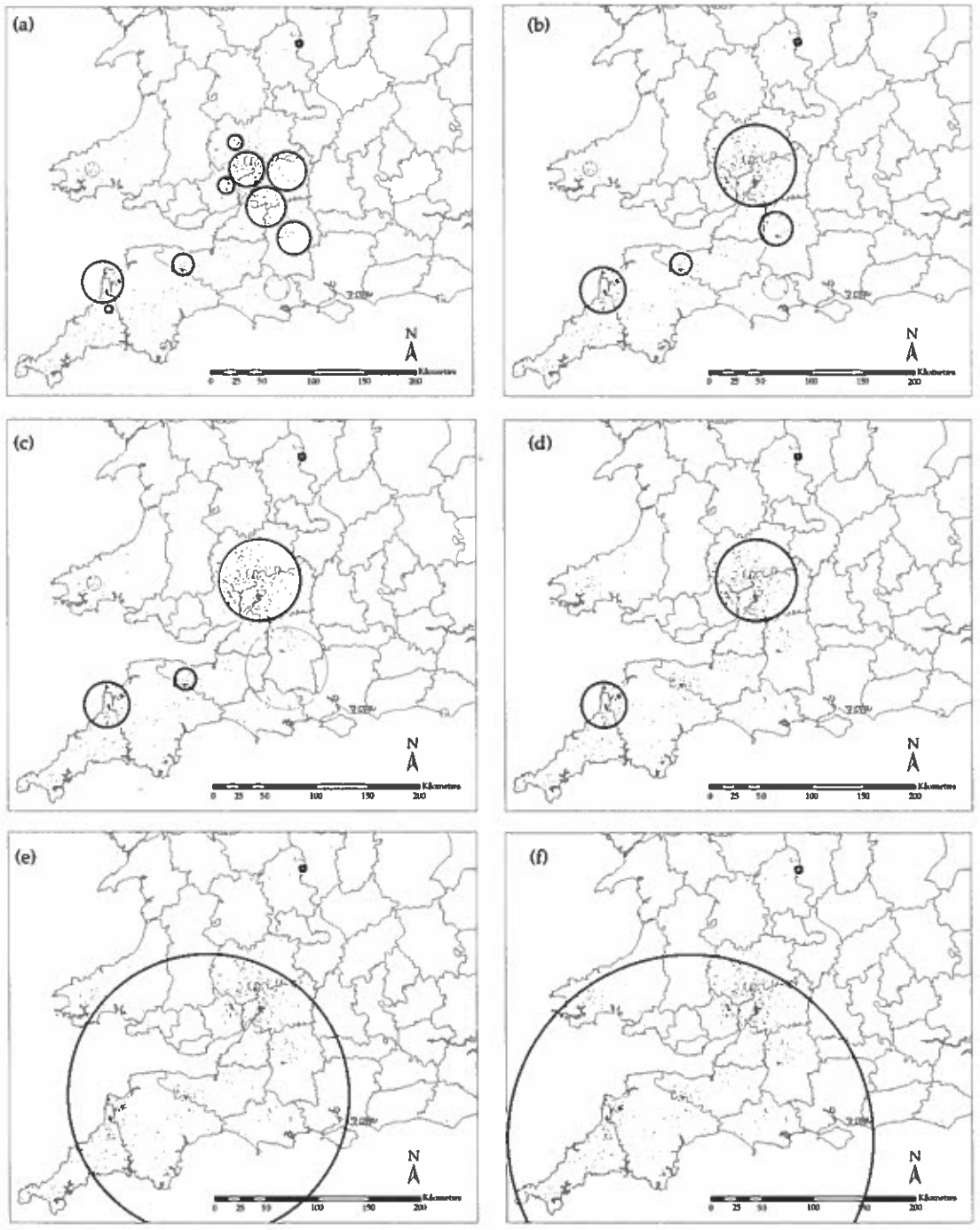


Figure 5.3 Cluster analysis comparing different maximum percentages of the population at risk to be included in a cluster, using all TB breakdown herds as cases and a 30% sample of control herds, in the southwest of Britain, in 1977. Maximum percentage of the population at risk to be included in clusters was a) 1%, b) 5%, c) 10%, d) 20%, e) 30% and f) 50%. Bold black circles indicate P-values of < 0.001; fine grey circles indicate P-values between 0.001 and 1.00.

Table 5.1 Details of the spatial clusters illustrated in Figures 5.3.b and 5.4. For each cluster the area is given, LLR is the log likelihood ratio, P (999) is the significance level based on 999 Monte Carlo replications (hence the plateau at 0.001), RR is the relative risk (observed/expected), which is directly proportional to Prev., (the prevalence corrected for sampling)

Cluster	Area (km²)	LLR	P (999)	RR	Prev.
1	4,879	280.5	0.001	8.3	0.048
2	1,516	115.8	0.001	9.7	0.056
3	32	33.7	0.001	35.7	0.206
4	821	32.6	0.001	7.3	0.042
5	337	32.4	0.001	11.2	0.065
6	37	18.4	0.002	20.8	0.120
7	149	14.4	0.010	8.2	0.048
8	52	13.3	0.030	41.6	0.240
9	479	12.5	0.052	7.8	0.048
10	82	9.7	0.427	15.3	0.088

