26

ANALYSIS OF SPATIAL DATA

Chapter contributed by Dirk Pfeiffer

OBJECTIVES

After reading this chapter, you will be able to:

1. Describe the specific characteristics of spatial data.
2. Generate maps expressing spatial variation in disease risk from point and area data.
3. Investigate spatial dependence and clustering for different data types, including space–time association.
4. Conduct regression analyses for describing spatial patterns and identifying risk factors in data that may be spatially dependent.

26.1 INTRODUCTION

Most disease events, risk factors, and other relevant attributes can be characterised by their geographical position. Epidemiological studies may then be aimed at investigating the spatial relationships, for example by assessing whether disease events express some level of spatial similarity between each other or whether disease occurrence is associated with potential risk factors that are in spatial proximity to the disease event. However, if there is spatial dependence in the attribute values of an epidemiological dataset, this will violate the assumption of independence between observations in a statistical analysis. In such instances, appropriate statistical algorithms will have to be used even if the analysis is not aimed at investigating spatial relationships.

Spatial analysis includes visualisation, exploration and modelling of spatial data. Chapter 25 describes visualisation, and the current chapter focuses on methods which involve statistical inference; namely, descriptive risk mapping, exploratory analysis and modelling of spatial data.

The objective of this chapter is to provide an introduction to the methods that can be applied in spatial analyses of epidemiological data. More detailed information is provided in various specialist textbooks such as: Bailey & Gatrell, (1995); Bivand *et al*, (2008); Haining, (2003); Lawson (2009); Pfeiffer *et al*, (2008); Waller & Gotway, (2004).

26.2 ISSUES SPECIFIC TO STATISTICAL ANALYSIS OF SPATIAL DATA

26.2.1 Spatial effects

A key characteristic of geo-referenced data is that it may be subject to 2 particular effects, **spatial dependence** and **spatial heterogeneity**. Spatial dependence means that observations closer to each other are more likely to be similar. For example, for diseases transmitted by direct or indirect means (such as avian influenza or foot and mouth disease), incidence levels on neighbouring farms are likely to be more similar compared with those further away. It is therefore a local process. On the other hand, spatial heterogeneity refers to patterns of variation across a larger geographical area, and a process is spatially heterogeneous if there is variation in mean values, and their variability, across the study area.

In statistical terms, 2 types of spatial relationships can be defined: first and second-order spatial effects. **First-order spatial effects** (global or large-scale trends) relate to variation in the mean value of a spatial process (*eg* the average density of the poultry population in northern Vietnam increases closer to the major population centre (Hanoi), resulting in spatial heterogeneity). **Second-order (local or small-scale) spatial effects** refer to the local dependence in data values (also called **spatial autocorrelation**), including effects such as overdispersion (see Section 16.12.4 for a general discussion of overdispersion). Second order effects may arise as a result of clustering of cases of an infectious disease or local environmental conditions such as levels of nitrate in soil as a consequence of level of fertilizer application. Chapter 20-24 describe how these variance effects can be included in regression models as random effects. It should be noted that many spatial variables will be subject to both effects. If positive spatial autocorrelation is ignored in statistical analysis, the strength of the evidence in the data for rejecting the null hypothesis will generally be exaggerated. This means that the true statistical power of a given number of correlated observations is lower than if they were from

uncorrelated ones. It is therefore often stated that autocorrelation results in a 'loss of information'. If both, first- and second-order spatial effects are affecting a variable, unbiased estimation of either effect requires removal of, or controlling for, the other. In this context, when modelling a second-order spatial effect, it is assumed to be **stationary** (or homogenous - *ie* there is no first order effect or spatial heterogeneity), which means that the statistical properties of the variable of interest such as its mean and variance do not vary between locations. This also implies that the absolute geographic locations are not important to describe the characteristics of the spatial variable. As an example, the spatial distribution of the density of an insect vector species over a study area could be estimated using a number of traps distributed across the area. In this example, second-order stationarity makes the assumption that the mean vector density is constant across the study area (there should be no first-order effect or spatial trend), and that the covariance between insect species counts from any pair of random sampling locations only depends on the distance between sampling locations or relative location, but not on absolute location. In this case, using the points in Fig. 26.1, the covariance between sampling points P_1 and P_2 should be similar to the one between sampling points P_3 and P_4, but not necessarily similar to the one between P_5 and P_6. A second-order effect is isotropic if the covariance between pairs of sampling point measurements only depends on distance but not on direction. This means that in Fig. 26.1, the covariance between the measurements taken at sampling points P_1 and P_2 is the same as between P_5 and P_6, as well as between P_3 and P_4. As an example, wind-borne spread of the foot and mouth disease virus is unlikely to be **isotropic** since it depends on the wind direction.

Edge effects result from observations near the edge of the study area having fewer neighbouring measurements than those deep within the study area. An irregular shape of the study area will further compound this effect. Edge effects are relevant if there is spatial autocorrelation in the data, since observations near the edge will be correlated with unmeasured observations outside the study area, and therefore any effect estimates generated using only the available data are likely to be biased. Even if point data are uncorrelated, but smooth surface representations of point densities need to be produced, density estimates for locations near the edge will be derived using fewer points than there really are (censored observations), and these therefore may have higher variability than in the centre of the study area.

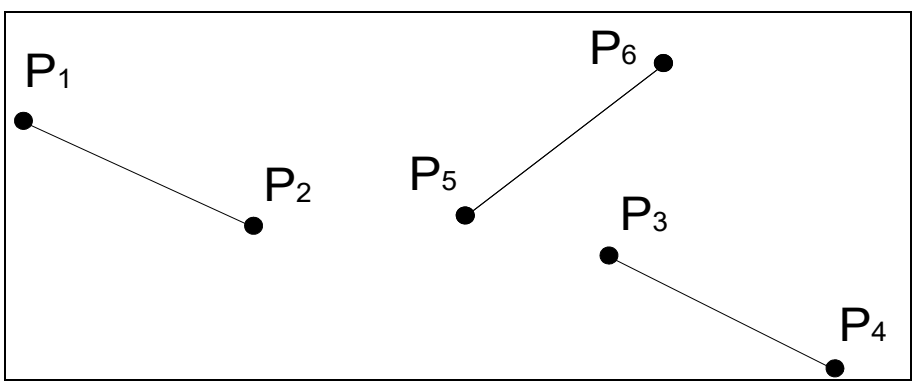


Fig. 26.1 Examples of point locations where those being equidistant are linked by lines

The modifiable areal unit problem (MAUP) is a manifestation of the ecological fallacy that can affect the results from spatial analyses (Waller & Gotway, 2004). It refers to conclusions from an analysis potentially being different, if the data are aggregated using different methods of aggregation. Two issues are usually considered: the **level of aggregation** (scale effect) and the selection of **boundaries** to be used in the aggregation process (zoning). As an example of the effect of level of aggregation, one may receive different analysis results if, in the multivariable analysis, the incidence of tuberculosis among cattle herds is aggregated at province rather than county level. An example of the zoning effect has been presented by Monmonier and de Blij (1996) using John Snow's data from the 1854 cholera epidemic in London. They showed that by aggregating the original point data using different geographic sub-divisions while keeping to the same scale (same size of areas), the cluster around the Broad Street pump could be made to visually appear or disappear. This is an example of a **zoning effect** and it is different from an aggregation effect which usually has a hierarchical structure. Both will influence any statistical inference. Waller and Gotway (2004) recommend that data should be collected at the 'resolution' at which it is to be analysed, so that the MAUP can be avoided. As a general principle, it is important to remember that any relationships detected in an analysis relate to the particular units of aggregation used for the analysis, but not necessarily for other units of aggregation even in the same study area.

26.2.2 Description of spatial arrangement

If spatial effects are to be taken into account in statistical analyses, the spatial arrangement of observations can be expressed using a **spatial weights** or **proximity matrix**. The interpretation of such a matrix depends on whether it describes spatially continuous fields or discrete spatial objects. Continuous fields can be defined by measurements taken at point sample locations or across a regular grid of pixels. In the case of point data, the matrix can represent the distances between all pairs of points. It can be simplified by representing actual interpoint distance as distance bands, and direction could be specified by defining segments within distance bands representing particular direction groupings. If data consist of pixels, raster or grid cells, the criterion can be whether or not pairs of pixels have common boundaries (first lag or **first-order neighbours**). This can be extended using the number of pixel boundaries that need to be crossed. For example, if 2 pixel boundaries have to be crossed, this would reflect a lag 2 or second-order neighbour. The arrangement of discrete spatial objects such as farm or administrative polygon areas is more complex to represent since closeness could mean that these share a common boundary of a given length or that object centroids are within a certain distance. A binary **spatial connectivity matrix** is the simplest neighbourhood representation and can also be used for area-type data. In the latter case, a symmetric weights matrix is used storing values of 1 where pairs of areas share a boundary, and 0 if they do not. An example of such a matrix is presented in Fig. 26.2 where the neighbourhood relationships between Australian states are represented as a binary connectivity matrix. A more detailed discussion of this topic is available in (Haining, 2003). Spatial weights matrices can be generated by various software packages such as GeoDA (http://geodacenter.asu.edu/software/downloads), R (http://www.r-project.org) and WinBUGS/GeoBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/).

26.3 EXPLORATORY SPATIAL ANALYSIS

Exploratory spatial analysis is aimed at improving our understanding of the data, typically involving use of statistical methods to identify departure from complete spatial randomness. It

	WA	NT	QLD	SA	NSW	VIC	ACT	TAS
WA	0	1	0	1	0	0	0	0
NT	1	0	1	1	0	0	0	0
QLD	0	1	0	1	1	0	0	0
SA	1	1	1	0	1	1	0	0
NSW	0	0	1	1	0	1	1	0
VIC	0	0	0	1	1	0	0	1
ACT	0	0	0	0	1	0	0	0
TAS	0	0	0	0	0	1	0	0

a) Binary spatial connectivity matrix

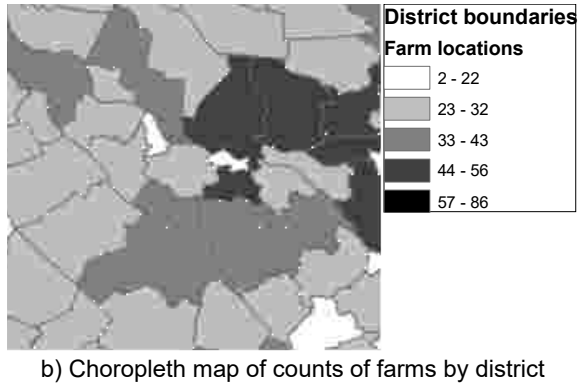
b) Map of states/territories in Australia

Fig. 26.2 Example of binary connectivity matrix

is applied to disease outcome data as well as potential risk factors. Exploratory analysis can be used to develop hypotheses in relation to causal relationships which are then tested formally using modelling approaches. Together with visualisation, exploratory analysis has the potential to become a standard component of disease surveillance systems. The types of data to be examined using exploratory analysis can be grouped into point and area data, and continuous spatial fields. The point patterns are either represented as actual point locations or aggregated at some administrative area level (see Fig. 26.3). The locations of the points are assumed to be generated by a process that has a random component. With continuous fields, the locations of the measurements are fixed and known in space. They represent sampling locations of an underlying spatially continuous process (continuous field) which means that, between the locations of any 2 sampling points, an infinite number of potential sampling locations exists at which values could be observed (see Fig. 26.4). The continuity is associated with the spatial aspect of such as a process, whereas the measurement values can be continuous or discrete (Schabenberger & Gotway, 2005).

Risk or probability maps can be produced for descriptive purposes to show spatial variation in disease occurrence, or explanatory variables can be considered to produce predictive risk maps

a) Point locations of farms



b) Choropleth map of counts of farms by district

Fig. 26.3 Farm locations represented as maps of point locations and aggregated by area

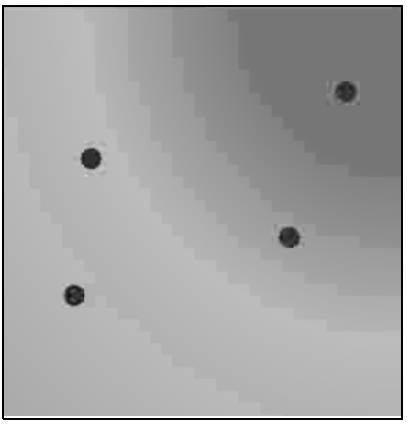


Fig. 26.4 Sampling points together with the underlying continuous field

(the latter will be further discussed in this chapter).

