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

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

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

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

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

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

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

CHAPTER 6

Spatial variation in risk

6.1 Introduction

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

6.2 Smoothing based on kernel functions

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

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

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

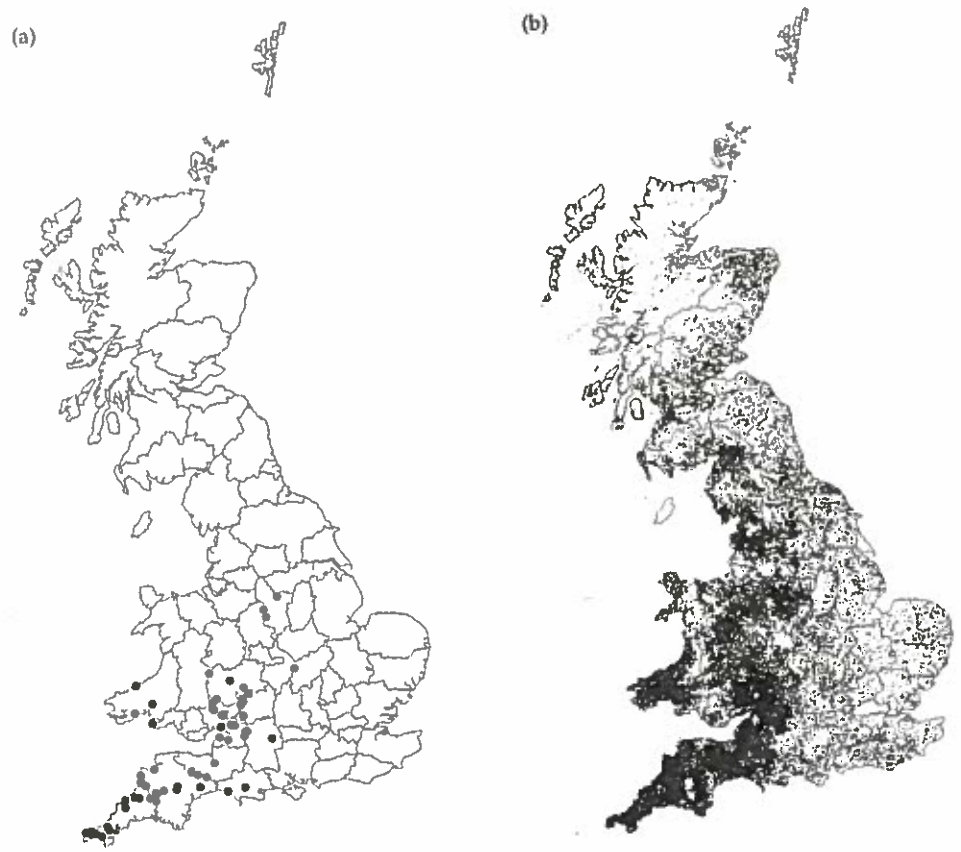


Figure 6.1 Point maps showing (a) the distribution of herds for which TB-positive cattle were identified at slaughter in 1996, and (b) locations of herds TB-tested in 1996, in Great Britain.

standard neighbourhood function in GIS (Bonham-Carter 1994). With epidemiological spatial data, they can be applied to point as well as aggregated data represented through a centroid. Talbot et al. (2000) describe the use of filters with fixed geographical size as well as with constant population size to generate smoothed map representations of disease ratios. They demonstrate that a filter with constant population size retains adequate spatial resolution in high density areas while at the same time producing stable rate estimates in low density areas.

Kernel density estimation is a special case of distance weighted map smoothing where a bivariate probability density function is applied to determine the intensity of a spatial point process (Bailey and Gatrell 1995). The following equation is used to perform the calculations:

$$\hat{\lambda}_\tau(s) = \frac{1}{\delta_\tau(s)} \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right) \qquad (6.1)$$

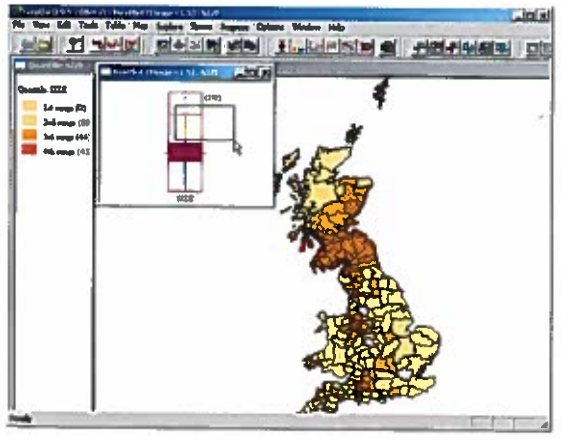
where k() represents the chosen bivariate probability density function, $\tau > 0$ is the bandwidth, s is the point on which a disc of radius τ is centred, s_i are the points within the disc's area, and τ represents an edge correction factor. The difference between density and intensity should be noted, in that the former integrates to one and estimates are interpreted as probabilities, whereas the latter estimates the mean number of events per unit area and does not integrate to one. One can be converted into the other by a multiplicative constant.

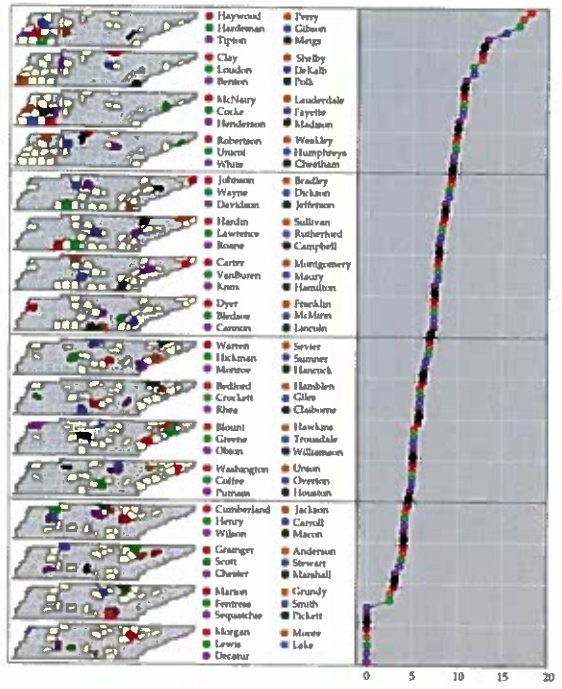
The resulting smoothed density surface is influenced by the choice of probability density function,

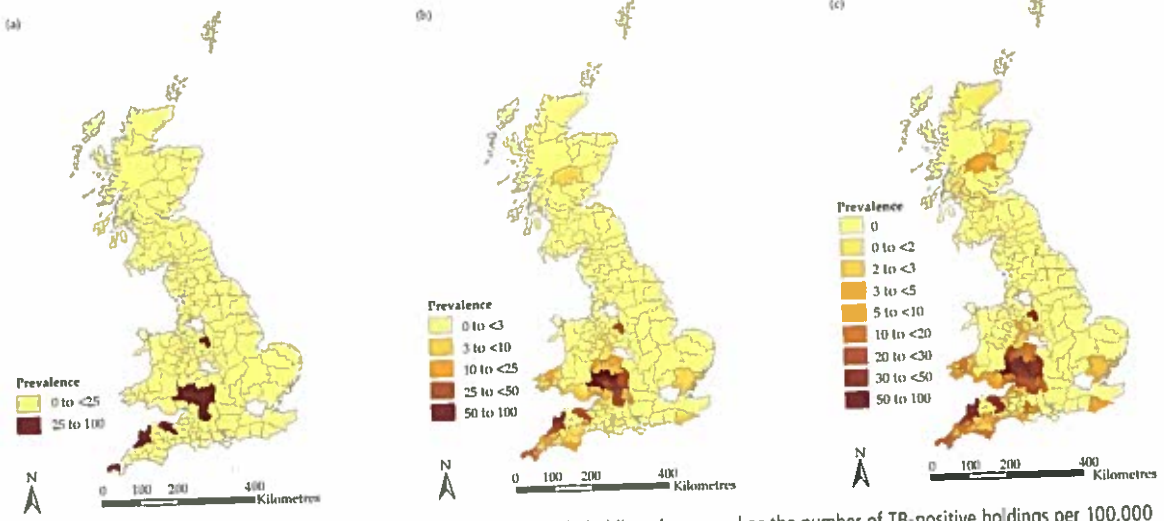
False colour composite of Fourier-processed air temperature variables for Great Britain. The average value (the 'zero-order' component) is displayed in red, the phase of the first-order component is displayed in green, and the amplitude of the first-order component is displayed in blue.

Dynamic exploratory spatial data analysis (ESDA) using the GeoDa software. The screen shot shows two windows: (main) a choropleth map of area-level herd size, and (inset) a box-and-whisker plot showing the distributional features of the same data.