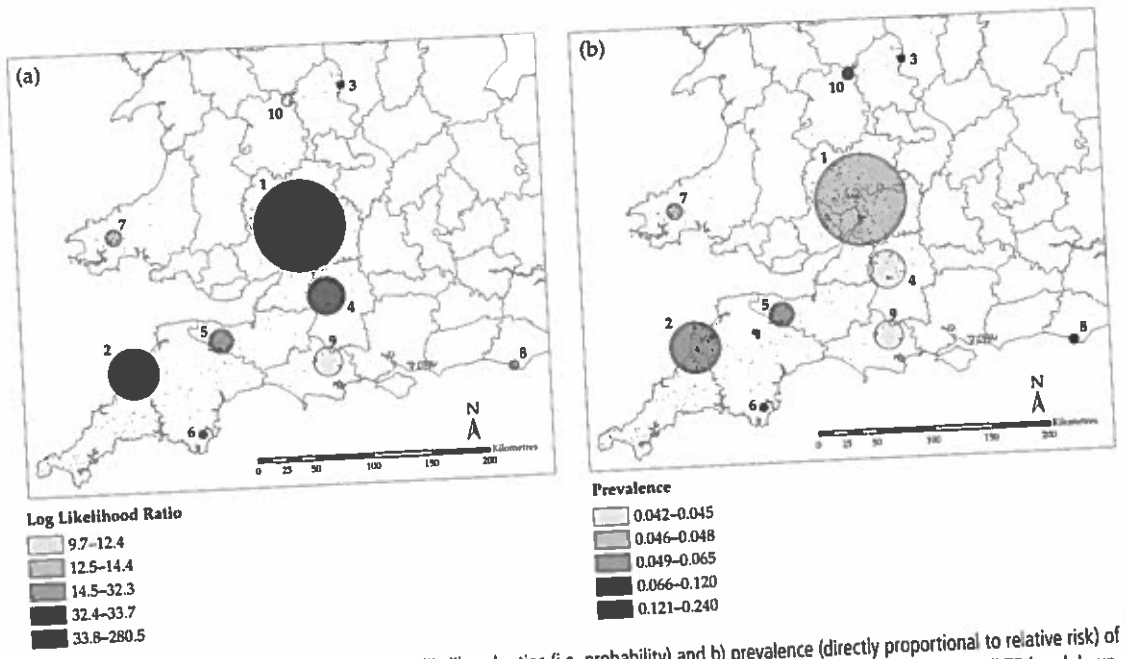


Figure 5.4 Detailed cluser analysis showing a) log likelihood ratios (i.e. probability) and b) prevalence (directly proportional to relative risk) of clusters of TB breakdown herds in the southwest of Britain, in 1977. Clusters were identified using the Bernoulli model using all TB breakdown herds as cases compared to 30% of control heads, with a maximum cluster size set to 5% of all points. See also Figure 5.3.b and Table 5.1.

example, cattle movements (Gilbert et al. 2005) into areas where prevailing environmental and epidemiological conditions allow new disease clusters to arise, as indicated for example in Wint et al. (2002). Clearly there are clusters of bovine TB in the southwest of Britain, some persistent, some fleeting, and some recurring.

5.4 Detecting clusters around a source (focused tests)

With focused tests the location of a cluster centre is specified *a priori*, and the likelihood of that location truly being a cluster centre is then determined. Localized clustering of disease around