26.3.1 Descriptive risk mapping of point data

Disease data recorded as point locations can be shown as such, if the number of points is limited and locations of individual points can be visually differentiated. If there are a large number of points, they can either be aggregated as area count data as discussed in the next section or converted into a point density surface using smoothing methods. They can then also be presented as a risk by relating them to a denominator.

The most commonly used **smoothing method** for point data is **non-parametric kernel density** estimation. It applies a bivariate probability density function to the point data to produce estimates of the intensity of a spatial point process (Bailey & Gatrell, 1995). The associated mathematical equation is as follows:

$$\lambda_\tau(s) = \frac{1}{\delta_\tau(s)} \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right) \qquad \textit{Eq 26.1}$$

where $k()$ represents the chosen bivariate probability density function, $\tau > 0$ is the bandwidth, s is the point on which a disc of radius τ is centred, s_i are the points within the disc's area and δ represents an edge correction factor.

The resulting values can be presented as a raster map with one density value for each grid cell. The most influential parameters in this calculation are the bandwidth τ and the size of the grid cells, whereas the choice of density function is less critical. The principle behind the process is that for each grid cell, the point density is calculated based on a distance-weighted average with the weights being determined by the **bandwidth** (width of kernel function) and the shape of the probability function. The larger the bandwidth, the smoother the resulting surface will be, and vice versa (see Example 26.1). The selection of bandwidth τ can be based on mathematical calculations or by subjective choice. It is important to apply an appropriate balance between the perceived objectivity of mathematical approaches and the option of subjective choice based on bandwidths that are able to reveal relevant underlying biological processes. It is therefore usually recommended to explore the effect of different bandwidths on the resulting smoothed

surface. Diggle (2000) recommends that a plot of the mean square error of the intensity estimator against different values of τ should be used to inform the choice of bandwidth, rather than an automatic procedure. With this method, the bandwidth value associated with the lowest mean squared error (MSE) could be a starting value for smoothing, but other values should also be experimented with. Diggle *et al* (2007) indicate that the MSE method often results in very small bandwidth values due to it having difficulty coping with substantial variation in intensity, and therefore recommend to only use it to guide decision-making. They suggest, for an infectious disease process, to use a bandwidth that is larger than the distance at which direct transmission can occur. Adaptive bandwidth selection methods can be used where the local bandwidth is varied during the estimation process to ensure that a minimum number of observations is included (Bailey & Gatrell, 1995). Further details and discussions in relation to bandwidth selection methods can be found in (Scott, 1992; Wand & Jones, 1995). Edge effects (also called spatial censoring) are likely to bias kernel estimation for locations towards the boundary of the study area (Lawson *et al*, 1999). Adjustments are then necessary where, for example, the area used in the calculation process is adjusted according to the overlap of the circular area defined by the bandwidth and the study region (Diggle, 2000). Kernel density estimation algorithms applied by commercial GIS software products such as ArcGIS (http://www.esri.com) will typically use the bivariate Gaussian kernel and a default algorithm to calculate the bandwidth and not apply corrections for edge effects. More flexible **kernel smoothing** can be conducted using the statistical functions developed for the "R programming language and software environment for statistical computing and graphics" (http://www.esri.com).

The choice of appropriate **grid cell sizes** should be guided by presentational, biological and numerical issues. The cells should not be too large, since the resulting map would not look smooth and may hide relevant patterns. They should not be smaller than what is a sensible resolution for the biological process being studied, for example the cells should not cover areas that are smaller than a typical farm, if farms are the unit of analysis. GIS software will use default algorithms for defining grid cell size, and it is sensible to review the results and experiment with different settings.

An assessment of the spatial heterogeneity of the density of cases of disease is not very meaningful if it is not considered in the context of the population at risk. If no data are available on the population at risk, it would be possible to use the spatial distribution of another disease, assuming it has a different etiology and there is no differential reporting bias (Lawson, 2001). The ratio between the intensity of cases and the population at risk becomes the **log disease risk**, whereas, if the denominator represents the intensity of controls (=non-cases), it is interpreted as a **log relative risk** (Kelsall & Diggle, 1995) (see Example 26.1 and Fig 26.5). To avoid division by zero in these calculations, a kernel function with a non-zero tail (*eg* Gaussian) should be used. Bandwidths may have to be used which are different from the ones for generating the separate case and population at risk surfaces. There are no clear guidelines with respect to whether the numerator and denominator surfaces should be generated using the same or different bandwidths (Bailey & Gatrell, 1995; Bithell, 1990; Diggle, 2000). Schabenberger and Gotway (2005) suggest the use of a visual exploratory approach for choosing the appropriate bandwidth and grid cell size, where the resolution and stability of the estimates is balanced in the context of their biological interpretation. Monte Carlo methods can be used to determine the statistical precision of the kernel estimates (Kelsall & Diggle, 1995).

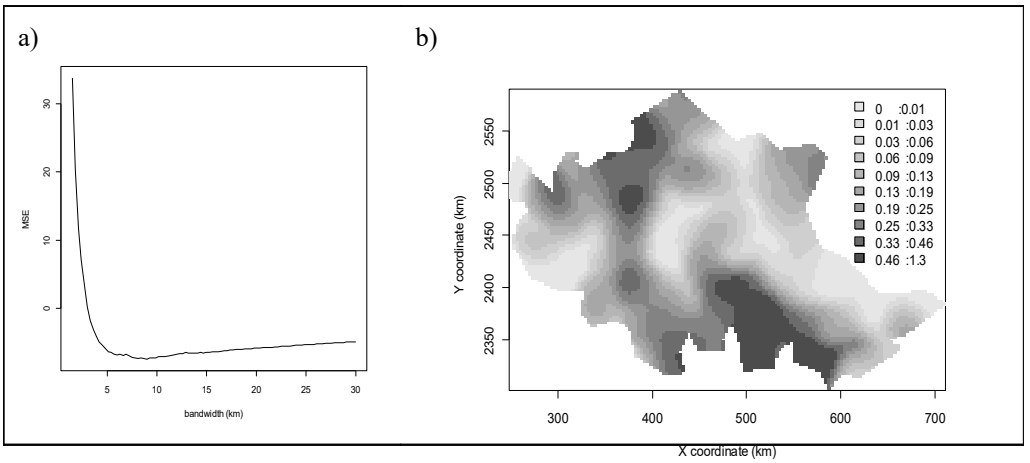


Fig. 26.5 Kernel smoothing for generating a relative risk surface of AI outbreak density in northern Vietnam, based on commune-level disease presence/absence information

26.3.2 Descriptive risk mapping of area data

If the disease and population at risk data are available and aggregated by individual areas within a study region, crude risks or risk ratios can be calculated for each area. The disadvantage of such an approach is that it will not adequately reflect the uncertainty associated with individual areas and its variability between areas. For this reason, empirical or fully Bayesian approaches can be used which will smooth local risk values toward either a global or a local average value. This is done by calculating a weighted average combining the local estimate with a global or local estimate. The calculation of the local average value makes use of the spatial weights matrix described above. The relative weight (*ie* the shrinkage factor) attributed to the local values and the neighbourhood or global values is calculated for each local value, and depends on the variance associated with the local value and the neighbourhood values. If the variance of the local value is high, then the neighbourhood or global value will be given more weight (*ie* the local value will be shrunk toward the neighbourhood or global mean), and vice versa. The global or neighbourhood values are considered to represent the Bayesian prior and the distribution of the observed data influences the likelihood. The combination of prior and likelihood results in the posterior value θ which is to be mapped. Following Bailey and Gatrell (1995), the observed **local rate** is calculated as:

$$r_i = \frac{y_i}{n_i} \qquad \textit{Eq 26.2}$$

where y_i represents the number of diseased animals and n_i the associated denominator for a particular area i.

In **empirical Bayesian analysis**, the mean of the prior γ_i and its variance Φ_i are calculated from the observed data which is described in more detail in Bailey and Gatrell (1995). In fully Bayesian analysis, variables are considered to be random variables with their own prior distributions, called **hyperpriors** (Waller & Gotway, 2004).

The weighting or shrinkage factor w_i is calculated as:

Example 26.1 Kernel smoothing to help visualise spatial heterogeneity of AI outbreaks in northern Vietnam
data = Vietnam

In this example, data in relation to avian influenza (AI) outbreaks between 2004 and 2006 in a northern region of Vietnam was extracted from a larger dataset for the whole country (Pfeiffer *et al*, 2007). The data consists of communes' centroid point locations and a dichotomous variable indicating whether at least one AI outbreak had occurred during the above time period. The objective of the analysis was to provide a more easily interpretable map of the spatial heterogeneity of AI outbreaks in this part of Vietnam. The calculations were performed using the R software with the splancs package.

The **kernel smoothing** method based on a quartic kernel with edge correction was used and the choice of bandwidth was informed by the MSE method mentioned above (Fig. 26.5a). While the lowest MSE was 8km for the outbreak locations, 3 times the value of that was used, as it became apparent on visual inspection of the resulting smoothed maps that it more appropriately reflected the large scale patterns. The same bandwidth was used for both, the outbreak and the non-outbreak locations. The ratio of both kernel smoothed maps is presented in Fig. 26.5b. It shows the increased risk in a relatively large part of the south and in an area to the north of this region. On the basis of this visual assessment, it can be concluded that the risk of outbreak occurrence is not uniform in this part of Vietnam. It seems to be highest toward the southern and north western border of the region.

$$w_i = \frac{\Phi_i}{(\Phi_i + \gamma_i / n_i)} \qquad \textit{Eq 26.3}$$

where Φ_i is the variance of the prior in area *i*.

The **posterior value** for local rate θ_i is calculated as:

$$\hat{\theta}_i = w_i r_i + (1 - w_i)\gamma_i \qquad \textit{Eq 26.4}$$

where the weighting factor w_i determines how the local rate value and the prior γ_i are to be summed.

While these methods provide a statistically more reliable representation of the underlying data, it is to be noted that the smoothed values will be different from the crude values, which could result in more difficult interpretation, and may hide some important local patterns. Nonetheless, in general, this smoothing method will allow for a more meaningful assessment of large-scale spatial patterns, as there will be less noise caused by local small sample sizes. (See Example 26.2 and Fig 26.6.)

26.3.3 Spatial cluster analysis

The spatial arrangement of infected herds or animals can be used to investigate hypotheses in relation to causal risk factors. If spatial proximity has no influence on the risk of infection, infection should be spatially randomly distributed. A clustered spatial arrangement suggests the presence of a contagious process or that there is a localized risk factor. The objective of spatial cluster analysis is therefore to test the null hypothesis of complete randomness of a spatial process. In the case of point locations, a random spatial process would follow a Poisson distribution. Since the occurrence of infectious diseases depends on the presence of susceptible herds/animals which usually are not randomly distributed in space, their spatial arrangement needs to be taken into account. Fig. 26.7 illustrates some of the key issues associated with

Example 26.2 Empirical and fully Bayesian smoothing to visualise spatial heterogeneity of AI outbreaks northern of Vietnam

data = Vietnam

The same data as described above was used for this example. But for this analysis, commune-level avian influenza outbreak occurrence was aggregated at district level as number of communes with and without outbreaks. The objective of this analysis is to describe the spatial pattern of disease risk in the area. This analysis was conducted using the R software with the spdep package and the statistical software WinBugs (http://www.mrc-bsu.cam.ac.uk/bugs/) for the fully Bayesian modelling.