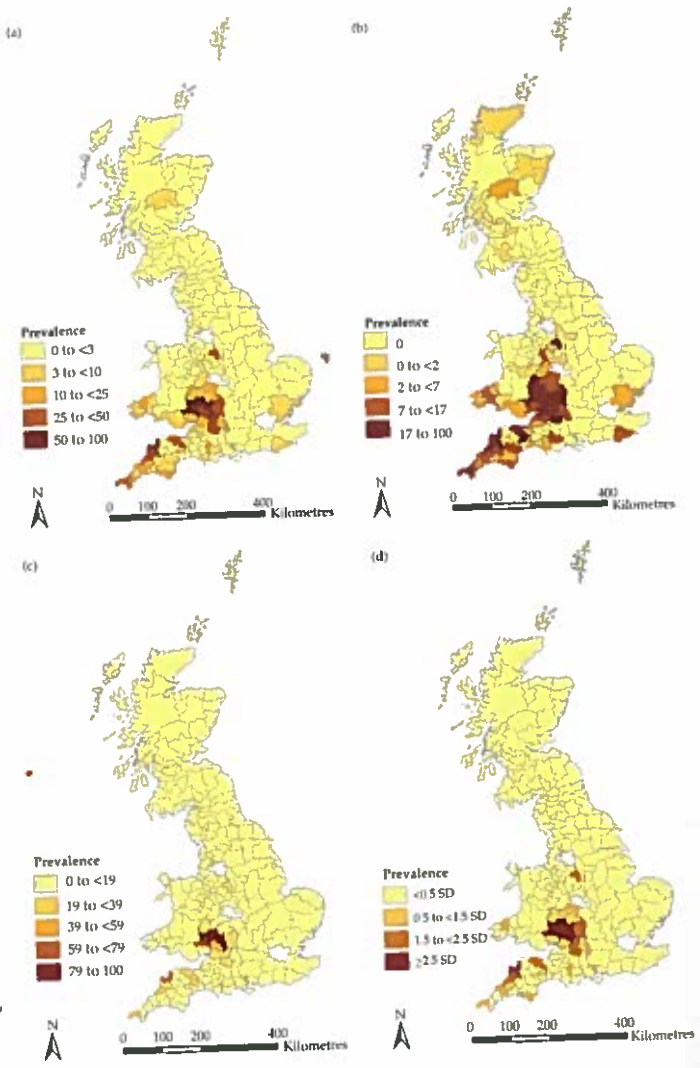
Conditioned choropleth map of median infant mortality rate per 1000 births in the state of Tennessee, USA between 1992 and 1997. Reproduced from Carr et al. (2000) with permission from *Statistics in Medicine*.



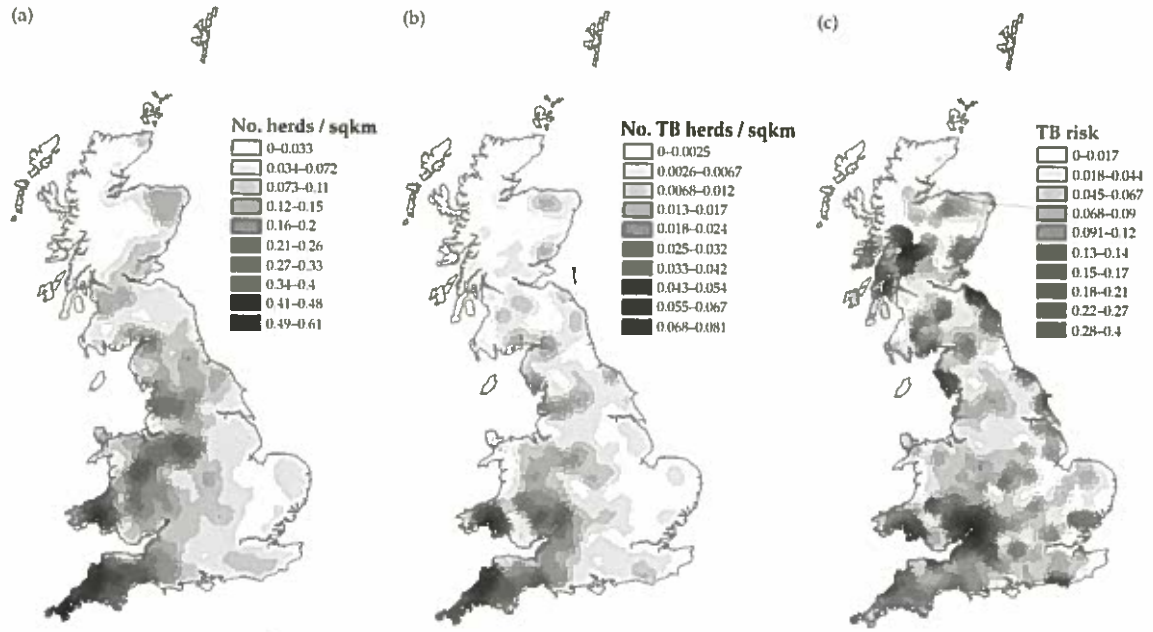




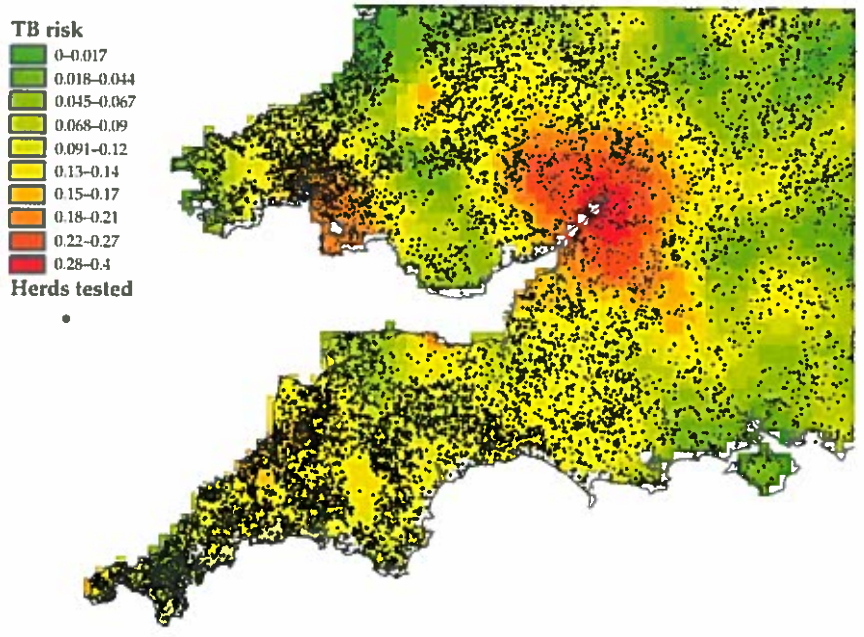
Choropleth maps showing area-level prevalence of TB on British cattle holdings (expressed as the number of TB-positive holdings per 100,000 holdings) for 1985–1997 plotted using (a) two, (b) five, and (c) nine cutpoints.