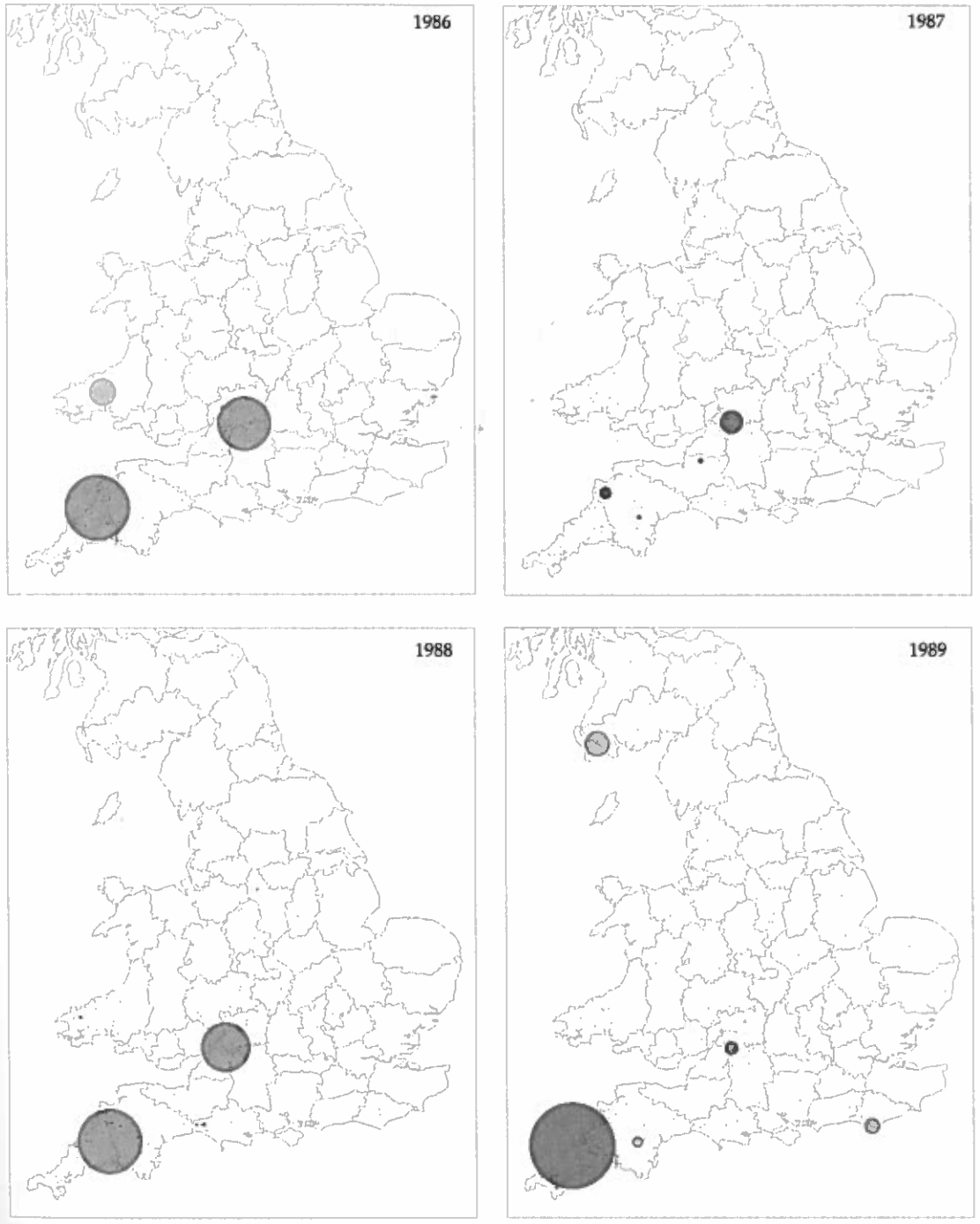
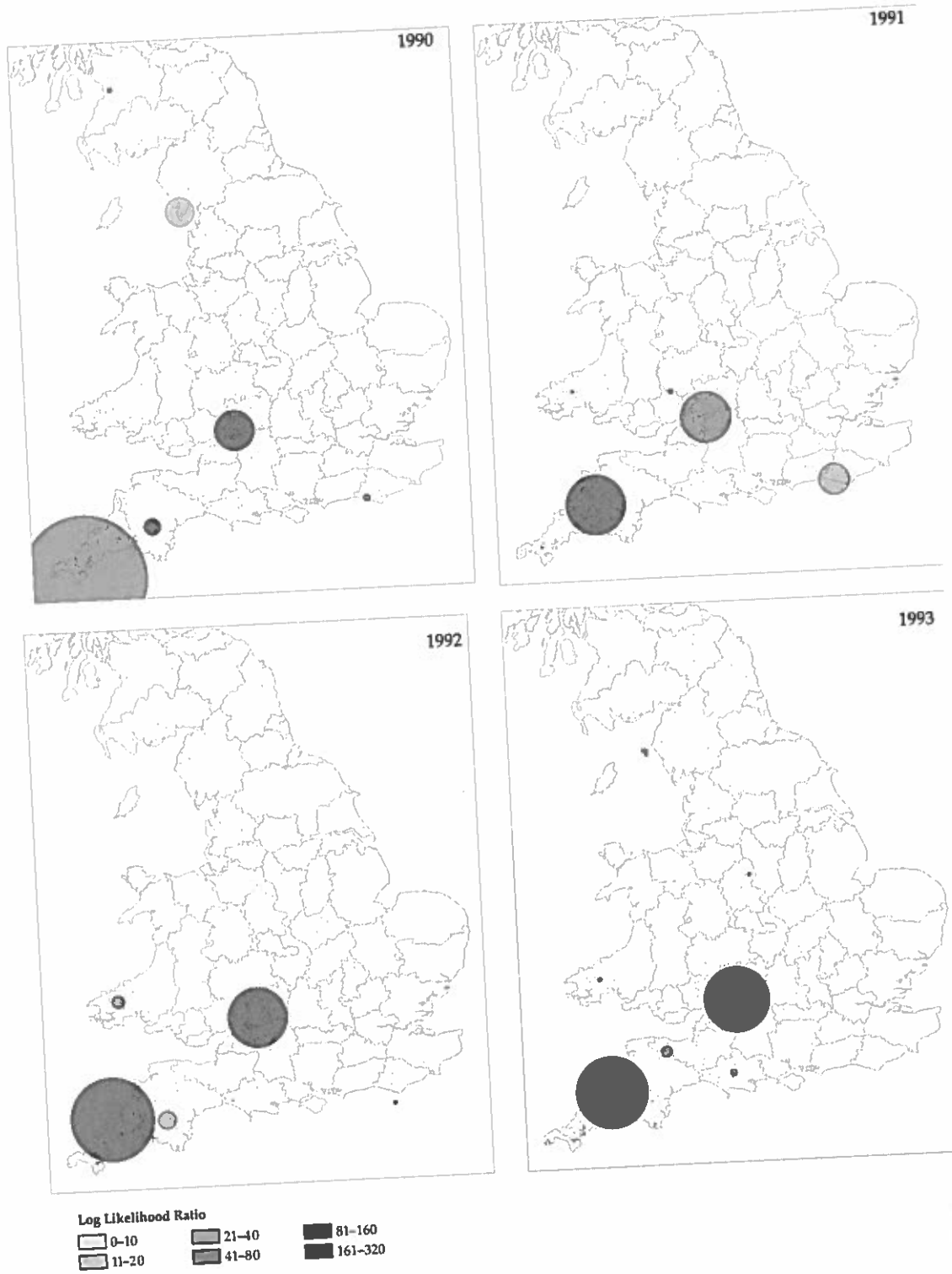
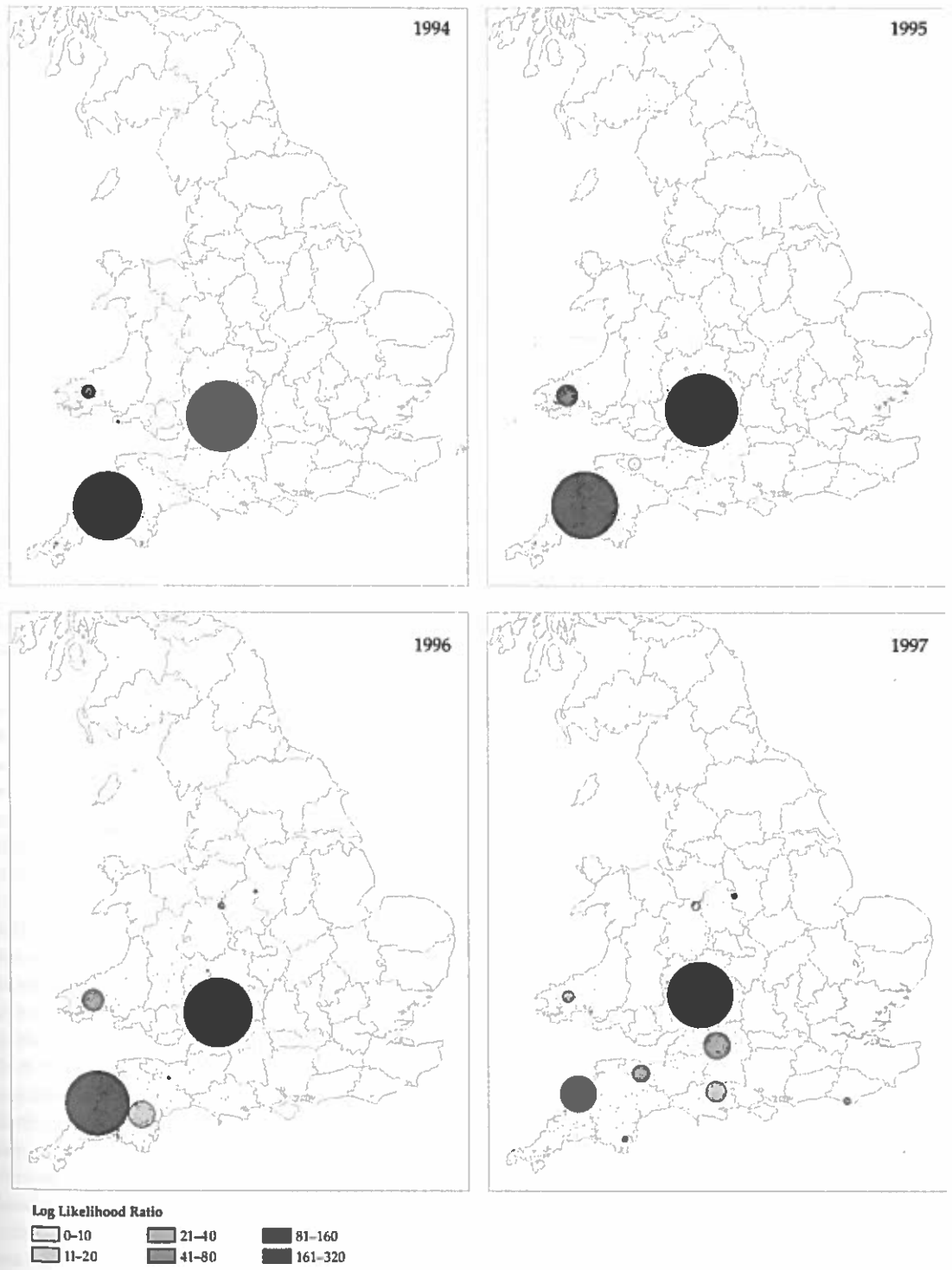