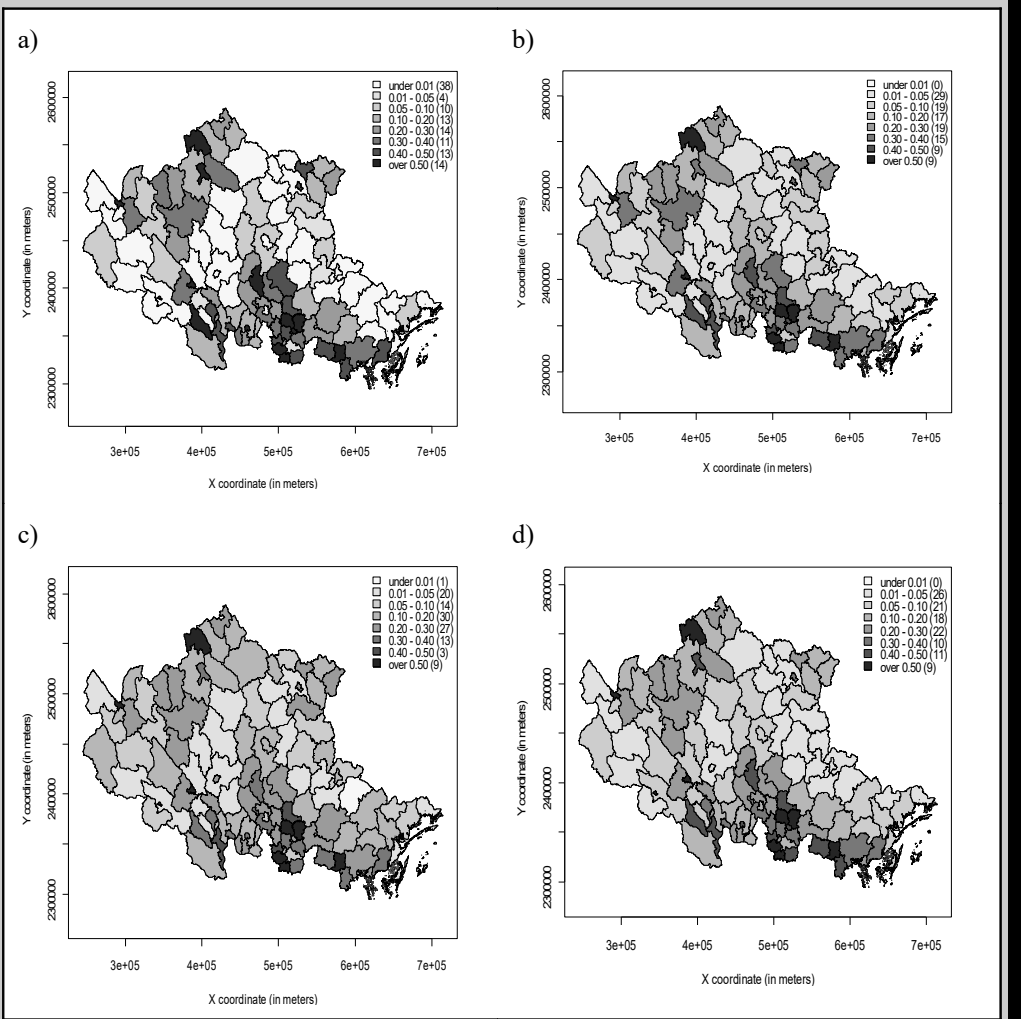


Fig. 26.6 Different approaches to presenting AI disease risk using choropleth maps, based on commune data aggregated at district-level

(*continued on next page*)

Example 26.2 Empirical and fully Bayesian smoothing (*continued*)

The map in Fig. 26.6a showing the crude risk estimates indicates significant variation between local estimates with higher risks toward the southern and the northern part of the area, whereas low risk in between. The globally smoothed empirical risk estimator based on the average risk across all districts and its variance as the prior is shown in Fig. 26.6b. It removes the extremely low and high values, and in particular the areas with zero risk. The locally smoothed map in Fig. 26.6c is based on including the values from the 4 nearest neighbouring districts into the prior, but the observed pattern is only marginally different from the one in Fig. 26.6b. A fully Bayesian approach was applied to the data in Fig. 26.6d, based on conditional autoregression allowing for a spatial and a non-spatial random effect (further details will be discussed in the spatial regression section below). In summary, in this instance the broad similarity of the patterns obtained from using the crude and the smoothed values suggests that the crude estimates are relatively stable due to high sample sizes (*ie* number of communes) in most districts. The results suggest the presence of disease clustering in the southern and northern part of the area, and this will be more formally investigated using the cluster investigation methods described in the next section.

identification of spatial clustering. A spatial distribution as presented in Fig. 26.7a is easy to identify, but occurs only rarely. More often the spatial distribution may be more like Fig. 26.7b, where some of the herds affected by a disease are located outside the obvious clusters, and it is therefore difficult to determine whether there is clustering. In reality, the affected herds will be a subset from a larger population at risk which in itself potentially occurs in a clustered fashion, as shown in Fig. 26.7c. And, in addition to that, there may be non-spatial risk factors, such as farm type, which influence the likelihood of a herd becoming infected (Fig. 26.7d). The methods described in this section will assist in addressing the problems presented in Fig. 26.7 b and c.

Statistical **significance testing** usually involves comparing an observed test statistic with a theoretical null hypothesis probability distribution of that test statistic representing all test values possible under randomness. The use of such theoretical distributions requires that specific assumptions are met, including for example a rectangular shape of the study area, independence of observations and a very high number of observations. Because in spatial analysis these are often difficult to achieve, randomisation or **Monte Carlo simulation** methods where the null hypothesis distribution will be generated from simulated data are usually applied. Note that the number of iterations specified in Monte Carlo simulation will define what the smallest detectable P-value will be. If 999 iterations are used, this means that a total of 1,000 values (999 for simulated plus 1 for observed data) for the test statistic will be available for generating the null hypothesis distribution of the test statistic. If the observed value is the most extreme value it has a chance not higher than 1/1000 to occur ($P <= 0.001$).

Methods used for spatial cluster analysis can be broadly categorised into global, local and focused statistics. **Global statistics** indicate whether there is clustering somewhere in a study area, **local statistics** will identify the location(s) of the clusters, and **focussed statistics** will test the presence of clustering around pre-specified spatial locations. In addition, it is possible to assess the presence of time-space clustering.

The null and alternative hypotheses tested by these methods vary, as outlined by Waller and Gotway (2004). The methods vary in their statistical power, and particularly when using aggregated data and when there is spatial heterogeneity in the population at risk. Within a study area, clusters are more likely to be detected in areas with large population sizes than where the

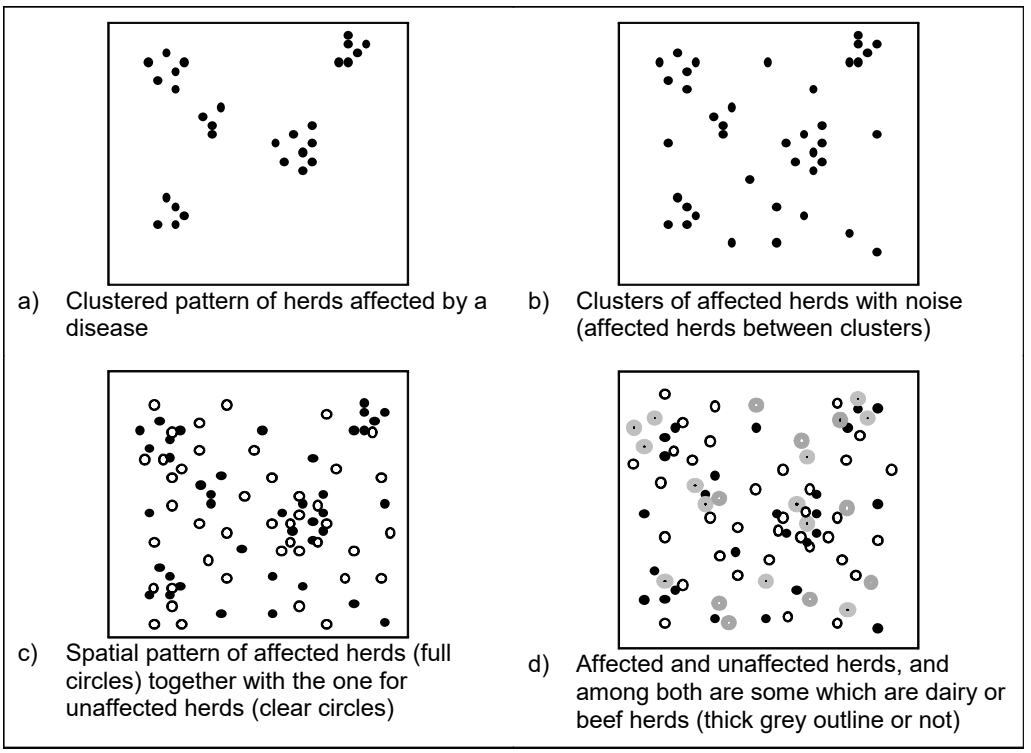


Fig. 26.7 Examples of spatial patterns

population at risk is sparse, which means that statistical power can be spatially heterogeneous.

Investigations of spatial clustering can be initiated as a result of suspicion or detected as part of routine surveillance. Such perceived clusters could then be subjected to a statistical analysis which, if they indicate a non-random distribution of disease, may justify more detailed causal epidemiological investigations. Centers for Disease Control (CDC) devised a formal framework for such investigations (Anon, 1990). As is the case with other epidemiological investigations, this approach is affected by data quality problems, bias, diagnostic sensitivity/specificity and statistical power. For this reason but mainly because perceived clusters are not necessarily real, it has been reported that only a very small percentage of 'cluster alarms' ever leads to identification of a causal factor (Wakefield *et al*, 2000).

26.4 GLOBAL SPATIAL CLUSTERING

The analysis of clustering may be based on point locations or on data presented in an aggregated fashion, such as counts of disease per district. For both data types, the observed pattern is to be compared with a spatial distribution that would be found if there was no clustering (*ie* null hypothesis distribution of a particular summary statistic). The term **spatial autocorrelation** is used when quantitative measures such as counts, risks or rates are being assessed. When interpreting the results of global spatial autocorrelation analyses, it needs to be recalled that they can be significantly biased by the presence of first-order effects in the data. Furthermore, these approaches are aimed at detecting effects averaged over the whole study

area which means that they may lack statistical power to detect single clusters within large study areas.

26.4.1 Methods for point data

There are 2 groups of methods included here. One assesses clustering in relation to the spatial distribution of a disease where each point represents, for example, the location of an affected herd. The other assesses clustering in relation to an attribute value, such as incidence of affected animals per herd.

Cuzick and Edwards test
If point locations of cases and controls are available, the **Cuzick and Edwards** test will assess the null hypothesis that the nearest neighbour to a randomly selected case is just as likely to be a control as another case (Cuzick & Edwards, 1990). The associated test statistic is the sum of all cases which have another case as nearest neighbour (see Example 26.3). The nearest neighbour criterion can be modified, in that higher orders of nearest neighbourhood (*eg* second, third or more nearest neighbour) can be used, so that different scales of clustering can be detected. Statistical significance is tested using an asymptotic test based on the Gaussian distribution. Alternatively, the null hypothesis distribution of the test statistic can be generated using Monte Carlo simulation involving random allocation of case/control status to each location, and repeating this many times, while recalculating the test statistic each time.

***K*-function**
The number of events of the same type occurring within a certain distance can be expressed using the *K*-function. It assumes stationarity and isotropy, and reflects second-order effects of a spatial process. The associated mathematical function for the *K*-function using the notation in Bailey and Gatrell (1995) is:

$$K(h)=\frac{1}{\lambda^2 R}\sum_{i=1}^{n}\sum_{j=1}^{n} I_h(d_{ij}) \qquad \textit{Eq 26.5}$$

where h stands for distance, λ represents the intensity of the spatial process for a given area R, d is the distance between events i and j within R and I_h is an indicator which is 1 for $d_{ij} \leq h$ and 0 otherwise (see Example 26.4 and Figs. 26.8 and 26.9).

Algorithms are available which take account of edge effects, such as through a proportional weighting process (Ripley, 1987). When calculating the *K*-function, a distance scale needs to be defined. Up to one third of the linear extent of the area should be used, since *K*-function estimates become inefficient for larger distances. The result can be visualised by plotting K against h, where under complete spatial randomness the function should have the shape of a parabola. To facilitate the interpretation, transformations of K have been recommended, so that a comparison with a straight line, reflecting randomness, can be used instead (Bailey & Gatrell, 1995). The *K*-function as such describes the spatial distribution of a population, but it does not take into account the likely spatial heterogeneity of other populations, such as a population at risk. This can be achieved by comparing the *K*-functions for the spatial distribution of case and control events through calculating a **difference function** *D(h)* as shown below.

$$D(h)=K_{\text{cases}}(h)-K_{\text{controls}}(h) \qquad \textit{Eq 26.6}$$

The difference function *D(h)* expresses the scaled expected number of excess cases occurring at a given distance from a random reference case. The significance of the difference function can

Example 26.3 Cuzik and Edwards test of clustering of AI outbreaks in Vietnam
data = Vietnam

This analysis is aimed at determining whether cases are more likely to have other cases as their nearest neighbours, and this is done here for between 1 and 5 nearest neighbours. The analysis was performed using the commercial spatial analysis software ClusterSeer Version 2.2.8.1 (www.terraseer.com).

The statistical significance was tested using Monte Carlo simulation based on random shuffling of case-control status for 999 times (*ie* smallest detectable P–value P = 0.001). The analysis included 446 case locations and 1,851 controls.