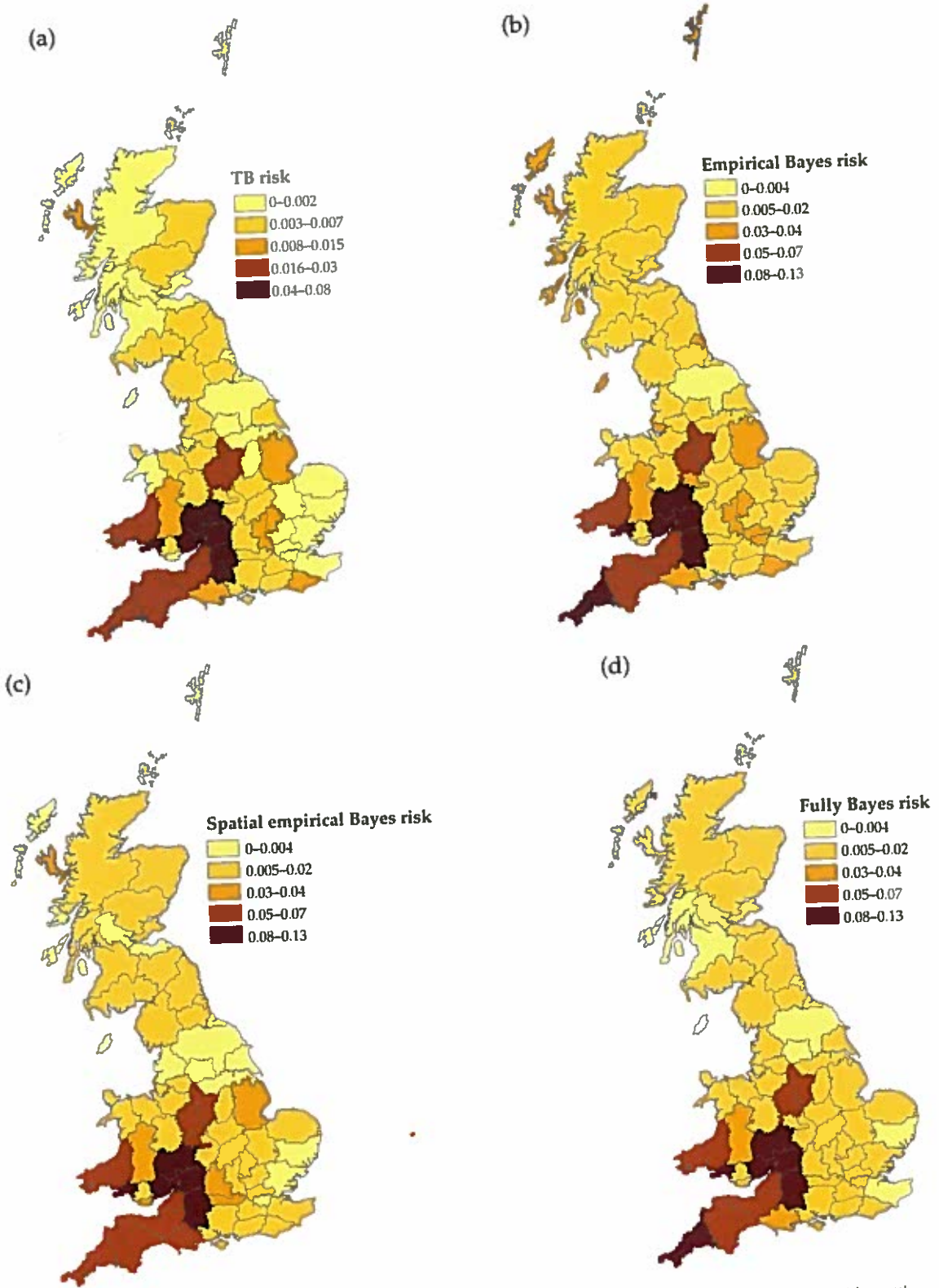
Choropleth maps showing area-level prevalence of TB on British cattle holdings (expressed as the number of TB-positive holdings per 100,000 holdings) for 1985–1997 using (a) natural breaks, (b) quintile breaks, (c) equal interval breaks, and (d) standard deviation classifications.



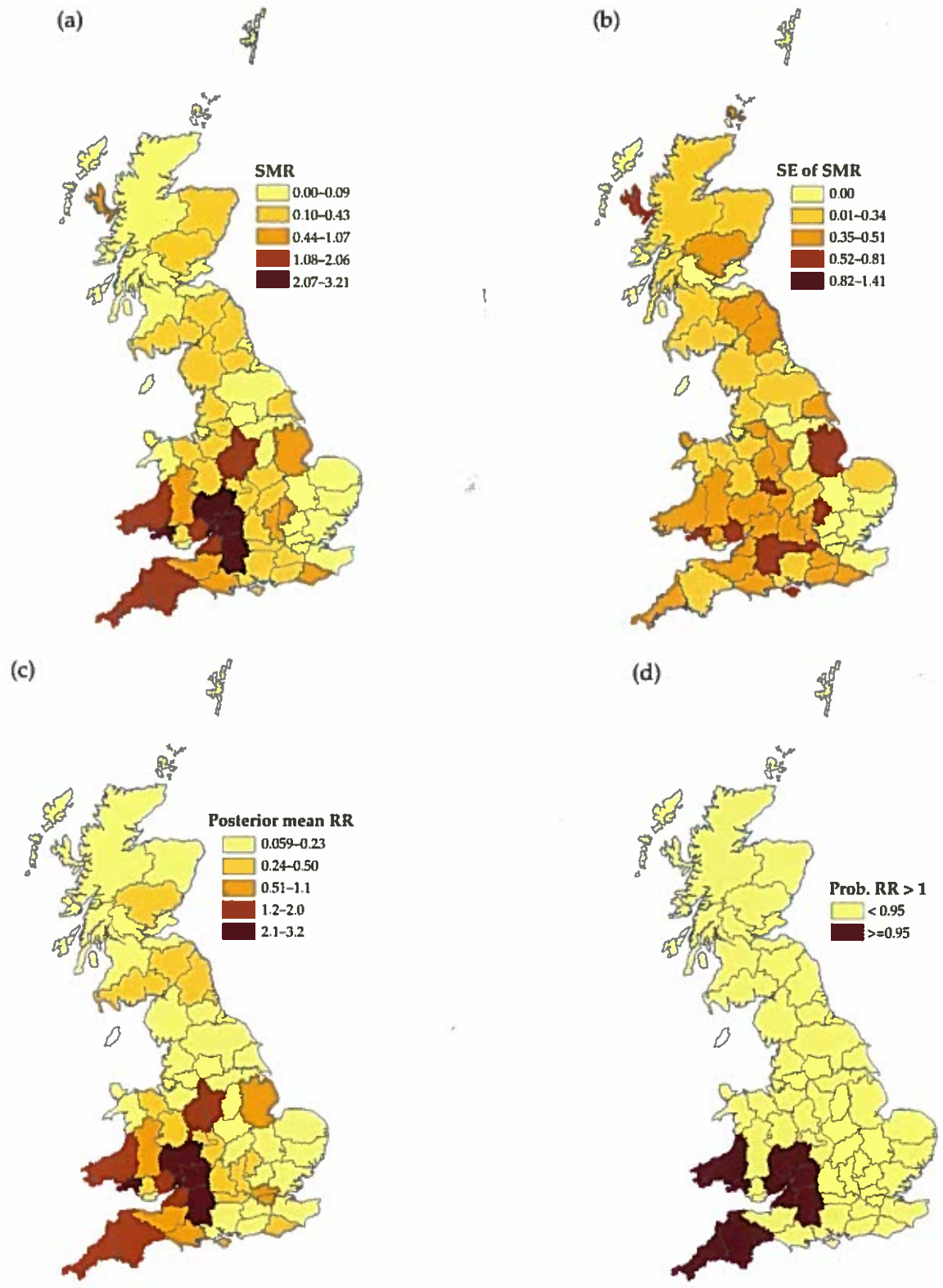
Kernel smoothed map representations of cattle herd densities in Great Britain. (a) Kernel smoothed intensity of **all** cattle herds tested using routine TB herd testing in 1996 (5 km grid cells, 30 km bandwidth), (b) kernel smoothed intensity of cattle herds tested **positive** during routine TB herd testing in 1996 (5 km grid cells, 30 km bandwidth), and (c) ratio of kernel smoothed intensities of test positive herds and all cattle herds tested in 1996.



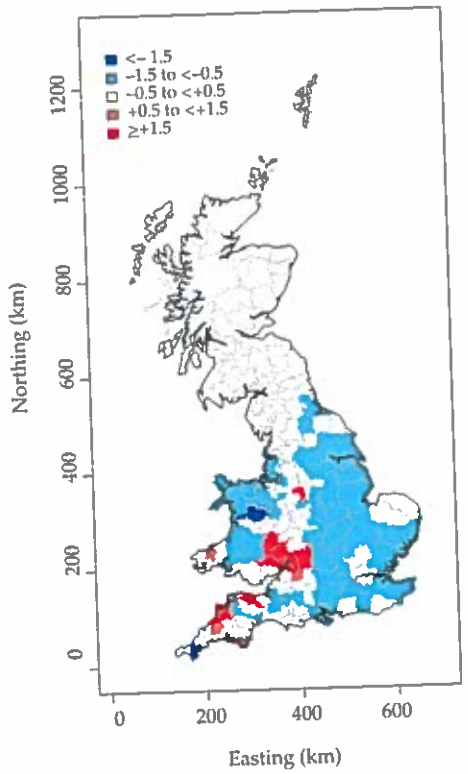
Kernel density ratio surface for the southwest of England showing the risk of cattle herds testing positive for TB during 1996 (the locations of all herds tested are shown as black dots).



Choropleth maps of TB-test results for cattle herds in Great Britian in 1999 aggregated by county. (a) Prevalence of TB-test positive cattle herds, (b) empirical Bayes risk of TB-test positive herds, (c) spatial empirical Bayes risk of TB-test positive herds, (d) fully Bayes risk of TB-test positive herds.