Figure 5.5 Cluster analysis of cattle TB data (1986–1997). Clusters were identified for each year using the spatial scan statistic and the Bernoulli model with all TB breakdown herds compared to 30% of control herds, with a maximum cluster size set to 5% of all points and no overlapping clusters permitted. Clusters are shaded by log likelihood ratio using the same scale for each year with darker clusters being the most likely. The distribution of breakdown herds in each year is superimposed on the cluster map.


1990
1991
1992
1993
Log Likelihood Ratio
0–10
11–20
21–40
41–80
81–160
161–320


1994
1995
1996
1997
Log Likelihood Ratio
0–10
11–20
21–40
41–80
81–160
161–320

environmental hazards is a sensitive issue, exemplified by the highly publicized investigations into elevated childhood leukaemia incidence around the Sellafield nuclear reprocessing plant in the 1980s. Suspicions were aroused by researchers for a television company in 1983 (Gardner 1989; 1992). The Yorkshire Television programme *'Windscale, the nuclear laundry'*, alleged that elevated levels of childhood leukaemia resulted from nuclear discharge from the plant into the Irish Sea, leading to a Government inquiry that resulted in the Black Report (Black 1984). This report confirmed a high incidence of cancer in the study area, which it maintained is 'unusual but not unique', and that no causal link between nuclear power and illness could be proved. Not surprisingly this sparked debate into the validity of the statistical methods applied and initiated a series of investigations into cancer clusters around nuclear and other 'high risk' sites (reviewed by Alexander and Boyle (2000)). Caldwell (1990) reports that investigations into 108 space–time clusters during 22 years of cancer cluster investigations at the Centres for Disease Control and Prevention in the United States, showed no consistent patterns or clues as to disease aetiology, and Alexander and Boyle (2000) state that 'whilst many statistically significant clusters have been identified, none has ever been explained by an environmental pollutant or infectious agent'.

Comprehensive reviews on cluster detection around putative point sources have been written, including those by Hills and Alexander (1989), Muirhead and Darby (1989), Marshall (1991a), Elliott et al. (1995), and Morris and Wakefield (2000). It is not the intention to go into great detail here, but rather to provide an overview. Morris and Wakefield (2000) provide a useful review of a variety of methods, applying them to a single example; the incidence of stomach cancer in relation to proximity to a municipal solid-waste incinerator in Great Britain (see also Elliott et al. (1996)). Besag and Newell (1991) warn against pre-selection bias arising from a point source being chosen for investigation based on a prior knowledge of a higher disease incidence in its vicinity. Hills and Alexander (1989) distinguish between situations where there is a prior hypothesis about a source and those where the hypothesis is reactive. They emphasize the problem of choosing the area around the source, and of accommodating extra-Poisson variability. Wartenberg and Greenberg (1993) highlight the importance of choosing appropriate techniques for given circumstances. In specifying the null hypothesis, usually that the disease events occur with a uniformly distributed probability of risk, they emphasize the importance of choosing a method that can account for any confounding variables such as age, gender, occupation, or ethnicity. In determining the alternative hypothesis they stress the importance of different clinical patterns of disease resulting from a single risk source as opposed to more general hotspots.

5.4.1 Stone's test

Stone (1988) developed a class of tests for trend, the maximum likelihood ratio (MLR) and Poisson maximum (Pmax) tests, both of which use the first isotonic regression estimator, working on the assumption that there will be a monotonic decay of risk with increasing distance from any point-source of a disease. Stone's Pmax test is applied to data on cancer incidence in the vicinities of the Siezewell and Sellafield nuclear installations (Bithell and Stone 1989). At Sellafield they demonstrate a highly significant trend in disease excess indicating a genuine association, but geographical proximity to the power station at Seascale does not appear to be important. Software for implementing Stone's test includes the Dcluster package in R.

Elliott et al. (1992b) use Stone's test to analyse the incidence of lung and laryngeal cancer around incinerators of waste solvents and oils at Charnock Richard in Lancashire (1972–1980), but find no evidence for a decreasing risk of cancer with increasing distance from the sites. In a later study Elliott et al. (1996) examine an extended cancer dataset for Great Britain in relation to solid-waste incinerators, finding no evidence for declining risk with distance from incinerators for cancer of the larynx or connective tissues (including soft-tissue sarcoma), nasal and nasopharyngeal cancer, and non-Hodgkin lymphomas. However, significant results are obtained for all cancers combined, and for stomach, colorectal, liver, and lung cancers, although these results are thought to be due to confounding effects such as deprivation.