					Upper-tail Monte Carlo	
k	T[k]	E[T]	Var[T]	z	P-value+	P-value
1	165	86.442	96.765	7.986	0.000	0.001
2	340	172.883	205.284	11.664	0.000	0.001
3	518	259.325	313.150	14.618	0.000	0.001
4	673	345.767	424.901	15.875	0.000	0.001
5	811	432.208	534.973	16.377	0.000	0.001

Combined P-value for Monte Carlo Randomization: Bonferroni P-value: 0.005

The above results indicate that there was clustering at all neighbourhood distances specified. In addition, the overall combined P-value (corrected for multiple testing using the Bonferroni method) was also significant. Therefore, communes affected by an AI outbreak are more likely to have another outbreak commune as nearest neighbour than an unaffected one. Since it is a global statistic, no information is generated about the number and location of any clusters.

be tested by using Monte Carlo simulation, where all events are randomly labelled as given numbers of case and control events. If this is done many times, the resulting difference functions can be used to define simulation envelopes with which the observed difference function D could be compared. The use of a Monte Carlo simulation approach allows statistical inference for $D(h)$ in the presence of non-stationarity (Diggle, 2000). If the latter extends beyond the simulation envelope, it can be concluded that the observed pattern of case events may represent clustering. Diggle (2000) emphasizes that the difference function can also be used to estimate the size of clustering, but the spatial scale at which clustering occurs needs to be small relative to the full extent of the study area. The **Diggle-Chetwynd test statistic** can be used to obtain an estimate of statistical significance of the departure of the observed pattern from randomness (Diggle & Chetwynd, 1991). Diggle *et al* (2007) extended the K-function to allow the assessment of clustering, for an inhomogeneous Poisson process, by including an estimate of the spatially varying intensity into Eq 26.5. The kernel smoothed intensity parameter is estimated using a logistic regression conditioning on locations of cases and controls and allowing for inclusion of confounding factors and true risk factors.

The K-functions presented in Fig. 26.9a show the values, obtained using Eq 26.5, for increasing distances between points. A visual comparison of the curve patterns for affected and unaffected commune centroid locations suggests that there is a slight difference in the expected number of observations at distances of up to 0.7 decimal degrees. The difference function $D(h)$ in Fig. 26.9b supports this observation since the function based on the observed data extends beyond

the simulation envelope generated by random labelling of the observations (n=99) as affected or unaffected by AI, and recalculating the resulting difference functions which are then summarised by the simulation envelope. The conclusion is therefore that, in this region, AI outbreaks were clustered. **Note** this analysis did not correct for edge effects. The Diggle-Chetwynd statistic *D* is 41.7, and is associated with a P-value lower than 0.0001 obtained using Monte Carlo sampling, also suggesting that the observed pattern is clustered in space (see Figure 26.9c). The non-homogeneous *K*-function based on a bandwidth value of 0.7 decimal degrees is shown in Fig. 26.9d. It mirrors the pattern shown for the difference function. The P-value is 0.01, confirming that the observed pattern is statistically significant.

26.4.2 Methods for aggregated data

Moran's *I*, Geary's *c* and Getis-Ord *G* statistic can be used for assessing spatial autocorrelation of attribute data measured at an ordinal or continuous scale. All are variations of a cross-product statistic which produces a similarity index weighted by proximity (Haining, 2003).

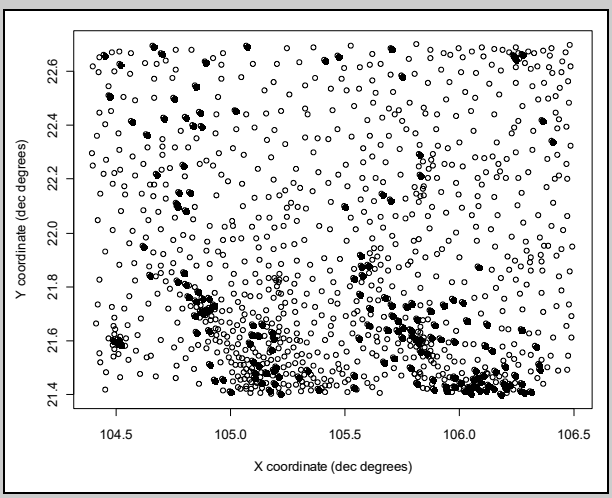
Moran's *I*

If data are available in an aggregated format, such as counts of diseased cases per district, global spatial autocorrelation can be estimated using the Moran's *I* index which is similar to the

Example 26.4 *K*-functions to detect spatial clustering of AI outbreaks in Vietnam
data = Vietnam

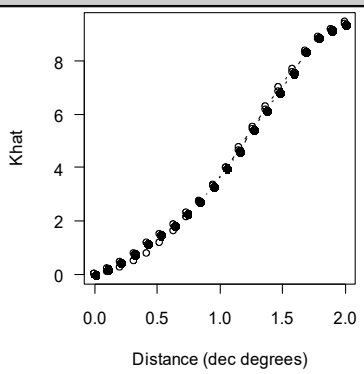
In this example, the same data (as above) on avian influenza outbreaks in the northern region of Vietnam was used (Pfeiffer *et al*, 2007). The objective of this analysis is to determine whether there is spatial clustering of avian influenza outbreaks at commune level. A rectangular area was selected to avoid complex edge effects in the analysis resulting from the uneven border of the complete region. This analysis was conducted using the R software with the splancs and spatialkernel packages. The R code for the inhomogeneous approach was based on Bivand *et al* (2008).

Fig. 26.8 Locations of communes in northern Vietnam affected (filled circles) or unaffected (empty

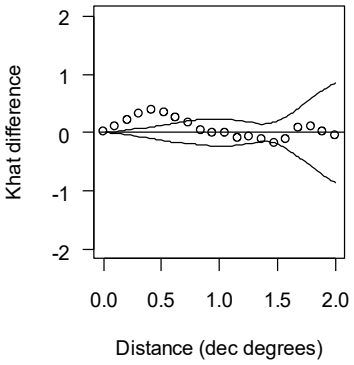


(*continued on next page*)

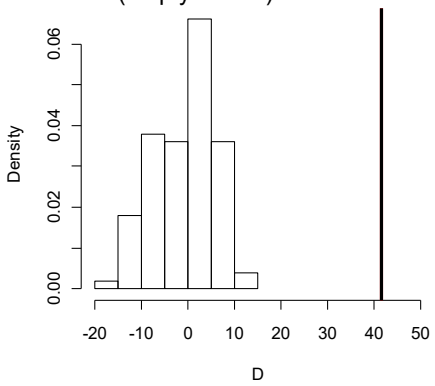
Example 26.4 *K*-functions (*continued*)



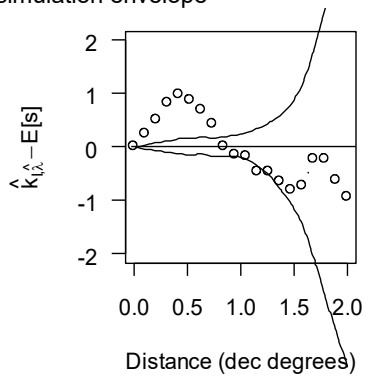
a) *K*-functions for affected (filled circles) and unaffected (empty circles) communes



b) Difference function *D*(*h*) including 95% simulation envelope



c) Histogram of observed value for Diggle-Chetwynd statistic (D) and distribution of values obtained under the null hypothesis of no clustering



d) *K* function for inhomogeneous population distributions including 95% simulation envelope

Fig. 26.9 *K*-function analysis of 2004-06 AI outbreaks for a northern region in Vietnam

global spatial autocorrelation can be estimated using the Moran's *I* index which is similar to the Pearson's product moment correlation coefficient, as shown below (Bailey & Gatrell, 1995):

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \hat{y})(y_j - \hat{y})}{\sum_{i=1}^{n}(y_i - \hat{y})^2 (\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})} \qquad \textit{Eq 26.7}$$

where n is the observation number, y_i and y_j are the respective values, $\hat{y}$ is the global mean and w_{ij} is the spatial proximity matrix.

Moran's *I* expresses the similarity between local values of continuous-scale measurements and their neighbours, based on their deviation from a global mean value (see Example 26.5 and Fig 26.10). A proximity matrix is required in order to express the spatial arrangement, and in this

case it is usually based on the neighbourhood or distance between locations. It is used to attribute weights to pairs of values which are then compared with respect to their similarity. When neighbouring values tend to be similar, *I* will be positive. If they tend to be different, *I* will be negative. Statistical significance can be tested by Monte Carlo simulation through random allocation of observed data values to locations available in the dataset. If the observed value of *I* is located in the tail of the simulated distribution, it is an indication of the presence of autocorrelation. The results from such analyses if conducted using disease count data will not take account of potential underlying heterogeneity of the population at risk. It will then be more useful to assess the data as disease proportions or rates. Since the denominators for these individual values will usually be different, the assumption of constant variance will be violated. In this case, adaptations of Moran's *I* can be used, such as Oden's method (Oden, 1995). Waller and Gotway (2004) recommend the method described by Walter (1992) which compares observed counts with expected counts under the constant risk hypothesis. The resulting statistic becomes a weighted cross-product of the difference between observed and expected counts. Its statistical significance can be tested using Monte Carlo simulation. Assunção and Reis (1999) proposed an empirical Bayes approach to adjust for differences in population sizes. It is also important to be aware that first-order or spatial-trend effects in the data may result in biased autocorrelation estimates.

Geary's *c*

In contrast to Moran's *I* which assesses the similarity in relation to deviations from the global mean, Geary's contiguity ratio (or Geary's *c*) is based on the average difference between pairs weighted as specified by the proximity matrix. The calculation is as follows (Waller & Gotway, 2004):

$$c = \frac{n-1}{2\sum_{i=1}^{n}(y_i-\hat{y})^2} * \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-y_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \qquad \textit{Eq 26.8}$$

where n is the observation number, y_i and y_j are the respective values, $\hat{y}$ is the global mean and w_{ij} is the spatial proximity matrix (see Example 26.6).

Values of 0 and 2 reflect perfect positive and negative autocorrelation, respectively. The closer the value is to 1, the more homogenous the spatial distribution of values will be. As in the case for Moran's *I*, heterogeneous distribution of the population at risk will adversely affect the validity of the analysis.

Spatial correlogram

Moran's *I* and Geary's *c* can be calculated for different distances (or spatial lags), and presented as a correlogram by plotting the resulting values against the corresponding lag (de Smith *et al*, 2007) (see Example 26.7 and Fig. 26.11). The information presented by plotting Geary's *c* values in a correlogram is similar to what is shown in a semi-variogram. Bailey and Gatrell (1995) point out that neighbouring values in such plots are highly correlated, and that correlations at larger lags are partly a function of correlation at smaller lags. Therefore, if there are any peaks at low lag values, further peaks at higher lags need to be interpreted with some caution. Furthermore, this analysis assumes that the spatial process is isotropic, and that first-order spatial effects are eliminated, since they will otherwise bias the plot.

Example 26.5 Moran's *I* to assess spatial correlation of AI outbreaks in Vietnam
data = Vietnam

The same dataset on avian influenza outbreak occurrence in a region of northern Vietnam (described above) was used to assess for the presence of spatial autocorrelation. The commune-level point data were aggregated at district level to represent the number of affected communes per district, as well as the number of communes at risk. This analysis was conducted using the R software with the spdep package.

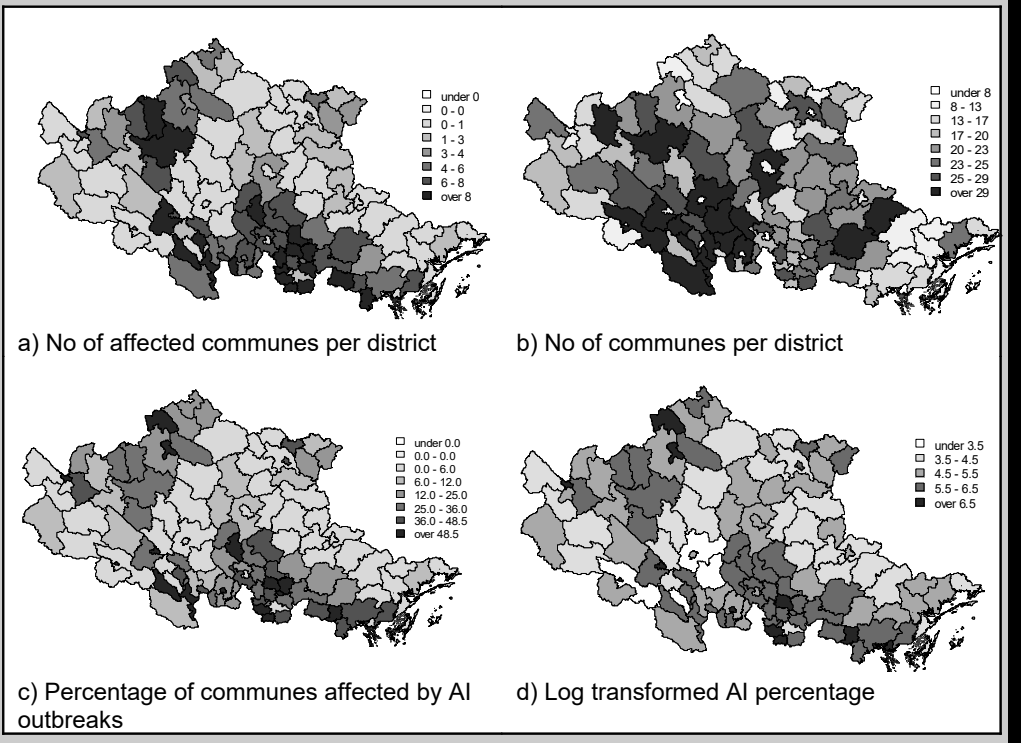


Fig. 26.10 Choropleth maps showing number of communes affected per district by AI outbreaks between 2004 and 2006, (total number of communes, percentage of communes with outbreaks as well as log transformed percentage for each district in a northern region of Vietnam).