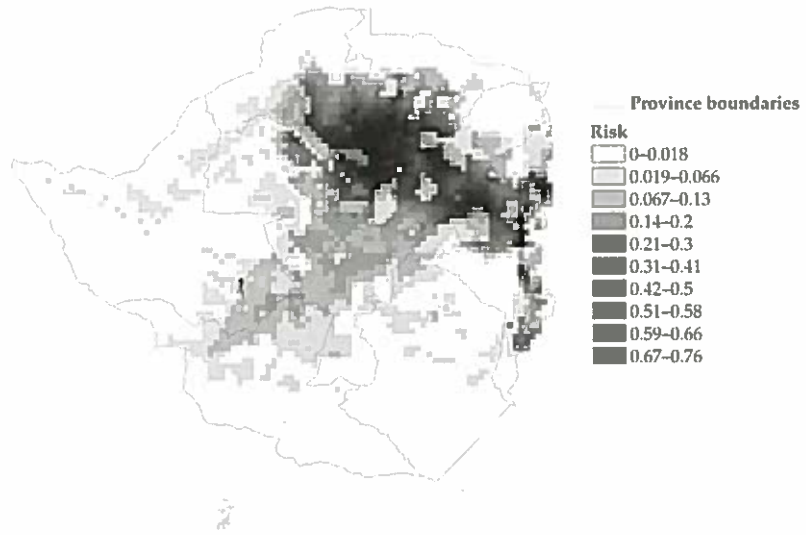


Choropeth maps of standardized morbidity ratios (SMR) and Bayesian relative risk estimates for TB herd test results for cattle in Great Britain in 1999 aggregated by county. (a) SMR comparing TB reactor risk between counties, (b) standard error of SMR, (c) fully Bayesian estimates of relative risk of TB-test reactor herds, and (d) statistically significant fully Bayesian relative risks of reactor herds.

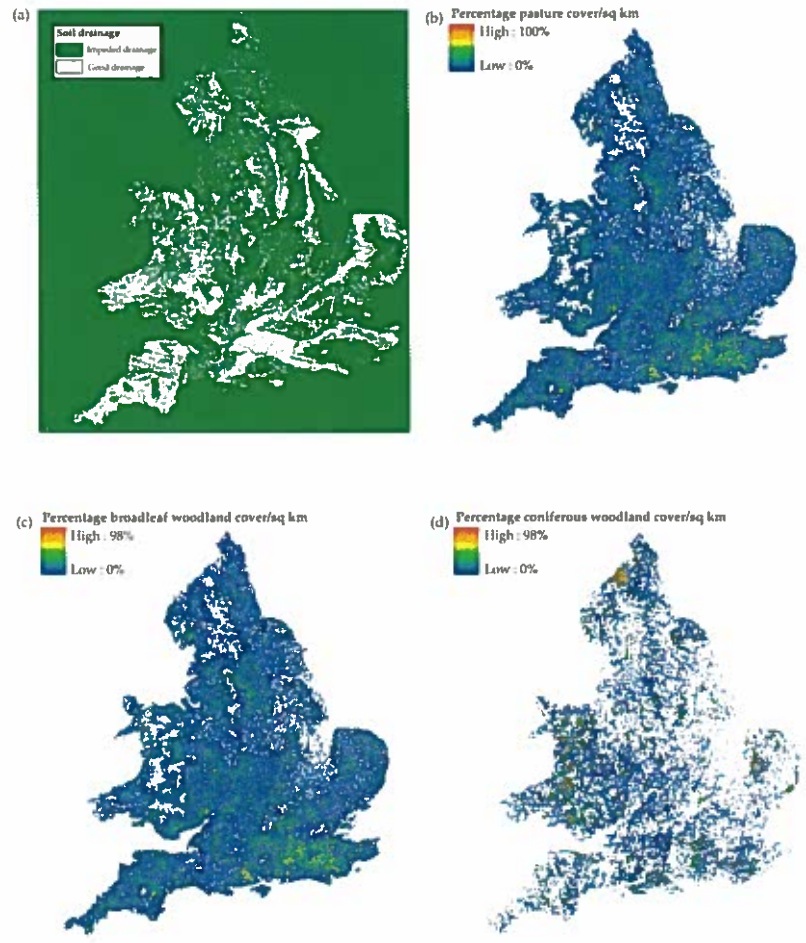


Choropleth map showing the spatial distribution of the area-level standardized residuals from the fixed-effects Poisson model of area-level TB risk in Great Britain shown in Table 7.2.

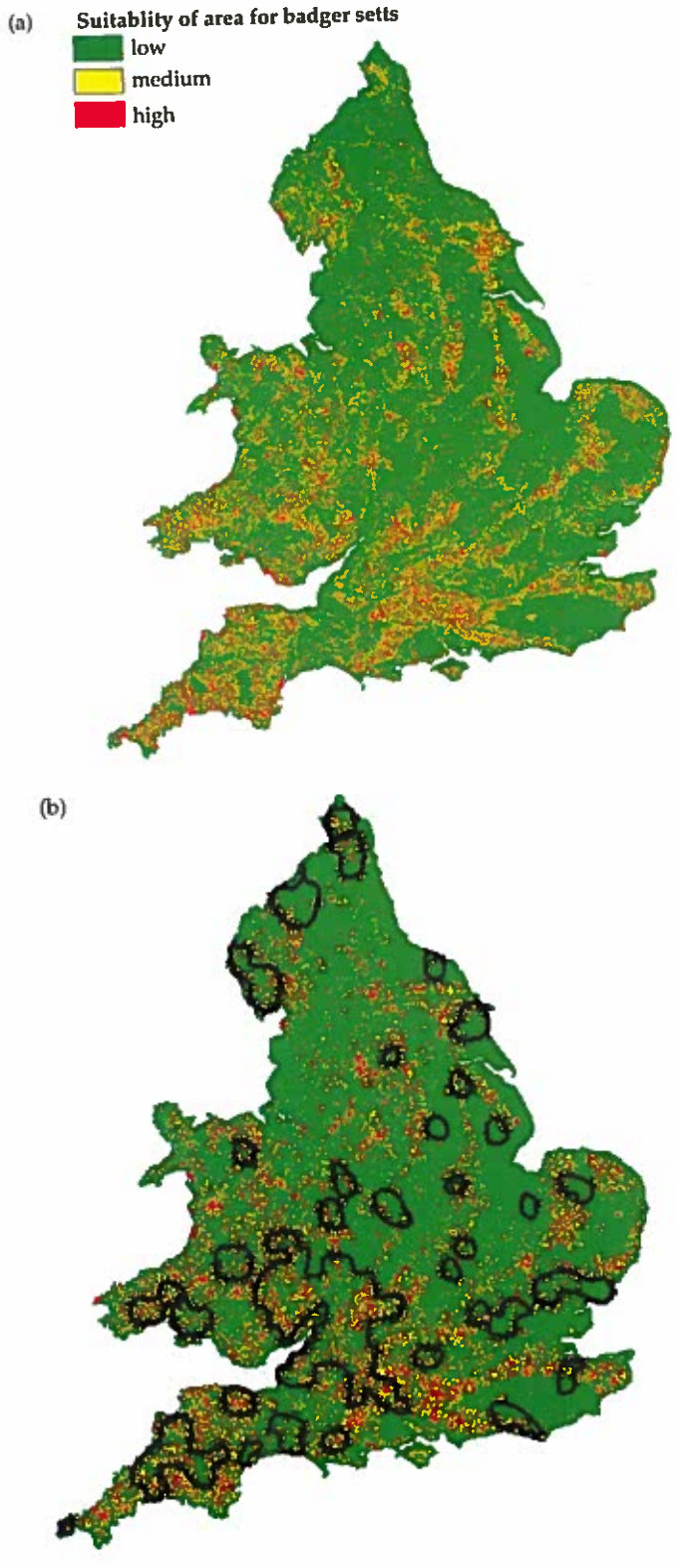
Mean (of 100) posterior probability risk map of bovine tuberculosis (TB) cases in Great Britain predicted by discriminant analytical models using step-wise inclusion of 10 variables to maximize the area under the curve (AUC) at each step. Five presence and five absence clusters were used for each bootstrap sample of 300 presence and 300 absence points sampled at random, with replacements, from a population of 482 presence and 3000 absence pixels (a presence pixel is a 0.01 degree pixel with at least one TB-positive farm in 1997; an absence pixel has farms, but no TB-positive ones in the same period). Posterior probability is on the green to red colour scale shown in the legend beneath the figure. The positive pixels are indicated by the black dots.



Risk map for theileriosis occurrence risk in Zimbabwe (Pfeiffer et al. 1997). Map showing probability of occurrence of outbreaks due to infection with *Theileria parva*. Reproduced from Pfeiffer et al. (1997)



Criterion maps used in the MCDA model. (a) Boolean map showing soils with good or impeded drainage, (b) a continuous scale map of pasture in England and Wales (percentage cover/km^2), (c) a continuous scale map of broadleaf woodland in England and Wales (percentage cover/km^2), and (d) a continuous scale map of coniferous woodland in England and Wales (percentage cover/km^2).