Stone's test has been applied to studies of cancer incidence near radio and television transmitters. In a study specific to the Sutton Coldfield transmitter in Great Britain, Dolk et al. (1997b) find the risk of adult leukaemia between 1974 and 1986 to be elevated within 2 km of the transmitter, and to decline significantly with increasing distance from the transmitter. They extend this study to explore the incidence of cancers in the vicinity of 20 high-power transmitters across Great Britain (1974–1986) (Dolk et al. 1997a) and whilst they do find evidence for a declining leukaemia risk with increasing distance from transmitters, conclude that the pattern around the Sutton Coldfield transmitter is uncommon. Michelozzi et al. (2002) use Stone's test to investigate the incidence of adult and childhood leukaemia in the vicinity of the Vatican Radio high-power radio transmitter. The result is significant for children and male adults, with a declining risk as a function of distance from the transmitter, but no explanation is provided for these findings.

A number of adaptations of the original method have been developed. For example, Morton-Jones et al. (1999) propose an extension that allows for covariate adjustments via a log-linear model, which they illustrate using data on the incidence of stomach cancers near municipal incinerators. Diggle et al. (1999) adapted Stone's isotonic regression method to incorporate case-control data in addition to covariate information. They illustrate the adaptations using data on lung and laryngeal cancers in the vicinity of an industrial incinerator, and on childhood asthma in relation to distance from major roads. Consistent with Diggle (1990) they show moderate evidence of increased risk of cancer of the larynx in the vicinity of the disused industrial incinerator, compared with the more common incidence of lung cancer, in the Chorley-Ribble area of south Lancashire, England between 1974 and 1983, although covariate information is not available for this dataset. They also find no association between asthma in children in North Derbyshire in 1992 and distance from major roads, after adjusting for the covariates of age and sex (in agreement with Diggle and Rowlingson (1994)).

5.4.2 The Lawson–Waller score test

Lawson (1993) and Waller et al. (1992) developed the Lawson–Waller score test, sometimes referred to as the uniformly most powerful (UMP) test. The score test detects a decreasing trend in disease frequency associated with declining exposure to a point-focus. The test statistic T_{LW} is given by:

$$T_{LW} = \sum_{i=1}^{I} g_i(O_i - E_i) \tag{5.8}$$

$$Var(T_{LW}) = \sum_{i=1}^{I} g_i^2 E_i - O_i\left(\sum_{i=1}^{I} \frac{p_i g_i}{p_+}\right)^2 \tag{5.9}$$

where g_i denotes the exposure to the focus for an individual residing in area i, O_i is the observed number of disease cases in area i, E_i is the expected number of disease cases in area i, p_i is the size of the population at risk in area i, and p_+ is the total population size. Monte Carlo simulation is used to evaluate the significance of the test statistic T_{LW}.

Waller et al. (1992) demonstrate the methodology using leukaemia incidence data in upstate New York (1978–1982), with waste sites containing trichloroethylene as putative hazardous point-sources. Lawson (1993) demonstrates the methodology using bronchitis mortality around a chemical reprocessing plant in Bonnybridge, central Scotland (1980–1982).

Michelozzi et al. (2002) apply the Lawson–Waller score test to investigate the incidence of adult and childhood leukaemia in the vicinity of the Vatican Radio high-power radio transmitter. Waller et al. (1992), Waller and Turnbull (1993), Waller and Lawson (1995), and Waller (1996) compare the Lawson–Waller score test with Besag and Newell's test (Besag and Newell 1991) (adapted as a focused clustering test) and with Stone's test (Stone 1988), concluding that the sensitivity of the methods depends on the level of aggregation of the underlying surveillance data (Waller 1996). From a number of simulations to compare the power of each test they conclude that for very rare diseases all tests have low power, with Stone's test performing least well, followed by Besag and Newell's test, and that the best by a small margin was the Lawson–Waller score test. Increasing the disease prevalence improved the power of all the tests. In this instance

the power of Stone's test surpasses that of Besag and Newell's test. They conclude that their own test outperforms the other two, in terms of power, under all conditions.

5.4.3 Bithell's linear risk score tests

Bithell et al. (1994) and Bithell (1995) developed a set of tests known as 'linear risk score tests', where disease incidence is weighted by some distance function from a point source (e.g. $1/d_i$, $1/d_i^2$, or $1/rank_i$). Using the reciprocal of distance is appropriate for detecting an environmental hazard that declines with distance from a source and is relatively insensitive to the precise location of the assumed source. Using the reciprocal of rank is more appropriate when the relative proximity of residence is important, rather than actual distance, but it is more sensitive to the precise location of the putative source. Bithell et al. (1994) compare the power of these tests against Stone's Pmax and MLR tests concluding that they are likely to be more powerful than Stone's tests, particularly in situations where non-uniformity of risk is only slight. They apply these tests to the distribution of childhood leukaemia around 23 nuclear installations in England and Wales (Bithell 1995) with only two, Sellafield and Burghfield, giving significant results. They conclude that there is no evidence for a general increase in childhood leukaemia in the vicinity of nuclear installations, and that apart from Sellafield, the evidence for distance-related risk is weak. Bithell's linear risk score tests can be implemented using the ClusterSeer software.

5.4.4 Diggle's test

Diggle (1990) developed a test that uses nonparametric kernel smoothing to describe natural variation in a disease (assuming a Poisson point process), and then a maximum likelihood test to evaluate the possibility of raised incidence around a pre-specified point source. Diggle (1990) demonstrates the approach, exploring links between a disused industrial incinerator and a possibly elevated incidence of laryngeal cancers, compared with the more common incidence of lung cancer, in the Chorley-Ribble area of south Lancashire. Kernel smoothing is used to describe the more common lung cancer incidence, providing an estimate of 'natural' spatial variation. A parametric maximum likelihood approach is then used to describe raised incidence in laryngeal cancer near the pre-specified source. This clustering is confirmed using Monte Carlo simulations. Diggle and Rowlingson (1994) developed a conditional approach, which converts the original point process model into a non-linear binary regression model for the spatial variation in risk. Using the same cancer datasets from south Lancashire they show similar results but conclude that these should be more reliable than those obtained from the original point process model. Moreover, this adaptation allows adjustment for covariates in a log-linear fashion. They further demonstrate the method using asthma data from North Derbyshire in 1992. Three putative point sources; a coking works, a chemical plant, and a waste treatment centre, are evaluated after adjusting for a number of covariates including distance from roads, presence of a cigarette-smoker in the household, dust problems, and hay fever. They find some evidence for association with the coking works and a significant association with the presence of hay fever. Software for implementing Diggle's test includes the splancs package in R, and ClusterSeer.