A spatial weights matrix was generated based on the 4 closest neighbouring districts. This was then used to calculate Moran's *I* under randomization, resulting in I=0.33 (P<0.0001), indicating the presence of positive autocorrelation between districts in the number of communes with AI outbreaks. Since the number of communes at risk varies between districts, the data was reanalysed using the empirical Bayes adjustment of the Moran's *I*. The resulting Moran's *I* value was 0.28 with a P-value of 0.001 based on 999 Monte Carlo simulations, indicating that even after taking account of heterogeneity in the number of communes per district, there is statistically significant positive autocorrelation. Using incidence proportion as the outcome variable, resulted in a Moran's *I* of 0.23 (P<0.0001) for the untransformed and in 0.30 (P<0.0001) for the log transformed incidence values. Of these 4 analyses, the empirical Bayes correction and the one based on the log transformed incidence values provide the statistically most robust estimates.

Example 26.6 Geary's *c* to assess spatial correlation of AI outbreaks in Vietnam
data = Vietnam

Geary's *c* was also calculated using a Monte Carlo simulation approach for the AI outbreak data as in the Moran's *I* example, resulting in c=0.64 (P<0.001) for the number of outbreaks and 0.77 (P< 0.001) for the incidence proportion. Since these values are less than 1, they indicate positive autocorrelation. This analysis was conducted using the R software with the spdep package.

26.5 LOCALISED SPATIAL CLUSTER DETECTION

Global indices of spatial association provide inference in relation to the whole study area. The presence of global positive autocorrelation needs to be interpreted cautiously, since it could have arisen from spatial trends (first-order effects) or local clustering (second-order effects). It is therefore important to reduce the influence of trends and variance heterogeneity on the index. But even then, a global index will not allow specifying clusters and may have insufficient statistical power for detecting single clusters (Waller & Gotway, 2004). As a consequence, various local indicators of spatial association have been developed which have particular relevance in hypothesis-driven epidemiological investigations as well as in disease surveillance.

26.5.1 Methods for point data

Spatial scan statistic
One of the most commonly used exploratory spatial analysis methods is the spatial scan statistic (Kulldorff *et al*, 1997). It is based on comparing the risk of disease within a circular window to

Example 26.7 Correlogram showing correlation of AI outbreaks at different distances in Vietnam
data = Vietnam

Moran's *I* values for different spatial lags were generated for the log transformed AI incidence per district in northern Vietnam. The results indicate that Moran's *I* is significant up to a lag of 2 neighbours, since beyond that the standard deviation bars include the value zero. This analysis was conducted using the R software with the spdep package.

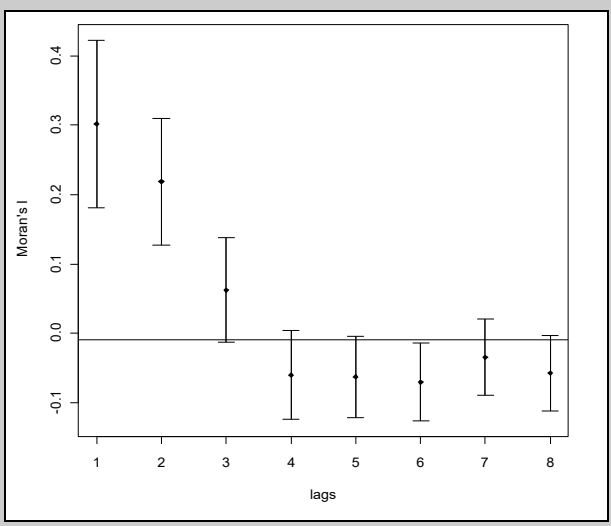


Fig. 26.11 Correlogram for Moran's I

Note Point estimate including +/- twice the square root of its variance) against different spatial lags (number of neighbours) applied to log transformed district level AI incidence data

that outside. The algorithm varies the radius of the windows up to a user-defined upper bound, and moves them around the study area. The centre of these windows can be either defined by a regular grid, or is based on the locations of the observed data, where the latter will more appropriately represent any possible spatial heterogeneity of the data. A local likelihood function is calculated from the product of the individual Bernoulli probabilities for each of the windows. The resulting **likelihood ratio statistic** is the maximum of these local functions under the alternative hypothesis (risk higher inside than outside the window) divided by its maximum value under the null hypothesis (risk inside equals risk outside) (see Example 26.8 and Fig. 26.12). As described by Waller and Gotway (2004) the overall test statistic T is proportional to:

$$T_{\text{scan}} = max\left(\frac{c_{\text{in}}}{n_{\text{in}}}\right)^{c_{\text{in}}}\left(\frac{c_{\text{out}}}{n_{\text{out}}}\right)^{c_{\text{out}}} \text{I}\left(\frac{c_{\text{in}}}{n_{\text{in}}} > \frac{c_{\text{out}}}{n_{\text{out}}}\right) \qquad \textit{Eq 26.9}$$

where c_{in} and c_{out} represents the number of cases inside and outside the window, and n_{in} and n_{out} the corresponding denominator. I is the indicator function with I=1 if the condition is met.

The window with the maximum of the local likelihood ratio functions indicates the location of the most likely cluster. Its statistical significance is determined using Monte Carlo simulation resulting in a single P-value and thereby avoiding multiple testing problems. The method also identifies secondary clusters for which adjusted statistical inference can now be obtained. The maximum population size to be included in any given window needs to be pre-defined, and there is no objective guideline on how to do this. The method can also be used to identify clusters of low risk, adjust for covariates, and it can search for overlapping or non-overlapping clusters. It can be used for Poisson counts (*ie* aggregated data) and to investigate time-space clustering. It has more recently been extended to allow use of ordinal, normal and exponential (=survival) type outcome data. The method also can now assess non-circular shaped windows. It is available as the public-domain software SaTScan (www.satscan.org).

With ordinal data, each case can belong to one of several outcome categories, and with exponential or survival data each case is defined by a censoring variable (1=event, 0=no event) and a continuous-scale variable expressing time to event or censoring.

26.5.2 Methods for aggregated data

This group of methods includes the **local indicators of spatial association** or LISAs. Anselin (1995) defined these local measures of similarity between neighbouring regions values. Mapping these values allows identification of areas of high and low local spatial association. This means that they indicate areas with similar data values, be they low, moderate or high. Anselin requires LISAs to link to a global indicator of spatial association, of which Moran's I is the most popular. An advantage of a local measure is that they are less affected by first-order effects than the associated global indicator. Low spatial association can also be used to identify outliers, which may be an indication of data errors. The statistical testing of these local indicators, even if using Monte Carlo simulation, is affected by several issues, including multiple testing (Waller & Gotway, 2004). Therefore testing should only be used for exploratory purposes. The spatial scan statistic can also be used for this type of data, although this method does not belong to the group of LISAs. In this case, the likelihood is based on the product of independent Poisson distributions. The centroid coordinate locations for each area, for each of which the number of cases and the population at risk will have to be specified, are used in this analysis.

Example 26.8 Spatial scan statistic to detect clusters of AI outbreaks in Vietnam
data = Vietnam

The data on AI outbreaks in northern Vietnam was used for this analysis, including 2,296 locations, 446 of which had outbreaks. The spatial scan statistic was used to determine the locations and size of potential spatial clusters. The analysis was conducted using the Bernoulli model for case-control point data, for 999 iterations, testing for high-risk clusters allowing for cluster sizes including a maximum of 50% of the population. The analysis was conducted allowing for circular as well as elliptic clusters, and was conducted using the SaTScan software version 8.0 (www.satscan.org).

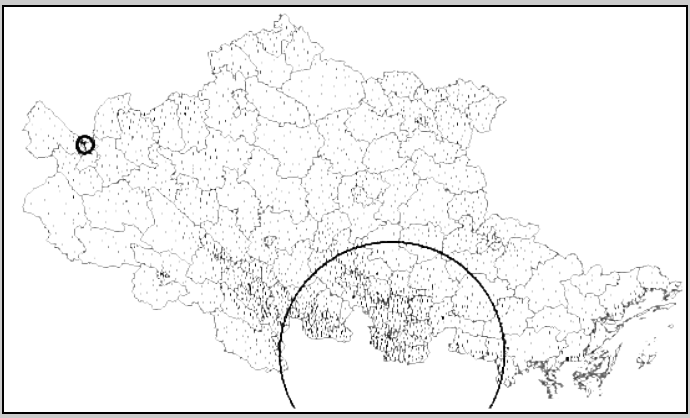


Fig. 26.12 Location of statistically significant clusters for AI outbreak occurrence between 2004 and 2006

(Map of districts in a northern region of Vietnam: large circle = most likely cluster; small circle = secondary cluster; black dots = communes with AI outbreak; grey dots = communes without AI)

The analysis identified two circular clusters (see Pfeiffer *et al*, 2007), one statistically significant, most likely cluster (*RR*=2.9; p<0.001) with a radius of 82 km and one statistically significant secondary cluster (*RR*=5.2; P=0.045) with a radius of 5.6 km. These results also illustrate the problem with circular clusters, since more than half of the area covered by the most likely cluster is outside the study area. No elliptic-shaped clusters were identified.

Local Moran test

The global Moran's *I* statistic can be disaggregated into a series of local Moran's *I* values of spatial autocorrelation for each area included in the study region, as shown below (Bivand *et al*, 2008):

$$I_i = \frac{(y_i - \hat{y}) \sum_{j=1}^{n} w_{ij} (y_j - \hat{y})}{\frac{\sum_{j=1}^{n} (y_i - \hat{y})^2}{n}} \qquad \textit{Eq 26.10}$$

where n is the observation number, y_i and y_j are the respective local values, $\hat{y}$ is the global mean and w_{ij} is the spatial proximity matrix.

The mean and variance for each resulting local Moran's *I* value can be obtained using

randomization. Standardization can be applied and, as a consequence of using different weights, may produce different numeric values when aggregated to the global Moran *I*. The **Moran scatter plot** provides an effective means of presenting the data (Anselin, 1995). The X-axis on the plot represents the standardized local value, and the Y-axis the weighted average of the standardized neighbouring values with neighbourhood defined by the weights matrix. Clustering is represented by data points in the lower-left (low-low) and upper-right (high-high) quadrants of the scatter plot. Values located in the remaining quadrants (high-low and low-high) are dissimilar from their neighbours, and could be spatial outliers. The slope of a regression line fitted to this data represents the global Moran's *I* statistic (see Example 26.9 and Figs 26.13 and 26.14).

26.5.3 Focused spatial cluster detection

Often in spatial cluster analysis, the research hypothesis may be that the risk of disease is increased in the proximity of pre-defined foci which may be particular geographical point, line or area feature, such as power lines or nuclear power plants. With testing such hypotheses, it is important to be aware of the pre-selection bias resulting from applying the 'Texas sharpshooter' principle, where one shoots the shed first and then places the bull's eye around the bullet hole. Potential foci associated with clusters should be defined based on causal hypotheses before any data analyses.

The **Lawson-Waller local score test** is a goodness-of-fit test, calculated by summing the deviation of observed from expected case numbers across areas weighted by exposure. Exposure levels can be specified in different ways, including inverse distance from the foci. Statistical significance can be tested using exact methods or Monte Carlo simulation. Morris and Wakefield (2000) provide an in-depth discussion of this topic.