(a) Map identifying areas in England and Wales of low, medium, and high suitabililty for the construction of badger setts, and (b) the suitability map with an overlay of the 1996 TB high-risk areas in England and Wales.

bandwidth τ, and size of the grid cells for each of which an individual density estimate is calculated. The mathematical function used to define the kernel specifies the pattern for down-weighting the influence of points further away from the point locations for which intensity is to be estimated. Different mathematical functions can be used but they are considered to have less influence on the estimation than the bandwidth, and the Gaussian kernel is therefore often used for the sake of computational simplicity (Waller and Gotway 2004). The latter is the width of the kernel function and therefore determines the distance over which points will be included in the calculation, and the larger it is the smoother the resulting surface. The bandwidth τ can be obtained using mathematical calculations or by subjective choice. Haining (2003) emphasizes that it is not about obtaining an optimal bandwidth based on theoretical considerations, but rather about generating a spatially smoothed surface that reveals insights into the underlying data. Characteristics of the biological process to be studied could therefore be used to guide the choice, and it is always recommended to explore surfaces generated by different bandwidth estimates. Diggle (2003) recommends that, rather than an automatic procedure, a plot of the mean square error of the non-parametric intensity estimator against different values of τ be used to inform the choice of bandwidth. Adaptive bandwidth selection methods vary the local bandwidth during the estimation process so that a minimum number of observations are included (Bailey and Gatrell 1995). Further details and discussions in relation to bandwidth selection methods can be found in Scott (1992) and Wand and Jones (1995). The kernel density calculation method should also take account of edge effects (also called spatial censoring) to minimize the bias of estimations close to the boundary of the study area (Lawson et al. 1999b). This can be achieved, for example, by adjusting the area used in the calculation process according to the overlap of the circular area defined by the bandwidth and the study region (Diggle 2003). When using kernel density estimation functions built into GIS software products such as ArcGIS (ESRI, Redlands, California, USA) the bivariate Gaussian kernel is typically used, a default algorithm is applied to calculate the bandwidth and no correction is made for edge effects. ArcGIS calculates the bandwidth as the minimum dimension (x or y) of the extent of the point theme divided by 30. More sophisticated methods for applying kernel smoothing can be implemented using the statistical functions developed for the software R. Stevenson et al. (2000) conduct a descriptive spatial analysis of BSE occurrence in Great Britain using kernel density estimation based on a Gaussian kernel and a fixed bandwidth of 30 km, estimated using the normal optimal method described by Bowman and Azzalini (1997) and implemented in the SM library for R by the same authors. Choice of the grid cell size for which the kernel-smoothed intensity estimates are to be produced needs to be considered from a presentational, biological, and numerical perspective. Since the objective is to achieve a balance between producing a smooth surface while still describing the relevant spatial variation in intensity, grid cells that are too small needlessly increase the data storage required, and if too large may hide relevant patterns. It is therefore sensible to make the grid cells larger than the geographical extent of the biological unit of interest. For example, if the density of farms is to be estimated, the grid cell size should be larger than the area covered by an average farm. If the grid cell size is too small, the intensity estimates will be very small numbers that may be more difficult to interpret. As with the bandwidth, the algorithm implemented in the software may use default settings to define grid cell size that are often not 'optimal' for the particular dataset. Fig. 6.2a shows kernel smoothed presentations of the density of all British cattle herds TB-tested in 1996 and Fig. 6.2b shows those that tested positive. Both estimations were made using a grid cell size of 5 km and a bandwidth of 30 km. This choice was based on the following considerations: firstly, from a national decision-making perspective, variability between areas above 25 km^2 was considered to be the focus of the analysis and secondly, local herd density estimates should not be influenced by densities at a distance of more than 30 km whereas bandwidth estimates below that would result in too much variability to allow meaningful policy-relevant interpretation. These

considerations were based to a large degree on subjective choices. Fig. 6.2b shows the spatial distribution of TB-test positive herds. In order to identify unusually high occurrences of these herds, the spatial distribution of the underlying population at risk (Fig. 6.2a) also needs to be considered. This is difficult to perform through visual comparison of the two maps.

As in the example presented above, the distribution of the population at risk is usually spatially heterogeneous and therefore, to be able to detect spatial clusters of unusually high or low case intensity, it is necessary to examine the spatial pattern of the proportionality between intensity of disease cases and non-cases, or population at risk (sum of cases and non-cases). This method has been called extraction mapping (Lawson and Williams 1993). If data on the intensity of the population at risk cannot be obtained, the spatial distribution of another disease can be used instead, as long as it is not subject to the same aetiological processes and there is no differential reporting bias (Lawson 2001b). 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 and Diggle 1995b). One requirement of such ratio calculations is that the denominator should not take on zero values, which means that the bivariate Gaussian kernel with its non-zero tail should be used. The optimal bandwidths chosen for producing the individual intensity surfaces may not be appropriate for generating the ratio surface. The resulting ratio surface is also much more sensitive to different choices of bandwidth than the individual surfaces. There is some debate as to whether the numerator and denominator in this calculation should be generated using the same or different bandwidths (Bithell 1990; Bailey and Gatrell 1995; Diggle 2000). The preference seems to be that it should be the same, even if the sample sizes differ between the two populations (Kelsall and Diggle 1995a). Schabenberger and Gotway (2005) describe a visual exploratory approach for choosing the appropriate bandwidth and size of grid cells, based mainly on balancing the resolution and stability of the estimates in the context of their biological interpretation. The statistical precision of the ratio estimates can be quantified using Monte Carlo methods (Kelsall and Diggle 1995b). Kernel regression methods have been proposed by Kelsall and Diggle (1998) as an alternative to the density ratio method. An advantage of this approach is that covariates can be included in the analysis.

Fig. 6.2c shows the ratio of the intensity surfaces of TB-test positive and all cattle herds tested as part of routine herd testing in Great Britain in 1996. It illustrates that it can be difficult to obtain stable risk estimates if intensities vary substantially across an area, such as with cattle-herd density in Great Britain where, particularly in Scotland, high risks were calculated based on very small numbers of herds at risk. This could be prevented by increasing the grid cell size and/or bandwidth of the kernel-smoothed maps but at the expense of spatial differentiation in the main areas of interest, namely Wales and the southwest of England. Fig. 6.3 shows the southwest of England where the risk estimates provide an interesting insight into the spatial pattern of positive routine herd testing. Cornwall and Devon have a lower risk of herds testing positive than the areas within the counties of Gwent, Gloucestershire, Avon, and Hereford and Worcester. There are also small clusters of increased risk in the counties of Dyfed, Powys, and Somerset.

6.3 Smoothing based on Bayesian models

Statistics used in the spatial representation of aggregated (e.g. province or district-level) disease risk data include relative risk, and standardized mortality/morbidity ratios (SMR). The SMR is the classic statistic used in representing spatial patterns of disease distribution. It standardizes the data by re-expressing them as the ratio between the observed number of cases and the number that would have been expected in a standard population. At whatever level of spatial aggregation (i), for a specified time period:

$$SMR_i = y_i / e_i \qquad (6.2)$$

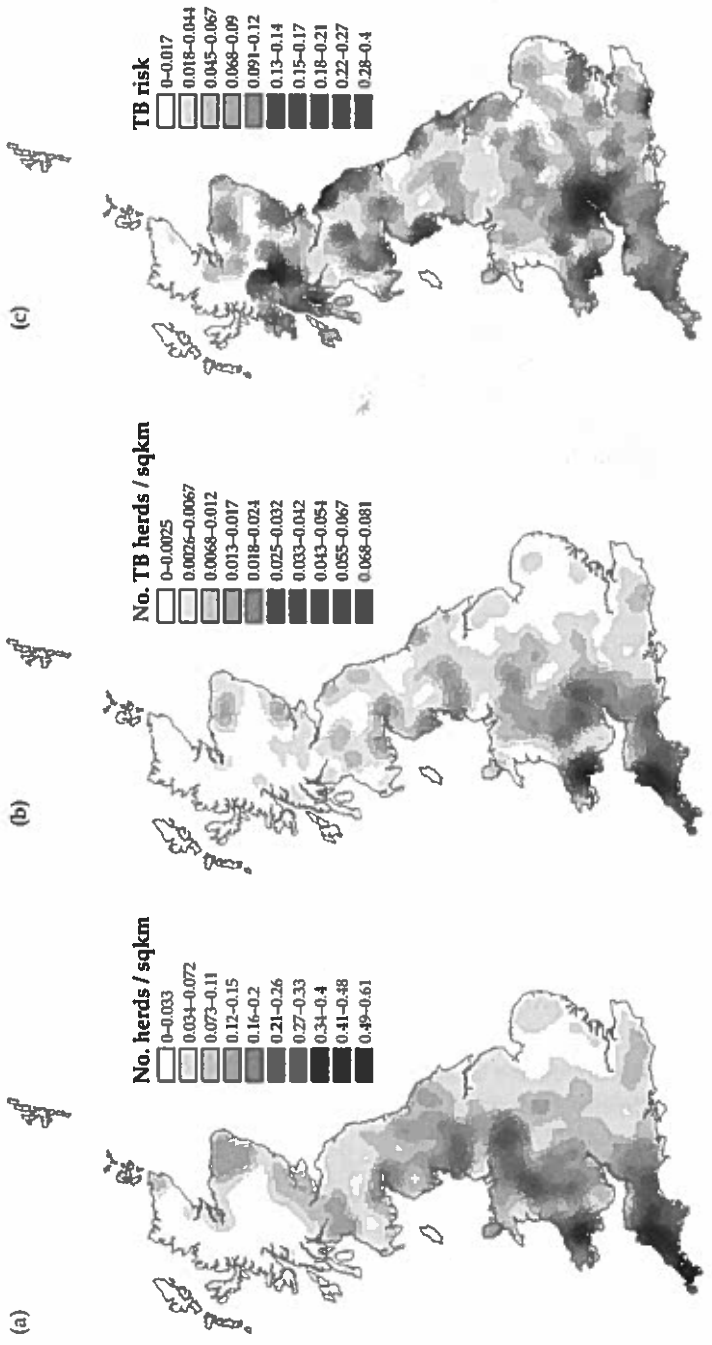


Figure 6.2 Kernel smoothed map representations of cattle herd densities in Great Britain. (a) Kernel smoothed intensity of **all** cattle herds tested using routine TB herd testing in 1996 (5 km grid cells, 30 km bandwidth), (b) kernel smoothed intensity of cattle herds tested **positive** during routine TB herd testing in 1996 (5 km grid cells, 30 km bandwidth), and (c) ratio of kernel smoothed intensities of test positive herds and all cattle herds tested in 1996 (see colour plate).

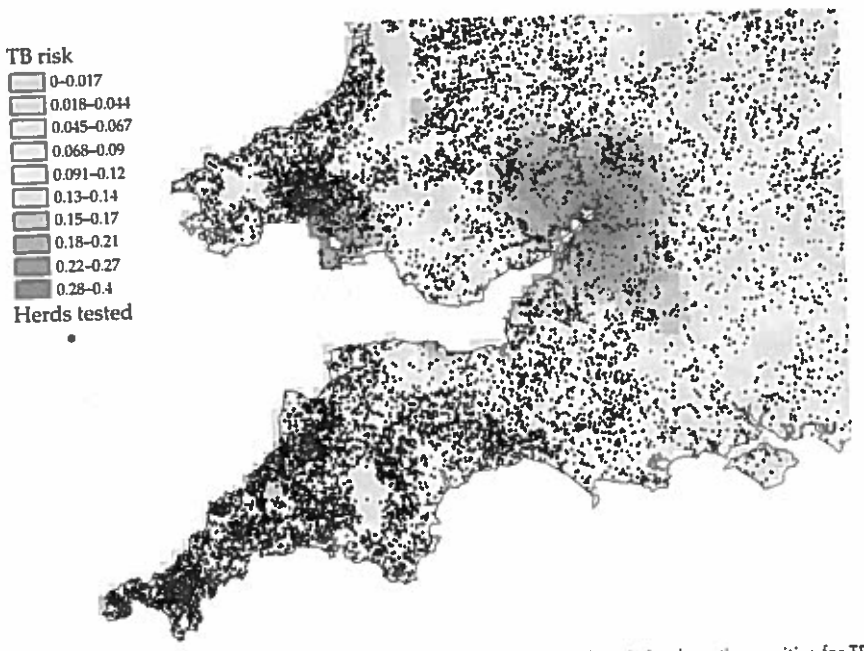


Figure 6.3 Kernel density ratio surface for the southwest of England showing the risk of cattle herds testing positive for TB during 1996 (the locations of all herds tested are shown as black dots) (see colour plate).

where y_i is the observed number of cases in area i, and the expected number of cases, e_i, is given by:

$$e_i = \frac{\sum y_i}{\sum n_i} \cdot n_i \tag{6.3}$$

where n_i is the observed population at risk in area i. National level disease risk or rate, for example, can be used to calculate the expected number of cases for each local area (Lawson and Williams 2001). The disadvantage of such SMR maps is that they tend to be dominated by areas with small numbers of observations, or if they focus on the statistical significance of the estimates, they tend to be dominated by areas with large populations, but potentially less important disease occurrence (Leyland and Davies 2005).

Clayton and Kaldor (1987) recognize that empirical Bayes methods can be used to generate maps that make use of the neighbourhood relationships in the data to produce statistically more precise risk estimates. Fully Bayesian estimation methods have been made possible through the development of the Markov Chain Monte Carlo (MCMC) algorithm, based on the Gibbs sampler (a special case of the Metropolis–Hastings algorithm; Gelman et al. 1995).

The basic principle of Bayesian methods is that uncertain data can be strengthened by combining them with prior information. In the case of empirical Bayes estimation of spatially-varying disease risk, posterior risk can be estimated from a weighted combination of the local risk (also called the likelihood) and the risk in surrounding areas, the latter representing the prior information (Clayton and Bernardinelli 1992; Wakefield et al. 2000). The relative weights for the two components depend on the sample size in the local area. If the local population size is large it will receive a strong weighting in the calculation process. If it is relatively small, its weighting is reduced and the derived estimate will tend towards the prior. The risk can be smoothed using a prior based on the global mean or on summarized data from the neighbouring areas. The smoothed risk is more stable and has higher specificity. It is recognized that Bayesian estimation always represents a trade-off between improved precision and the introduction of bias (Best et al. 2005).

The following formula is used to perform empirical Bayes calculations of disease risk (Bailey and Gatrell 1995):

$$\hat{\theta}_i = w_i r_i + (1 - w_i)\gamma_i \tag{6.4}$$

where θ_i is the empirical Bayes estimate for area i, w_i are the weights applied to the local and neighbourhood estimates, r_i is the local risk in area i and γ_i is the mean of the prior, and r_i is the local risk in area i.

$$r_i = \frac{y_i}{n_i} \tag{6.5}$$

where y_i is the number of cases and n_i the population at risk in area i. The weights, w_i, in (6.4) are estimated as:

$$w_i = \frac{\phi_i}{(\phi_i + \gamma_i / n_i)} \tag{6.6}$$

where Φ_i is the variance of the prior, γ_i is the mean of the prior, and n_i the population at risk in area i.

The empirical Bayes prior specified above is purely spatial and is defined using the hyperparameters γ_i and Φ_i. The estimation is made by simplified posterior distributions through likelihood or integral approximations (Lawson et al. 2003). In fully Bayesian estimation, the hyperparameters have hyperprior distributions resulting in the estimates for each area better approximating the true value. Empirical estimates conversely, are inexact and tend to oversmooth towards the global mean. In addition, fully Bayesian methods generate credibility intervals for each local estimate. It is also possible to have a hierarchical or multilevel structure to the prior. Besag et al. (1991) developed the convolution prior which consists of an unstructured and a structured spatial component. Parameter distributions are estimated using MCMC spatial modelling. This method involves taking very large numbers of samples from the posterior distributions of model parameters. Sampling based on the Gibbs sampler is performed using a Markov Chain where successive samples are dependent on each other. A key issue with MCMC modelling is to decide when to finish sampling (i.e. when posterior distributions of the various parameters have converged). Detailed introductions to Bayesian spatial analysis can be found in Lawson et al. (2003), Banerjee et al. (2004), and Congdon (2003). Chapter 7 explores Bayesian approaches in the context of identifying risk factors associated with disease.

Fig. 6.4a is a choropleth map showing the proportion of TB-positive cattle herds per county in Great Britain in 1999. While this type of map is easy to interpret it has the disadvantage that the size of the districts and the position of their boundaries is typically a reflection of administrative requirements rather than of the spatial distribution of epidemiological factors. Fig. 6.4b shows a map generated using empirical Bayes estimation with the prior being the national level herd infection prevalence. Fig. 6.4c employed a neighbourhood empirical Bayes approach where the local estimate has shrunk towards the mean prevalence estimates of the neighbouring counties. The fully Bayesian approach based on a convolution prior was used to generate the map in Fig. 6.4d (Besag et al. 1991). Examining the four different maps shows that some of the counties have quite different risk levels, when considering prior information. While most of the high risk areas (around Gloucestershire) remain consistent, Cornwall is only in the highest risk category when using empirical non-spatial or fully Bayes estimation. An alternative approach for assessing these data would be to focus on the differences between the counties and the national average. Fig. 6.5a presents SMR estimates for TB-infected herds aggregated at county level. As with the prevalence map (Fig. 6.4a), this is largely a function of sample size and it is therefore appropriate to accompany the map with a presentation of the variability of the estimates, such as in Fig. 6.5b. Fig. 6.5c shows the posterior mean relative risks for each county based on fully Bayesian estimation. The proportion of estimates with a relative risk of greater than one is presented in Fig. 6.5d. It identifies three groupings of counties with an elevated risk of TB-infected herds relative to the rest of the country.