5.4.5 Kulldorff's focused spatial scan statistic

The spatial scan statistic can be used to search for clusters around a point source by making the point source the only coordinate pair in the special grid file. In this way Viel et al. (2000) demonstrate clustering of cases of soft-tissue sarcomas and non-Hodgkin's lymphomas, but not of Hodgkin's disease, around a solid-waste incinerator with high emission levels of dioxin in France. The authors conclude that further studies would be needed to confirm the relationship between dioxin exposure and soft-tissue sarcoma and non-Hodgkin's lymphoma risk. In two recent applications the focused spatial scan statistic has been used to link BSE cases to specific feed mills. Using a number of models, Sheridan et al. (2005) confirm spatio-temporal clustering of cattle herds in Ireland (1996–2000). The spatial component is dominant and a focused test identifies high-risk feed mills at the centroids of clusters. Schwermer et al. (2007), investigating BSE cases in Switzerland (1996–2001), use as cluster

centres the locations of feed producers that tested positive for contamination of cattle feed with meat and bone meal. They conclude that whilst a causal link is suggested by their results, other factors must also play a part since BSE clusters occurred around only some producers of contaminated feed.

A particularly innovative application of the focused spatial scan statistic is its use to establish the origins of a *Trypanosoma brucei rhodesiense* sleeping sickness outbreak in the Sorroti area of eastern Uganda. In an 18-month observational study, following an outbreak of sleeping sickness in Sorroti, Fevre et al. (2001) find that over half the cattle traded at the Brookes Corner cattle market originated from endemic sleeping sickness areas. A subsequent case-control study identifies distance to the cattle market as a highly significant risk factor for sleeping sickness, and by using the focused spatial scan statistic they demonstrate significant clustering of cases close to the market at the start of the outbreak. Software for implementing the focused spatial scan statistic includes SaTScan, the Dcluster package in R, and ClusterSeer.

5.5 Space–time cluster detection

5.5.1 Kulldorff's space–time scan statistic

The spatial scan statistic has been adapted to look for clusters in space and time by extending the idea of a two-dimensional circular window to that of a cylinder passing through time. This has been applied, using the Poisson model, to explore brain cancer distribution in Los Alamos, a remote New Mexican community established in 1943 to provide a workforce to the Los Alamos National Laboratory, a nuclear research and design facility (Kulldorff et al. 1998a). Kulldorff (2001) proposes the prospective use of the space–time scan statistic, with repeated time periodic analysis, as part of a surveillance system to track active clusters of disease, for detecting the geographic location of emerging clusters and evaluating their significance. He demonstrates this with the Los Alamos data. In a later development of the original 'prospective space–time scan statistic' (Kulldorff 2001), Kulldorff et al. (2005) develop the 'space–time permutation scan statistic', which does not require data on the background population at risk, but estimates expected disease occurrence based only on case data. They demonstrate this adaptation using daily diarrhoea surveillance data from hospital emergency departments in New York City between November 2001 and November 2002.

Hjalmars et al. (1999) use the space–time scan statistic to investigate clustering of childhood malignant brain tumours in Sweden, demonstrating an increase in rates of these cancers during the period 1973 to 1992, but find no evidence for clustering in space or time. Viel et al. (2000), using a Poisson model, explore space–time clustering of soft-tissue sarcomas, non-Hodgkin's lymphomas, and Hodgkin's disease, around a solid-waste incinerator in France. Smith et al. (2000) use a Bernoulli model to explore space–time clustering of anthrax strains in the Kruger National Park in South Africa. The spatial and temporal distributions of two genotypes indicate the anthrax epidemic foci to be independent of each other. Sheehan et al. (2000) find significant clustering of breast cancer in western Massachusetts, USA. Ward (2001) uses the space–time scan statistic (Poisson model) to investigate clustering of sheep blowfly strike, controlling for the effect of flock size and composition, the presence of fly strike control measures, and rainfall. Other, more recent examples of the space–time scan statistic include explorations of clustering in viral diseases of farmed fish (Knuesel et al. 2003), leptospirosis in dogs (Ward 2002), and different molecular subtypes of human listeriosis (Sauders et al. 2003).

In an investigation of outbreaks of acute respiratory disease in Norwegian cattle herds, Norström et al. (2000) compare the space–time scan statistic to both Knox's method and Jacquez's k nearest neighbour test. They find that all the methods identify significant space–time clusters, but highlight the disadvantage of the subjective selection of space and time thresholds in the Knox test, and the disadvantage, common to both the Knox test and to the Jacquez k nearest neighbours test, that they assume the population size does not change through time. The prospective space–time scan statistic does not make this assumption, it can accommodate confounding covariates in the analysis and moreover, it has the advantage in that it identifies the actual locations of space–time clusters.

Kleinman et al. (2005) investigate various adjustments to the space–time scan statistic to account for underlying spatial and temporal trends in diseases. They find that when adjusting for day of the week, month, holidays, and local history of disease, smaller numbers of clusters of lower respiratory complaints are found (2% of days) compared to when unadjusted census population data are used (26% of days). Other recent examples of the application of space–time scan statistics include Recuenco et al. (2007), who identify a number of significant space–time clusters of enzootic racoon rabies in New York State (1997–2003), Sheehan and De Chello (2005) who demonstrate clusters of high and low incidence of late-stage breast cancer in Massachusetts (1988–1997), and Jones et al. (2006) who use the prospective space–time scan statistic to reveal clustering of shigellosis in Chicago, in 2002. Software for implementing the space-time scan statistic includes SaTScan, the Dcluster package in R, and ClusterSeer.