26.6 SPACE-TIME ASSOCIATION

Methods investigating space-time association of disease occurrence can be broadly categorized into those which focus on detection of clusters and those aimed at space-time interaction. **Space-time clusters** are present if disease occurrence is not only clustered in absolute space, but also in time. The procedure usually requires data on cases as well as non-cases. If a disease process is infectious, proximity of cases in both space and time is likely to occur. In this situation, only case data are required, and the presence of space-time interaction can be tested. Such a process may actually move spatially, as long as, at any time, new cases occur in spatial proximity to other cases occurring at the same time (not those occurring at other times).

26.6.1 Space-time interaction tests

All these tests require data about the time and location for cases of disease, but not data on the population at risk and its geographical distribution. Therefore they are not sensitive to a heterogeneous distribution of the population risk, but bias will occur if the density of the population changes spatially at different rates (Kulldorff, 1998). Knox (1964) developed a simple space-time test which required categorising all pairs of cases into whether they occurred close or far apart in space and time, thereby producing a 2X2 table. The criteria for specifying closeness in either the temporal or spatial dimension are subjective, and the method is only appropriate for diseases with short incubation or latency period. The statistical hypothesis of

independence in space and time is tested using a Poisson distribution of the counts in the 2X2 table. Norström *et al* (2000) used the **Knox test** to examine the pattern of an outbreak of acute respiratory disease in cattle in Norway, and it allowed them to test the hypothesis of potential airborne transmission mechanism for this infectious disease. As an alternative analytical method for this type of data, the **Mantel test** or regression uses the numeric distance in space and time between all pairs of cases (Mantel, 1967). The statistical hypothesis of independence in space-time is assessed using permutation or Monte Carlo tests. Constant values can be added to the space-time distance measures to bring them on the same scale, or they can be transformed

Example 26.9 Local Moran test of AI outbreaks in Vietnam
data = Vietnam

Using the AI outbreak data, the local Moran *I* can be calculated for log transformed AI incidence data. The objective of this analysis is to identify potential clusters and outliers. The analysis was performed using the spdep package of the R software.

The Moran scatter plot in Fig. 26.13 indicates that some observations in this dataset potentially are outliers (mainly those in the top-left and bottom-right quadrant). The points in the top-right and bottom-left quadrant of the plot reflect districts which are autocorrelated with their neighbourhood. The map in Fig. 26.14 shows the districts for which the local Moran statistic is statistically significant, indicating local clustering of outbreaks in the corresponding districts. It is notable that the pattern is very different from what was identified using the spatial scan statistic, but that is partly due to the fact that it also identifies areas where low values are surrounded by low values. It needs to be recognized that the spatial scan statistic compares risks between inside and outside circular or elliptically shaped areas. In contrast, local Moran *I* examines local spatial autocorrelation between values within an area and the average of the neighbourhood; hence, it focuses on similarity of values rather than actual risk comparison.

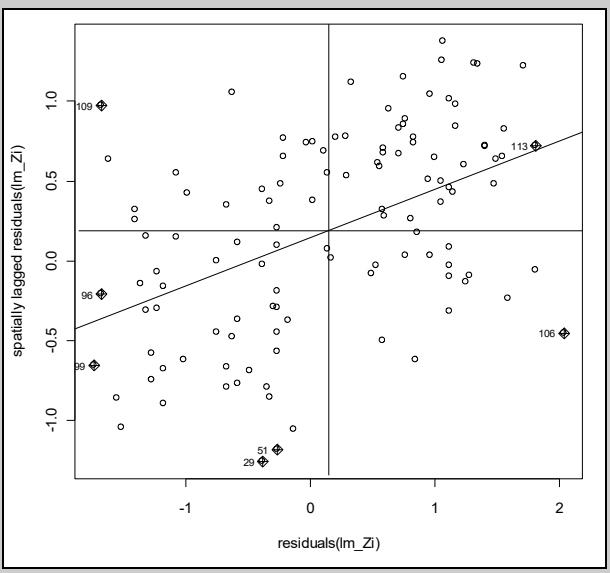


Fig. 26.13 Moran scatterplot for log transformed AI outbreak incidence data

(*continued on next page*)

Example 26.9 Local Moran test (*continued*)

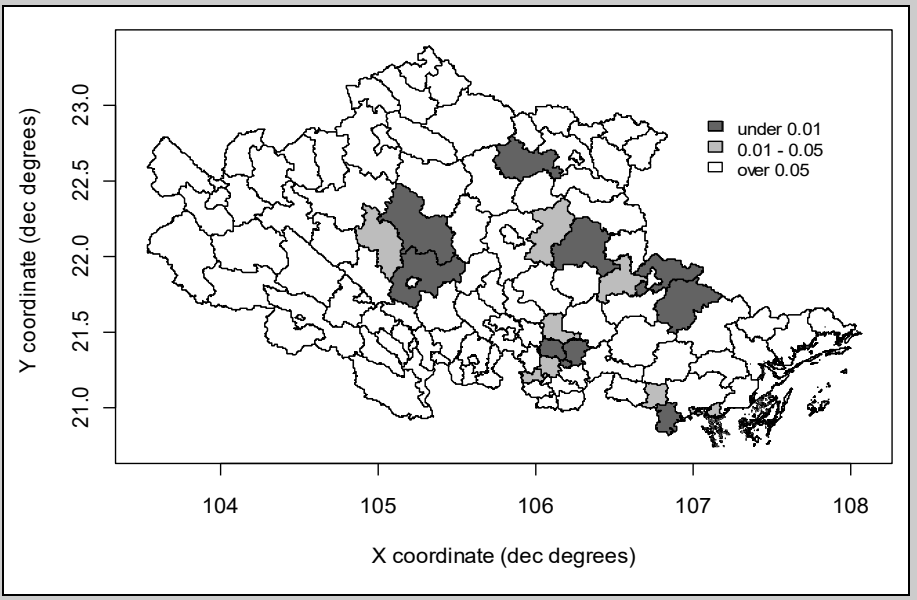


Fig. 26.14 Choropleth map indicating districts in northern Vietnam with statistically significant local Moran *I* values for log transformed AI incidence

to reduce the influence of outlying values. The results of the analysis will be influenced by such somewhat arbitrary decisions. The Mantel test can be used for assessing association between any pair of distance matrices, including space-genetic distances. It is also possible to generate a partial Mantel test where a third distance matrix can be taken into account in the analysis (Goldberg *et al*, 2000). The **Jacquez *k*-nearest neighbour test** expresses proximity at k nearest neighbours rather than absolute distance, and will therefore be less affected by spatial variation in the density of the population at risk than the methods above (Jacquez, 1996). It can be calculated for a single k value or using a series of such values for producing a summary statistic. Kulldorff (1998) notes that the above methods all assume that the spatial distribution of the population at risk does not change differently in space over time. It needs to be emphasized that if these statistics are not significant, it may still be that temporal or spatial clustering is present. Ward and Carpenter (2000) provide a detailed discussion with examples of the Knox, Mantel and k-nearest neighbour tests.

The **bivariate space-time *K*-function** $K(h,t)$ expresses the expected number of events occurring within a given spatial (h) and temporal (t) distance from a random event, relative to the intensity of events per unit, space and time. This could be visualised as a space-time cylinder centred on the random event (Diggle *et al*, 1995). The function can be corrected for edge effects. If there is no space-time interaction then $K(h,t)$ should be the product between $K(h)$ and $K(t)$. An appropriate statistical test would be to calculate the difference $D(h,t)$ between $K(h,t)$ and the product term $K(h) * K(t)$. The resulting values can be presented as a three-dimensional plot of $D(h,t)$ against spatial and temporal distance values. High $D(h,t)$ values indicate space-

time interaction. Monte Carlo methods are used to generate the null-hypothesis distribution for this statistic. The sum of all observed $D(h,t)$ is then compared with this generated empirical distribution. The test is robust to non-stationarity of the underlying process. Diggle (1995) notes that the space-time K-function only provides useful information at values for h and t which are small relative to the total study area. Porphyre *et al* (2007) applied the space-time K-function to investigate factors influencing the occurrence of TB outbreaks in a specific geographical area in New Zealand. The use of this method assisted them in concluding that farm-to-farm spread was less important than other potential sources of infection.

26.6.2 Space-time cluster detection

The spatial scan statistic described above can be extended to include time, with the basic principle being that a cylinder instead of circle or ellipse is used to calculate likelihood ratios (Kulldorff, 1998). This technique will search for clusters where more cases than expected under randomness occurred within a particular area and time interval. The method can be used for retrospective as well as for prospective data analysis (Kulldorff, 2001). If no population at risk data are available, such as could be be the case with a surveillance system, the space-time permutation statistic is used purely with case-reporting data (Kulldorff *et al*, 2005). The basic principle is the same as for the space-time scan statistic, except that the permutations are based on randomly shuffling the spatial and temporal attributes of the cases. It controls for purely temporal and purely spatial clusters. The space-time permutation method will be biased by spatially heterogeneous changes in population density over time. The method is sensitive to missing data, and should only be used to replace the space-time scan statistic if no, or only poor quality, population-at-risk data are available. Abatih *et al* (2009) used the space-time scan statistic to identify clustering of ampicillin resistant *Escherichia coli* within 3 islands of Denmark.

26.6.3 Continuous spatial fields

The methods presented above can only be used for discretely defined spatial features. If the spatial variation in a factor is continuous, the focus of the analysis becomes the similarity of measurements depending on their distance from each other. The assumption is being made that the measurements or attribute values are collected at randomly selected point locations. An **empirical semi-variogram** summarises covariation in attribute values between point observations at different distances apart, by calculating half (therefore semi-variogram) of the average of the squared difference between paired values (semi-variance) within each distance range. Before the empirical semi-variogram is determined, a variogram cloud can be constructed which shows the actual semi-variance for all pairs of values. The empirical semi-variogram graphically summarises the spatial dependence for the spatial process as a scatter plot where distance (spatial lag) is presented on the X axis and semi-variance on the Y axis. It is calculated as follows:

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{N(h)} [Z(s_i) - Z(s_j)]^2 \qquad \textit{Eq 26.11}$$

where $N(h)$ are the set of distinct pairs of values separated by distance h, $|N(h)|$ is the number of distinct pairs in $N(h)$.

An important assumption is that the underlying process is stationary. If the dependence varies

between different directions and the spatial process is therefore **anisotropic**, it will be necessary to produce separate semi-variograms for different directions. The typical shape of a semi-variogram for a spatially autocorrelated process has low values for small lag values (see Fig. 26.15), indicating small differences between attribute values. The value at which the function crosses the y-axis is named the **nugget**. The function increases to a value on the y-axis called the **sill** which it reaches at a distance denoted as the **range**. If the function does not reach a sill, this suggests the presence of non-stationarity. If there is no spatial autocorrelation, the semi-variogram plot should be a horizontal line. The empirical semi-variogram can be used to define a theoretical semi-variogram which represents the observed dependence relationship through a theoretical mathematical function. This theoretical function can then be used for kriging interpolation. Co-variograms and correlograms present similar information to the semi-variogram, but the latter is used most commonly (see Example 26.10 and Fig 26.16).

Semi-variograms are often used to assess whether there is spatial dependence in the residuals from a regression model. It should be noted that the resulting semi-variogram should be interpreted cautiously, when based on an analysis of raw residuals generated using ordinary or generalized least squares regression models (Schabenberger & Gotway, 2005).

26.7 MODELLING

Models for spatial data can be used for the purpose of describing spatial variability in a single variable or for explaining variability in an outcome variable with exposure variables which may be subject to spatial effects.

26.7.1 Modeling to describe spatial variation

Models that describe spatial variation in a variable are used to interpolate or predict attribute values at unmeasured locations. The data type and the type of spatial effect present will influence which method is to be used. For example, **trend surface regression** can be used to describe first-order (or large-scale) spatial effects in a continuous-scale variable based on a linear combination of polynomial functions of coordinate locations. This method can be applied to continuous spatial fields as well as area data (Haining, 2003). Second-order (or small-scale) spatial effects for stationary continuous spatial fields are commonly modelled using inverse-distance weighted interpolation or **kriging**. The advantage of kriging is that it allows interpolation including uncertainty estimates of the predicted values. With this method, a system of kriging equations is developed which derive their distance weights from a semi-variogram (discussed above). For each location value to be interpolated, neighbouring known values are used to estimate the local value including its variance. Different types of kriging technique are now available, including probability kriging which is used for binary data. These methods are covered in some detail in Waller and Gotway(2004). Berke (2004) used kriging to generate a smooth disease risk surface from tape worm infection data in foxes which were surveyed in Lower Saxony, Germany, between 1991 and 1997, aggregated at the level of 43 administrative regions. The spatial dependence captured in a kriging model can also be used to model the covariance structure in a random effects model, such as described below. Clements *et al* (2007) present an example of such an application in Bayesian regression modelling.