6.4 Spatial interpolation

Spatially continuous surfaces for attribute values, such as risk of disease vector presence or environmental exposure variables, are often produced from data collected at a sample of spatially distributed sites. These could be, for example, traps for capturing the *Culicoides* midges that transmit the bluetongue virus, or epidemiological field surveys

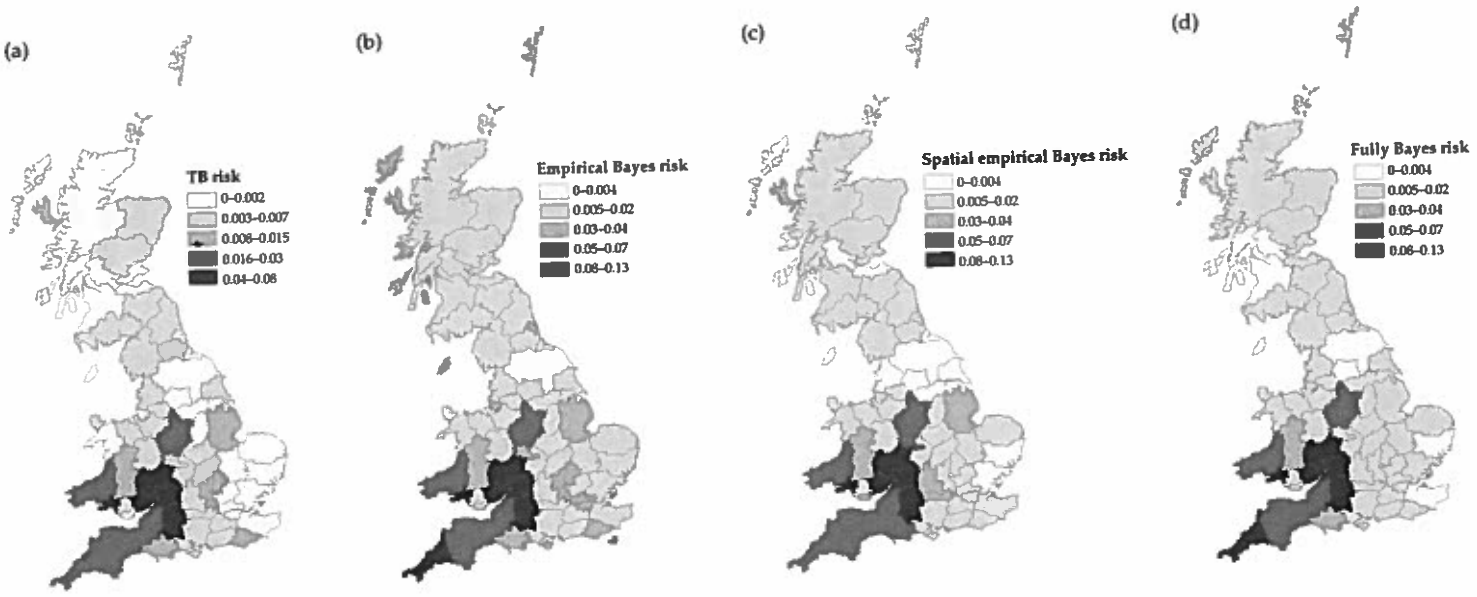


Figure 6.4 Choropleth maps of TB-test results for cattle herds in Great Britian in 1999 aggregated by county. (a) Prevalence of TB-test positive cattle herds, (b) empirical Bayes risk of TB-test positive herds, (c) spatial empirical Bayes risk of TB-test positive herds, (d) fully Bayes risk of TB-test positive herds (see colour plate).

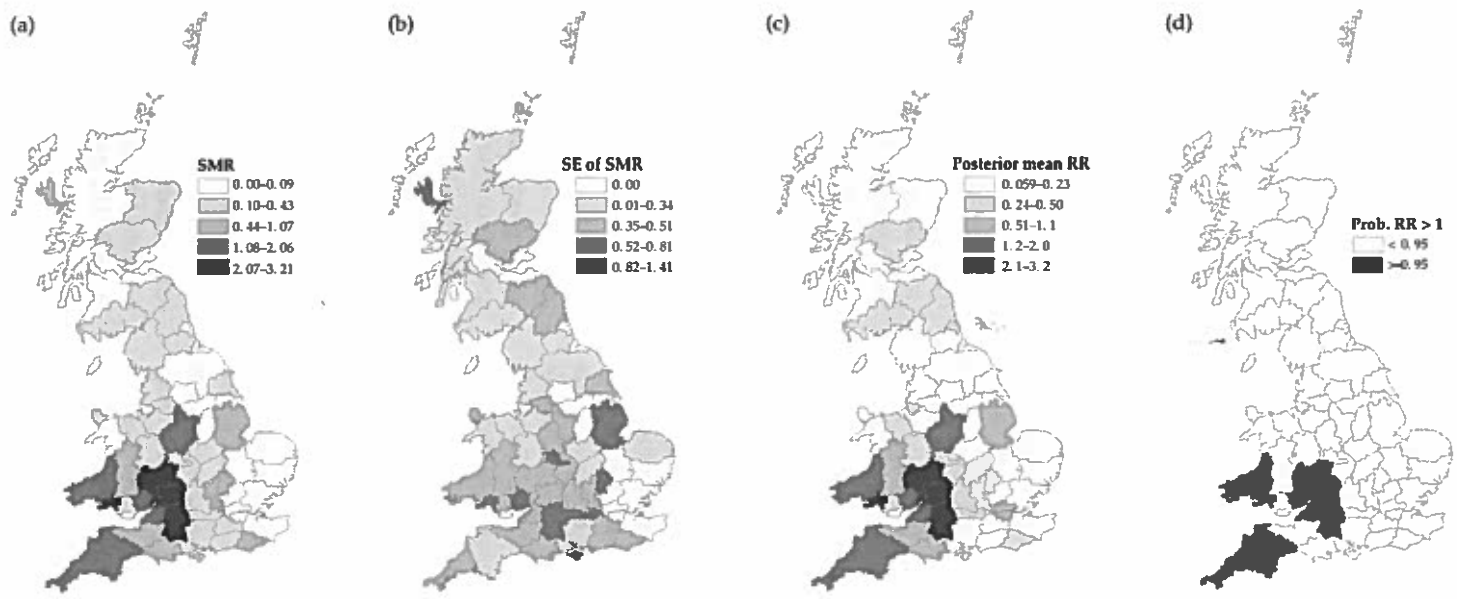


Figure 6.5 Choropeth maps of standardized morbidity ratios (SMR) and Bayesian relative risk estimates for TB herd test results for cattle in Great Britain in 1999 aggregated by county. (a) SMR comparing TB reactor risk between counties, (b) standard error of SMR, (c) fully Bayesian estimates of relative risk of TB-test reactor herds, and (d) statistically significant fully Bayesian relative risks of reactor herds (see colour plate).

that determine disease prevalence or incidence in a sample of herds or communities. Many environmental exposure variables describing, for example, water or air quality are measured at particular monitoring stations, but represent a spatially continuous phenomenon. One specific characteristic of this type of data is that its locations are fixed and known, and not random as is the case with the cluster analyses described in Chapters 4 and 5. Several methods can be used to convert these point location data into surface information. The simplest is to assign to each sampling point a polygon including all the hypothetically possible point locations that are closest to the sampling point. Dirichlet or Thiessen polygons fulfil this purpose but have the disadvantage that the resulting surface is discontinuous. The inverse distance weighting (IDW) interpolation method generates a continuous surface by calculating missing values from distance-weighted sample measurements. This method can be biased by the presence of local clusters. Kriging has the advantage in that it allows the errors of the imputed values to be estimated (Haining 2003). It can do this by not only accounting for the actual sample measurement values and their distance to the location to be predicted, but by also incorporating a mathematical model of the spatial dependence amongst sample measurements. Kriging was originally developed to describe continuous-scale outcome variables, such as the concentration of particular metals in the soil. Its use for non-Gaussian-type variables has been explored by several authors (Carrat and Valleron 1992; Webster et al. 1994; Diggle and Tawn 1998; Berke 2004; Graham et al. 2005; Goovaerts 2006). Stationarity is an important assumption of kriging (Bailey and Gatrell 1995; Graham et al. 2005), which means that the spatial correlation structure should be constant across locations. Typically however, this is not the case with data on disease counts, rates, and risks (Gotway and Wolfinger 2003), but the same authors conclude that, despite non-stationarity, mis-specified variogram functions, and assumptions of linear models for rate and count data, kriging estimates are relatively unbiased. The validity of using data aggregated to an area, such as in the case of disease frequency estimates per county, for producing point estimates using kriging has been confirmed by Kyriakidis (2004) and discussed further by Goovaerts (2006).

The basic model for interpolating a surface is based on the following equation (Waller and Gotway 2004):

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i) \qquad (6.7)$$

where $Z(s_i)$ is the measured value at the i^{th} location, λ_i is the weight attributed to the measured value at the i^{th} location, and s_0 is the prediction location.