5.5.2 Example of space–time cluster detection

The space–time permutation of Kulldorff's spatial scan statistic was applied to the British cattle TB data for the period 1986 to 1997. The Bernoulli model was used and all TB-breakdown herds were included as cases compared to 30% of control herds, with the maximum cluster size set to 5% of all points (mainly to facilitate comparison with other analyses performed using these settings). Table 5.2 provides details of the 12 space–time clusters of TB breakdowns that were located, and Fig. 5.6 illustrates some details of these clusters: (a) log likelihood ratio, (b) prevalence, and (c) duration.

The results of the analysis are dominated by two large and highly significant clusters; Cluster 1 (labelled in Fig. 5.6), with its centre in Gloucestershire, and Cluster 2 spanning the Devon/Cornwall boundary. Both were relatively long-lasting but Cluster 2 had a much lower prevalence over this time period illustrating the need to consider all characteristics of clusters when making comparisons.

The results from cluster analysis of individual years (Fig. 5.5) can thus be compared with a single analysis over the entire period (Fig. 5.6). The latter presents the results of the analysis in a more condensed form, but at the risk of missing some of the subtleties that can be seen by looking at individual years.

5.6 Conclusion

This chapter has explored a variety of methods by which spatial and spatio-temporal disease clusters

Table 5.2 Details of space–time clusters illustrated in Figure 5.6. For each cluster the area is given, the start and end date, LLR is the log likelihood ratio, P(999) is the significance level based on 999 Monte Carlo replications (hence the plateau at 0.001), RR is the relative risk (observed/expected), which is directly proportional to Prev., (the prevalence corrected for sampling) over the duration of the cluster

Cluster	Area (km²)	Start	End	LLR	P(999)	RR	Prev.
1	5,830	1992	1997	1,150,4	0.001	9.5	0.042
2	5,148	1989	1997	743.6	0.001	5.9	0.002
3	316	1991	1997	212.1	0.001	13.1	0.006
4	209	1993	1997	78.2	0.001	15.1	0.009
5	194	1997	1997	41.3	0.001	20.5	0.063
6	194	1989	1997	41.1	0.001	33.8	0.012
7	21	1987	1996	24.3	0.001	43.1	0.013
8	27	1987	1994	21.8	0.001	11.2	0.004
9	47	1995	1997	20.6	0.003	39.9	0.041
10	479	1997	1997	18.1	0.017	14.5	0.045
11	15	1990	1992	14.3	0.382	37.3	0.038
12	13	1987	1996	13.7	0.721	97.0	0.030

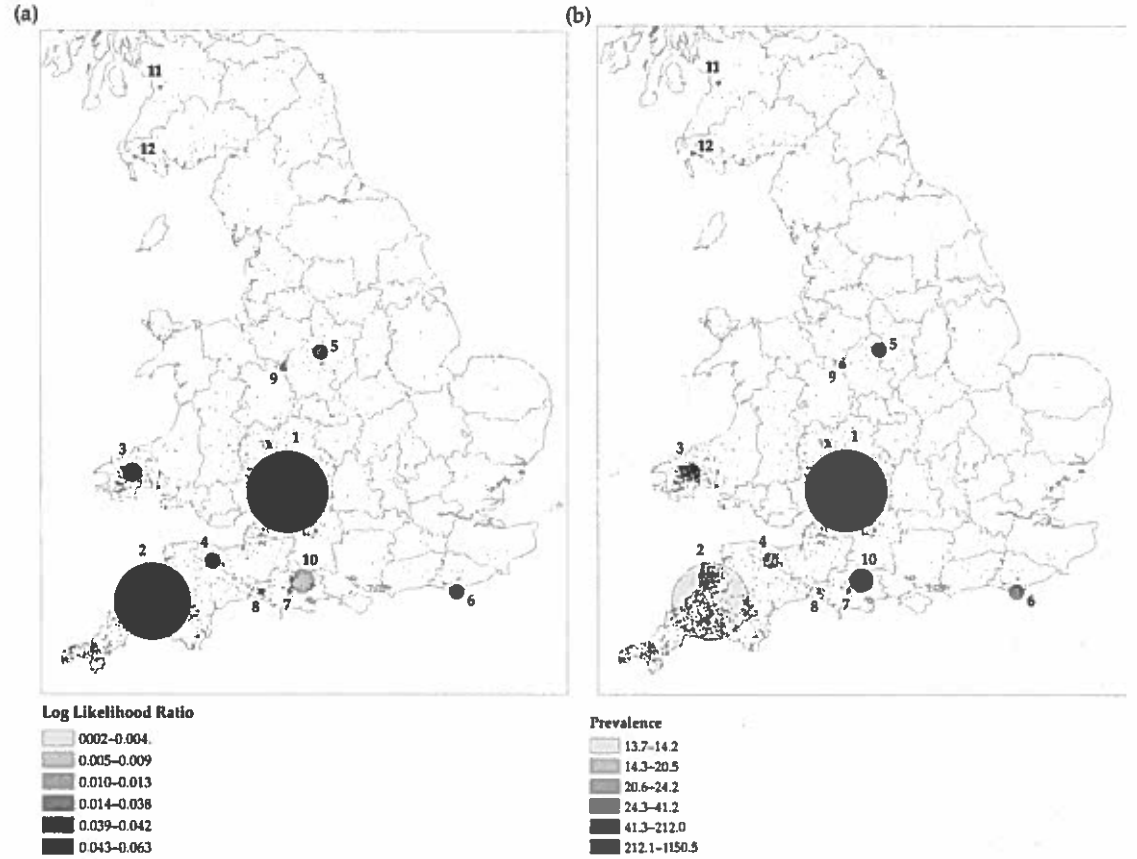


Figure 5.6 Space–time cluster analysis showing a) log likelihood ratios (i.e. probability), b) prevalence (directly proportional to relative risk), and c) duration of space–time clusters of TB breakdown herds in Britain from 1986 to 1977. Clusters were identified using the Bernoulli model of the space-time scan statistic with all TB breakdown herds as cases compared to 30% of control herds, with a maximum cluster size set to 5% of all points. See also Table 5.2.

can be identified statistically. New methods for detecting local clustering are continuously being developed. For example, Aamodt et al. (2006) compare the spatial scan statistic against a method based on generalized additive models (Hastie and Tibshirani 1991) and one using Bayesian approaches that have emerged from the Besag, York, and Mollié model (Besag et al. 1991). Other Bayesian approaches to cluster detection include those demonstrated by Neill et al. (2006) and Lawson (2006b), but these tend towards the modelling approaches that are described in detail in Chapters 6 and 7. Another recent advance is the development of a spatial hazard model for cluster detection (Gay et al. 2007), which can account for risk factors and for spatial heterogeneity in the population at risk. In terms of statistical methodology, the field of cluster detection is advancing rapidly.