Continuous-scale data for areas can be modelled using **conditional (CAR)** or **simultaneous autoregressive (SAR)** methods. If the area data are discrete values, auto-logistic , binomial or Poisson models can be used, and the basic principle of the modelling approach is that the local

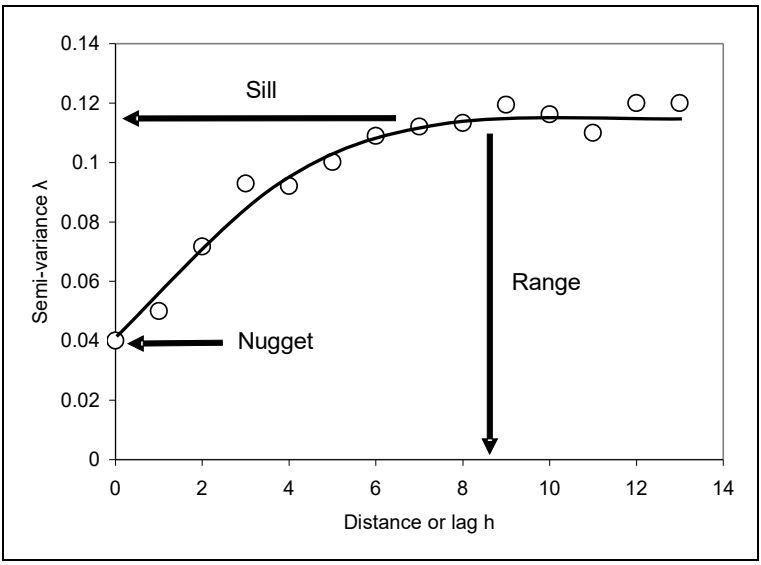


Fig. 26.15 Example of an empirical (circles) and theoretical (curve) semi-variogram

value is conditional on the values in the neighbourhood. This means that the models include fixed effect parameter estimates for local as well as neighbourhood effects. The parameters can represent directional or trend effects and potential interactions. Haining (2003) provides a detailed discussion of the approaches for continuous- and discrete-scale area data.

While the methods described above are used to generate parameters describing the spatial effects for area data, hierarchical modelling aims at obtaining more precise local estimates of spatially varying parameters such as disease risk by making use of the information in neighbouring areas. In this case, spatially structured random effects are used. Examples are presented above under empirical and fully Bayesian smoothing.

26.7.2 Modelling to explain variability (spatial and non-spatial)

In principle, the models described above become explanatory regression models when exposure variables are included (see Example 26.11 and Fig. 26.17). The methodology for the relevant approaches used to deal with different types of dependent and independent variable are described in Chapters 14-24. The key assumptions for these approaches are the independence of the observations which constitute the data that are used to derive the regression relationship and an unchanging relationship between exposure and response variables across space. The presence of spatial autocorrelation in the residuals from a model is an indication that the assumptions have been violated. Spatial dependence may occur in the response variable, in the exposure variables or in both.

Linear regression will produce unbiased effect estimates with correct confidence intervals, if any spatial dependence in the response variable is completely explained by the spatial pattern in the exposure variables (*ie* there is no unmeasured factor which is spatially correlated). If there is still residual spatial dependence, it will often be reflected in spatial autocorrelation in the regression residuals. It should be noted that response variables related to infectious diseases are

Example 26.10 Variogram analysis of AI outbreaks in northern Vietnam
data = Vietnam

For the region in the north of Vietnam, height above sea level was measured at a single location within each district. The current analysis is aimed at summarising the spatial dependence between height measurements at locations different distances apart using semi-variogram analysis, and was conducted using the gstat package of the R software.

The semi-variance values of all pairs of observations are shown in Fig. 26.16a. It suggests that there may be some outliers for smaller distances. Fig. 26.16b presents the semi-variogram for the data. It shows that there is spatial dependence since semi-variance increases with distance between pairs of point observations. As the values do not seem to level out at a particular semi-variance value, it is likely that the data are subject to non-stationarity. Fig. 26.16c represents the results after detrending the values (*ie* removing a possible longer distance trend). The resulting pattern seems more likely to reach an upper maximum value or sill, but the theoretical variogram function does not level out. Fig. 26.16d shows a set of 4 directional variograms based on the detrended values. While the patterns vary slightly, it cannot be concluded that there is strong directionality. But there is a difference in the variability of semi-variance estimates which are lowest for the 135 degree angle.

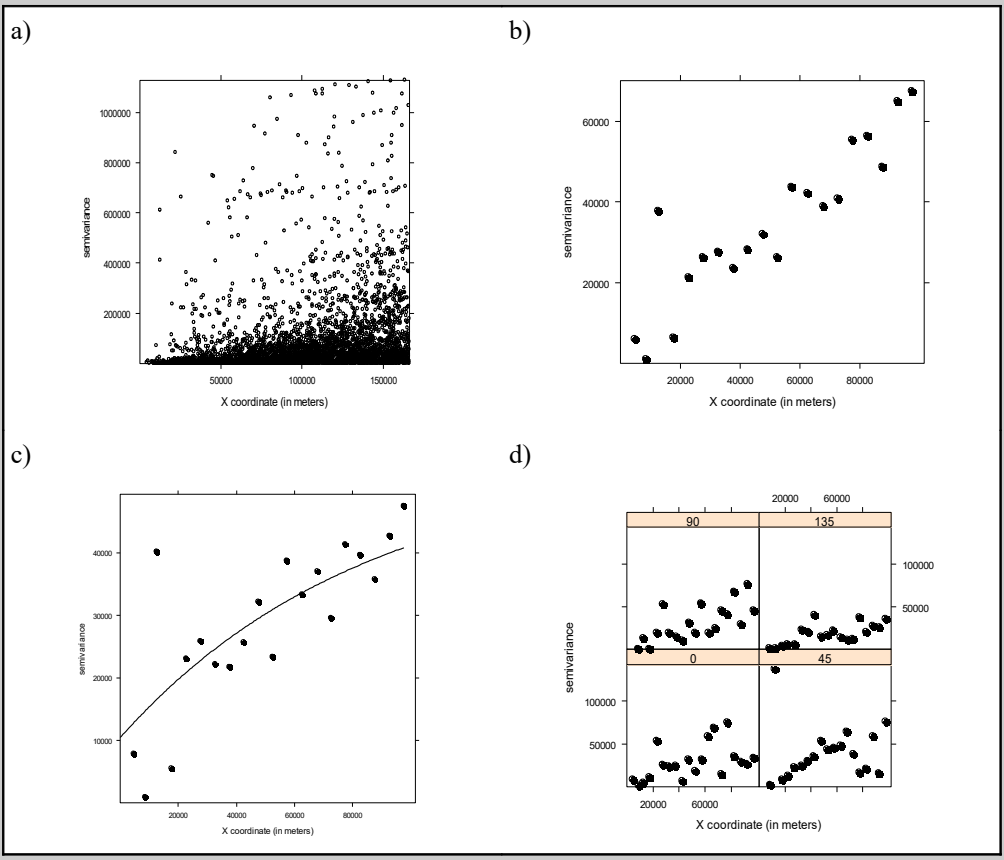


Fig. 26.16 Variogram analysis for height above sea level in a northern region of Vietnam

often subject to inherent spatial dependence which is unlikely to ever be completely explained by exposure factors. First-order (or large-scale) spatial effects could be modelled using coordinate locations as covariates, and second-order (or small-scale effects) using the covariance structure. A spatial covariance structure can be parameterised using parametric functions of distance which are informed by kriging models (also called geostatistical approach). If the data relate to areas (aggregated at centroid locations for each area), this approach would inappropriately assume that observations have been made at locations between neighbouring area centroids. In this case, spatially autoregressive models are more suitable, which reflect the similarity in response values in the covariance structure. This is done through SAR or CAR autoregression models. The dependence can be expressed using spatial moving average or spatial lag models (the latter might be lagged on response or specific exposure variables). Both, Waller & Gotway (2004) and Schabenberger & Gotway (2005) provide examples of applications of linear regression modelling techniques for spatially autocorrelated data. In linear mixed modelling, the observed exposure factors are represented as fixed effects and the unobserved ones as random effects. The latter means that instead of aggregating all unexplained variation into the general error term, a random effect is used to separate out any structured error component. In the presence of spatial variation which cannot be captured by fixed effect exposure factors, the random effect can have a spatial covariance structure. The associated estimates are obtained using restricted maximum likelihood estimation. A detailed discussion of the approach is provided in Littell R *et al (*2006) and Schabenberger & Gotway (2005).

With binary, binomial or count response variables, the generalised linear modelling (GLM) approach is used, and second-order (or small-scale) spatial variation can be modelled through the covariance structure implemented as a random effect in a generalised linear mixed models (GLMM). The estimation algorithms which need to be used include quasi- and pseudo-likelihood, as well as Bayesian methods. Within a Bayesian analysis framework, Besag *et al* (1991) recommended the use of a spatial trend term, a spatially correlated heterogeneity term and an uncorrelated term (together called convolution prior). A detailed description of developing various types of fully **Bayesian spatial regression model** using R and WinBUGS is provided in (Lawson, 2009). Among the non-Bayesian approaches, **generalised additive mixed models** (GAM) and **geographically weighted regression** (GWR) methods can also be applied to take account of spatial dependence. The principle of GWR is that data are weighted according to distance from any point in the dataset by using a spatial kernel (Fotheringham *et al*, 2002). It could be considered a disadvantage of this method that the mathematical form for the kernel and its bandwidth need to be specified, and particularly the latter will have strong influence on the model's coefficient estimates. GAM apply non-linear functions to the predictor variables, and they can also include random effects (Wood, 2006). As with GWR, somewhat subjective choices have to be made in relation to the mathematical form of the non-linear relationships which will then often strongly influence the analysis result.

The choice for a particular model is informed by a variety of methods, including assessment of the residuals and likelihood-based information criteria such as Akaike's information criterion (AIC) or Bayes' information criterion (BIC). These are calculated on the basis of both the deviance and the number of variables in the model and usually the smaller the value, the 'better' the model fit. Schabenberger and Gotway (2005) describe various scenarios where these values need to be used cautiously when comparing different models — in particular when applied to multilevel models. Neither of them is appropriate for comparing Bayesian models, and the deviance information criterion (DIC) should be applied instead (Banerjee *et al*, 2004).

The statistical significance of the coefficients is assessed using the Wald test or likelihood ratio tests. The residuals need to be examined for the presence of spatial variation. Such an assessment can be performed by visual examination of mapped residuals. A quantitative analysis is usually done by calculating Moran's *I* or generating an empirical semi-variogram from the residuals. The results should be interpreted cautiously, as it is not possible to clearly determine whether the particular result was caused by the underlying spatial process or is an artifact resulting from the regression estimation process (Schabenberger & Gotway, 2005).

Example 26.11 Spatial regression modeling of AI outbreaks in northern Vietnam
data = Vietnam

The same data on AI outbreaks in a northern region of Vietnam was used for the current analysis. The response variable is the proportion of communes that experienced AI outbreaks between 2004 and 2006 for each district within the study region. Based on the descriptive risk maps presented in Pfeiffer *et al*, (2007) it appears that disease risk is not randomly distributed which was later confirmed statistically by applying various spatial cluster analysis methods, including the spatial scan statistic and Moran's *I*. The objective now is to identify factors that might be associated with AI outbreak risk and potentially explain the spatial heterogeneity of disease risk. The factor considered in this analysis is the density of ducks at district level. The analyses were conducted using the packages nlme, lme4 in the R software, and the statistical software WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/) for the fully Bayesian modelling.