A frequently quoted limitation common to many cluster detection tests (e.g. Besag and Newell (1991) and Cuzick and Edwards (1990)) is the *a priori* choice of cluster size, as testing for a variety of cluster sizes results in problems of multiple inference. Analysis of the British cattle TB data suggests this to be a central problem and there do not seem to be clear guidelines on how to deal with it.

Kulldorff et al. (1997) suggest, and demonstrate using an analysis of breast cancer in the northeast United States, a method of decomposing a significant cluster into non-overlapping sub-clusters, each of which would allow rejection of the null hypothesis on its own strength by sequentially limiting the maximum cluster size (the very approach adopted in Section 5.3.6). However, in the same paper he states that one of the reasons for the superiority of the spatial scan statistic over other cluster detection algorithms is that 'by searching for clusters without specifying their size or location, the method ameliorates the problem of pre-selection bias'.

Chaput et al. (2002), in a spatial analysis of HGE near Lyme, Connecticut, vary the maximum cluster size from 50 to 25% of the population at risk, in order to look for 'sub-clusters'. A large and highly significant cluster ($p = 0.001$) is identified at the 50% level, and with the 25% threshold two 'sub-clusters' emerged, one highly significant ($p = 0.001$) and the other considerably less so ($p = 0.16$). They argue that by taking the default value of 50%, pre-selection bias is avoided and therefore the first analysis better represents the areas of increased risk for infection based on the spatial distribution of HGE in the area.

Jemal et al. (2002) report on the geographic analysis of prostate cancer mortality between 1970 and 1989 in the United States. They identify fairly large, significant clusters in a 'first round of analyses' and smaller sub-clusters in a 'second round of analyses'. Based on the very large size of a primary cluster of prostate cancer among white males in the northwest of the United States it must be assumed that they select 50% of the population as the upper limit. What is not clear is whether the sub-clusters are located by considering the clusters identified in the first round as independent populations (for new analyses), or by reducing the maximum cluster size (proportion of the population included) by some unspecified amount, in order to identify a larger number of smaller clusters (as was done in Section 5.3.6).

Boscoe et al. (2003) point out that those following Kulldorff's approach tend to select clusters with large areas containing large populations but only small elevations in disease risk, since these exhibit the highest levels of statistical power. Smaller areas, with lower levels of significance but higher prevalence or incidence rates, tend to be overlooked. They propose an adaptation to Kulldorff's approach that involves stratifying the set of statistically significant circles by relative risk. Within each risk stratum the non-overlapping circles with the highest likelihood are displayed in a nested manner. They demonstrate this method using prostate cancer mortality data from the USA.

As shown in Section 5.3.6 (specifically in Fig. 5.3), *a priori* choice of cluster size can have profound effects on the results. The grave concern arises that by exploring a range of maximum cluster sizes an upper cluster size threshold can be chosen that presents a pattern of clustering best suited to support a particular argument, rather than that which best reflects reality. This may cast doubt on the validity of the numerous studies that have been reported using scan circles, and in particular those based on the spatial scan statistic.

CHAPTER 6

Spatial variation in risk

6.1 Introduction

Whenever possible, epidemiological disease investigations should include an assessment of the spatial variation of disease risk, as this may provide important clues leading to causal explanations. The information presented may be, for example, the density of disease cases or spatial variation in an attribute value such as disease risk. The objective is to produce a map representation of the important spatial effects present in the data while simultaneously removing any distracting noise or extreme values. The resulting smoothed map should have increased precision without introducing significant bias (Haining 2003). The method used to analyse the data depends on how they have been recorded. If the data occur as point locations (e.g. outbreaks of disease) kernel smoothing methods can be applied to facilitate visual assessment of the pattern. In the case of data representing, for example, the incidence of infection within administrative areas, Bayesian methods can be applied to take account of the uncertainty of the local measurement and spatial dependence between neighbouring measurements. If the data represent sample point locations used to describe continuous fields, such as disease vector presence, interpolation methods such as kriging can be applied to generate spatially continuous representations.

6.2 Smoothing based on kernel functions

Visual analysis of point data density ranges from a simple display of locations to the use of smoothing methods for generating point density surface representations in raster format. In general, the first method should only be used if the number of points is limited so that different densities can still be differentiated visually. The map in Fig. 6.1a, showing the point locations of TB-infected herds, can be readily interpreted. However, Fig. 6.1b shows the point locations of all cattle herds TB-tested in Great Britain in 1996, and as a result of the large number of points it is not possible to differentiate between areas with high densities of points. In such situations, it is necessary either to generate estimates aggregated at some administrative level or to apply smoothing methods.

Spatial smoothing can be achieved by calculating simple localized averages or by applying more complex mathematical functions to the data. With both approaches a window, typically of fixed size, centres on each data point or polygon centroid. The degree of smoothing is influenced by the size and shape of the window (also called filter or kernel), as well as the mathematical function applied to the values within the window. The larger the window, the more information is 'borrowed' from the neighbouring areas, and the more statistically precise is the resulting smoothed estimate. The disadvantage is that in a spatially heterogeneous environment the estimates could be biased as they will be more strongly influenced by data values from locations some distance away (Burrough and McDonnell 1998; Haining 2003). As a consequence, important local clustering of increased or reduced point density may disappear in the kernel-smoothed representation. The results produced by spatial smoothing methods can also be biased by edge effects.

Spatial filters are applied in image enhancement to remove random noise, but are also available as a