(*continued on next page*)

Example 26.11 (*continued*)

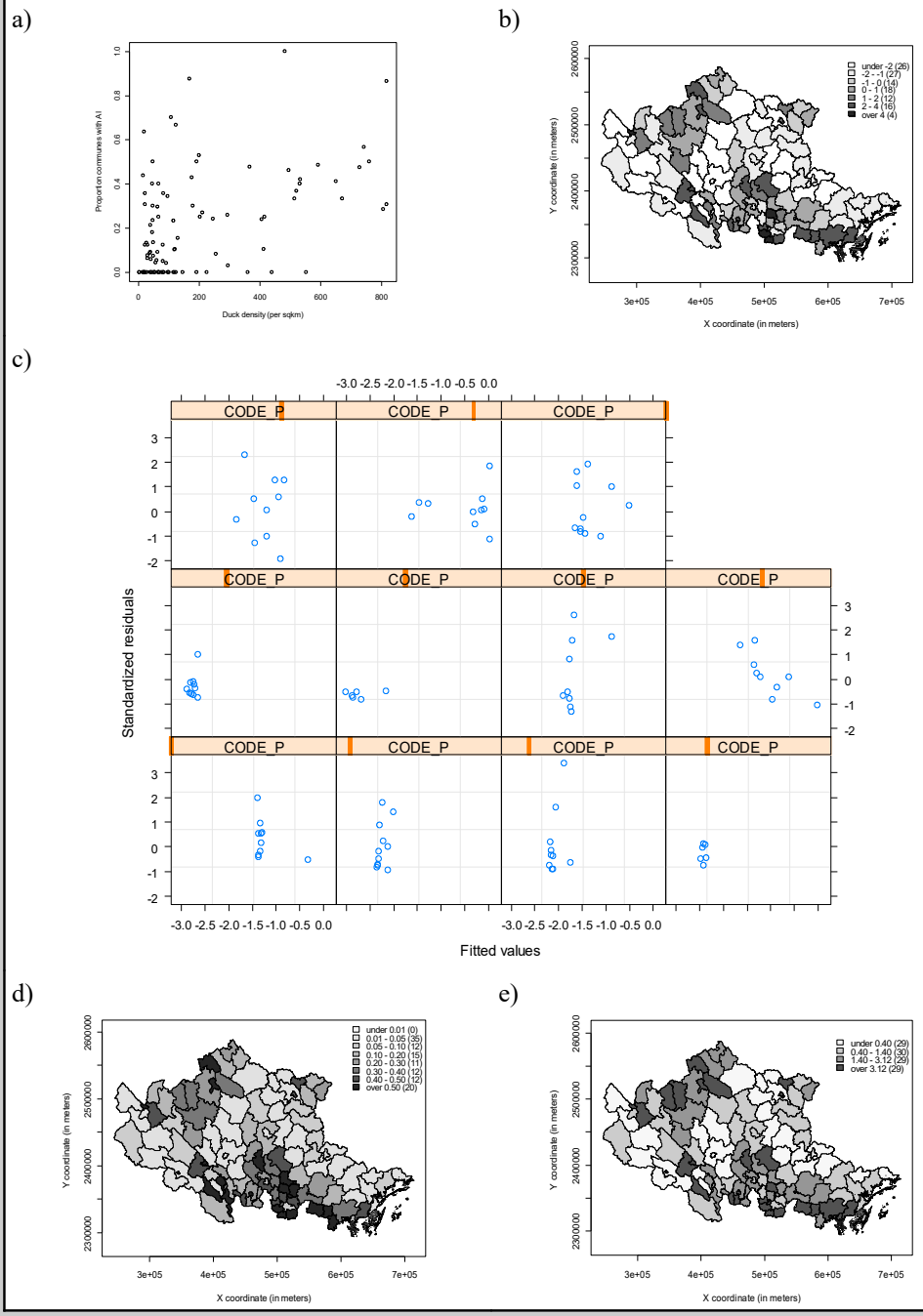


Fig. 26.17 Spatial regression modelling of the relationship between the proportion of communes with AI outbreaks and duck density at district level within a northern region of Vietnam

(*continued on next page*)

Example 26.11 (*continued*)

The relationship between the two variables AI outbreak proportion and duck density is presented in the scatter plot in Fig. 26.17a. As a first step, a generalized linear model analysis with a binomial link function was conducted. The linearity of the relationship between the two variables was tested as follows. The continuous-scale duck-density variable was re-expressed as a new variable recoded into 4 equal-interval groups which were then included in separate models as a continuous and as a factor variable and model fit was compared with the likelihood ratio statistic. Including duck density as a factor variable did not improve the model fit and it was therefore concluded that it could be modelled as a linear effect, using the original continuous-scale variable.

The results of this analysis indicate that duck density is positively associated with AI proportion (Model 1: β=0.0021, P<0.001). Controlling for overdispersion resulted in only a slight change of the P-value for the coefficient of the fixed effect variable duck density. The residuals of this model appear to be clustered in space as shown in Fig. 26.17b. The visual impression is confirmed by applying Moran's *I* to the residuals (I=0.25, p<0.001). Inclusion of province as a random effect in the model results in a coefficient for duck density of 0.0018 (Model 2: p<0.001). The random effect was then further adapted to the data by using a spatial covariance structure which was estimated using the penalized quasi-likelihood method. A Gaussian distance weighted covariance structure was used, and the resulting coefficient for duck density was 0.002 (Model 3: p<0.001).

The residuals for this model are clustered by province as shown in the trellis plot in Fig. 26.17c. Finally, a fully Bayesian approach resulted in the risk map presented in Fig. 26.17d. In this analysis, a convolution prior involving a spatial and a non-spatial random effect was applied as described by Besag *et al* (1991). A nearest neighbour matrix was used to represent the spatial dependence in the data, and the regression effect for duck density became 0.0022 (Model 4: Bayesian 95% CI 0.0005-0.004).

The simulation was based on two simulation chains, each consisting of 20,000 iterations in addition to a burn-in phase of 5,000 iterations. The potential scale reduction factor Rhat (indicating potential further reduction in the confidence interval given an infinite number of iterations) was between 1 and 1.1 for all parameters indicating good mixing and convergence of the chains, with only one parameter having an effective sample size of less than 100 (Gelman A & Hill J, 2007). The residuals are presented in Fig. 26.7e, and show clustering of high values along the northern and southern borders of the study area. Comparing the results from these 4 models indicates that duck density at district level is associated with an increased risk of communes within districts experiencing AI outbreaks, and that given the similarity of the coefficients and their significance levels both the effect estimate and its variance are unaffected by spatial dependence.

REFERENCES

Abatih EN, Ersbøll AK, Lo Fo Wong DMA, Emborg HD. Space-time clustering of ampicillin resistant Escherichia coli isolated from Danish pigs at slaughter between 1997 and 2005. Prev Vet Med. 2009; 89: 90-101.

Anon. Guidelines for investigating clusters of health events. Morb Mort Weekly Rep. 1990; 39: 1-16.

Anselin L. Local indicators of spatial association - LISA Geographical Analysis. 1995; 27: 93-115.

Assuncao R, Reis E. A new proposal to adjust Moran's *I* for population density. Stat Med. 1999; 18: 2147-62.

Bailey T, Gatrell A. Interactive Spatial Data Analysis. Longman Group; Harlow, Essex, England. 1995.

Banerjee S, Carlin B, Gelfand A. Hierarchical Modeling and Analysis of Spatial Data. Chapman & Hall/CRC; Boca Raton, Florida, USA. 2004.

Berke O. Exploratory disease mapping: kriging the spatial risk function from regional count data. Int. J.Health Geogr. 2004; 3: 18.

Besag, J., York, J., Mollié, A. Bayesian image restoration with two applications in spatial statistics. Annals Inst Stat and Math. 1991; 43, 1-59.

Bithell J. An application of density estimation to geographical epidemiology. Stat Med. 1990; 9: 691-701.

Bivand R, Pebesma E, Gomez-Rubio V. Applied Spatial Data Analysis with R. Springer; New York, USA. 2008.

Clements A, Pfeiffer D, Martin V, Pittliglio C, Best N, Thiongane Y. Spatial risk assessment of Rift Valley fever in Senegal. Vector Borne Zoonotic Dis. 2007; 7: 203-16.

Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. J Royal Stat Soc B. 1990; 52: 73-104.

de Smith M, Goodchild M, Longley P. Geospatial Analysis. Matador; Leicester: 2007.

Diggle P. Overview of statistical methods for disease mapping and its relationship to cluster detection. In: Spatial epidemiology - Methods and applications. Elliott P, Wakefield J, Best N, Briggs D (Eds). Oxford University Press; Oxford. 2000.

Diggle P, Chetwynd A. Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics. 1991; 47: 1155-63.

Diggle P, Chetwynd A, Haggkvist R, Morris S. Second-order analysis of space-time clustering. Stat Meth Med Res. 1995; 4: 124-36.

Diggle P, Gomez-Rubio V, Brown P, Chetwynd A, Gooding S. Second-order analysis of inhomogeneous spatial point processes using case-control data. Biometrics. 2007; 63: 550-7.

Fotheringham A, Brunsdon C, Charlton M. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons; Chichester, UK. 2002.

Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press; Cambridge, UK. 2007.

Goldberg T, Hahn E, Weigel R, Scherba G. Genetic, geographical and temporal variation of porcine reproductive and respiratory syndrome virus in Illinois. J Gen.Virol. 2000; 81: 171-9.

Haining R. Spatial Data Analysis - Theory and Practice. Cambridge University Press; Cambridge, UK. 2003.

Jacquez G. A k nearest neighbour test for space-time interaction. Stat Med. 1996; 15: 1935-49.

Kelsall J, Diggle P. Non-parametric estimation of spatial variation in relative risk. Stat Med.

1995; 14: 2335-42.

Knox E. The detection of space-time interaction. Applied Statistics. 1964; 13: 25-9.

Kulldorff M. Statistical Methods for Spatial Epidemiology: Tests for Randomness. In: GIS and Health. Gatrell A, Löytönen M (Eds). Taylor and Francis; London. 1998.

Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic J Royal Stat Soc A. 2001; 164: 61-72.

Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast Cancer Clusters in the Northeast United States: A Geographic Analysis. Am J Epidemiol. 1997; 146: 161-70.

Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS.Med. 2005; 2: e59.

Lawson A. Disease map reconstruction. Stat Med. 2001; 20: 2183-204.

Lawson A. Bayesian Disease Mapping. Chapman & Hall / CRC; Boca Raton, Florida, USA. 2009.

Lawson A, Biggeri A, Dreassi E. Edge effects in disease mapping. In: Disease mapping and risk assessment in public health. Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel J, Bertollini R (Eds). John Wiley and Sons; Chichester. 1999.

Littell R, Milliken G, Stroup W, Wolfinger R, Schabenberger O. SAS System for Mixed Models. SAS Institute; Cary, North Carolina. 2006.

Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Research. 1967; 27: 209-20.

Monmonier M, de Blij H. How to Lie with Maps. Chicago: University of Chicago Press; Chicago, USA. 1996.

Morris S, Wakefield J. Assessment of Disease Risk in Relation to a Pre-specified Source. In: Spatial epidemiology - Methods and applications. Elliott P, Wakefield J, Best N, Briggs D (Eds). Oxford University Press; Oxford. 2000.

Norström M, Skjerve E, Jarp J. Risk factors for epidemic respiratory disease in Norwegian cattle herds. Prev Vet Med. 2000; 44: 87-96.

Oden N. Adjusting Moran's I for population density. Stat Med. 1995; 14: 17-26.

Pfeiffer D, Robinson T, Stevenson M, Stevens K, Rogers D, Clements A. Spatial Analysis in Epidemiology. Oxford University Press; Oxford. 2008.

Pfeiffer DU, Minh PQ, Martin V, Epprecht M, Otte MJ. An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. The Vet J. 2007; 174: 302-9.

Porphyre T, McKenzie J, Stevenson M. A descriptive spatial analysis of bovine tuberculosis in intensively controlled cattle farms in New Zealand. Vet Res. 2007; 38: 465-79.

Ripley B. Spatial point pattern analysis in ecology. In: Developments in Numerical Ecology. Legendre P (Eds). Springer Verlag; Berlin. 1987.

Schabenberger O, Gotway C. Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC; Boca Raton, Florida. 2005.

Scott D. Multivariate Density Estimation: Theory, Practice and Visualisation. John Wiley & Sons; New York. 1992.

Wakefield J, Kelsall J, Morris S. Clustering, cluster detection and spatial variation in risk. In: Spatial epidemiology - Methods and applications. Elliott P, Wakefield J, Best N, Briggs D (Eds). Oxford University Press; Oxford. 2000.

Waller L, Gotway C. Applied Spatial Statistics for Public Health Data. John Wiley & Sons; Hoboken, New Jersey. 2004.

Walter S. The analysis of regional patterns in health data. I. Distributional considerations. Am J Epidemiol. 1992; 136: 730-41.

Wand M, Jones M. Kernel Smoothing. Chapman & Hall / CRC; Boca Raton, Florida. 1995.

Ward M, Carpenter T. Analysis of time-space clustering in veterinary epidemiology. Prev Vet Med. 2000; 43: 225-37.

Wood S. Generalized additive models: An introduction with R. Chapman & Hall/CRC; Boca Raton, Florida. 2006.