This formula is used for both the IDW and kriging interpolation procedures, but they use different methods to calculate λ_i. With IDW it depends only on the distance to the prediction location whereas with kriging it is determined by the semivariogram, the distance to the prediction location and the spatial relationships between measurements around the prediction location. The empirical semivariogram shows the spatial dependence in the variable of interest as a scatterplot. Distance (spatial lag) is presented on the x- and semivariance on the y-axis. The semivariance is calculated as follows:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} \left[Z(s_i) - Z(s_j) \right]^2 \qquad (6.8)$$

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

The semivariogram can also be expressed as a covariance function but use of the former is preferred in spatial analysis. Schabenberger and Gotway (2005) provide a more detailed discussion of the relationship. The formula above describes a stationary and isotropic empirical semivariogram. If a spatial process is anisotropic, that is the spatial dependence varies with direction, different semivariograms can be estimated for each direction. Plotting the semi-variance produces a curve typically rising from a point on the y-axis (the nugget), to a maximum semi-variance value or 'sill' within a certain lag distance on the x-axis (the range) (see Fig. 6.6). The nugget describes the spatially uncorrelated variation or noise in the data. The larger this value the less spatial dependence there is amongst the attribute values and the less useful kriging interpolation will be. To provide a useful representation

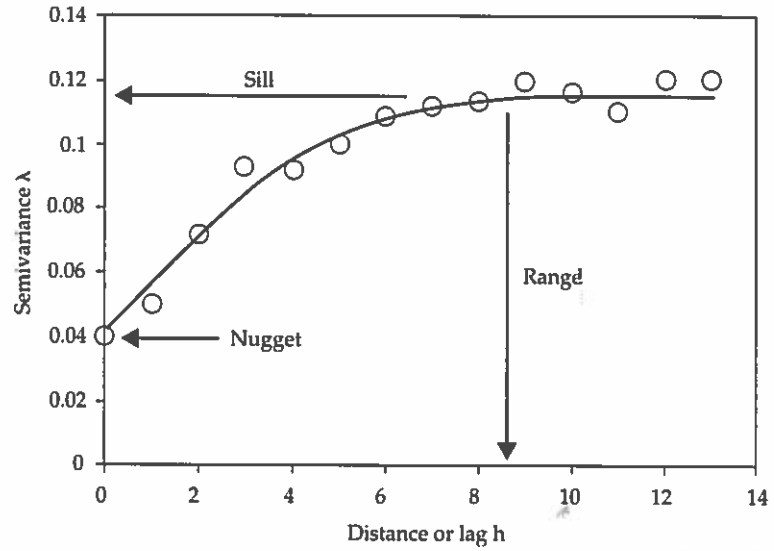


Figure 6.6 An empirical (circles) and theoretical (curve) semi-variogram

of the correlation structure, semi-variance should be calculated for about 10 to 20 lags with a maximum lag distance of about half the maximum separation distance between points (Waller and Gotway 2004; Schabenberger and Gotway 2005). (See Chapter 4 for an explanation of the difference between a semivariogram and a correlogram.)

Apart from providing information for kriging, the empirical semivariogram can be used as a generic tool for visually assessing the presence of spatial dependence or autocorrelation in a continuous process. This can also be useful for testing the assumption of independence for the residuals generated by a regression model. The empirical semivariogram cannot be used directly for kriging. Instead, a mathematical function has to be fitted to the curve, resulting in a theoretical semivariogram function that can then be used in the kriging process. The function can be derived from the variogram using statistical fitting algorithms or by visual assessment of the fit of standard mathematical functions. A flat shape of the resulting theoretical variogram function would suggest absence of spatial dependence. A function with, for example, an exponential shape with values increasing with distance reflects the presence of spatial dependence. Graham et al. (2005) use the shape of the semivariogram to draw conclusions about the relative importance of spatial dependence and local herd effects on *Salmonella enterica* herd antibody levels.

Universal kriging can be used for non-stationary data, as it combines trend surface analysis and ordinary kriging into a single analysis. Amongst the various types of kriging, indicator, and co-kriging have been used to estimate disease risk surfaces. Indicator kriging models the probability of an event of interest. This could be, for example, the probability of an environmental contaminant being above a certain threshold level and is derived from input indicator data that needs to be in a binary format (zero indicating absence of, and one indicating presence of, the characteristic). The associated kriging variance reflects the uncertainty associated with the generated probability surface resulting from the interpolation process. Valencia et al. (2005) use this approach to predict the risk of ascariasis in children in Brazil. They report that they were able to improve the prediction by using co-kriging, which considered multiple covariates. Count data are interpolated by Ali et al. (2006) using Poisson kriging and Banerjee et al. (2004) describe Bayesian approaches to kriging. A more complete discussion of different kriging methods is provided in Waller and Gotway (2004).

Fig. 6.7b shows an example where ordinary kriging was used to generate a continuous surface

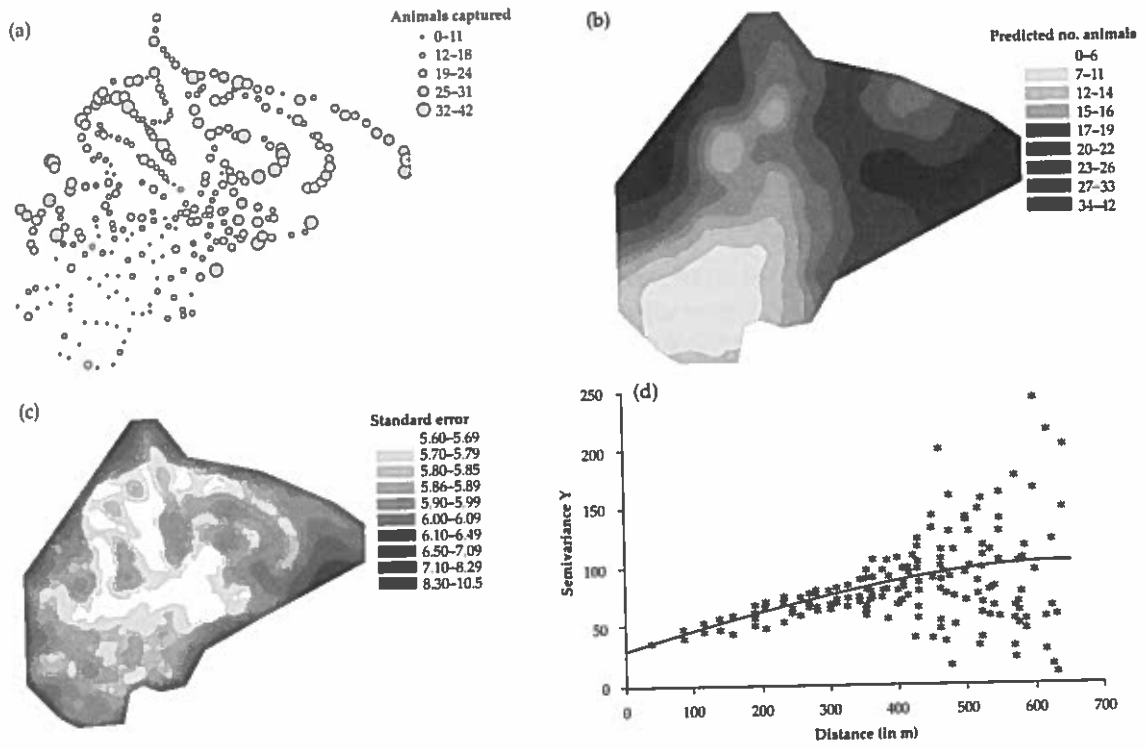


Figure 6.7 Example of ordinary kriging used to generate a continuous surface showing wild possum density in New Zealand derived from possum trap capture data. (a) Point map of trap locations with circle radius representing number of animals captured, (b) kriging surface map of predicted possum numbers, (c) standard error map for kriging estimates, and (d) semivariogram including model used to generate kriging surface.

of wild possum numbers based on trap-capture data for a study area (Fig. 6.7a). There were several problems associated with applying this technique to this particular dataset. Firstly, the traps did not cover the study area evenly and secondly, the data contained biological edge effects since the traps on the boundary would have attracted possums from outside the study area. The theoretical semivariogram function is based on fairly stable empirical semi-variance estimates up to a distance of 400m between point locations (Fig. 6.7d). No directional effects were detected when performing a directional semivariogram analysis using a semivariogram surface (not shown here). The standard error map (Fig. 6.7c) suggests that the error was relatively evenly distributed across the study area, but increased towards the edge. In this particular case, kriging interpolation improved upon the information shown in the point map, resulting in a better impression of the variation in animal density across space. It does however need to be emphasized that, due to the data limitations described above, the predicted surface should only be used for broad interpretation of the underlying spatial variation in possum density.

Universal kriging has been applied in the example presented in Fig. 6.8 (which is presented in more detail in Berke (2004)). The semivariogram was fitted to tapeworm infection data in foxes, which were surveyed in Lower Saxony, Germany, between 1991 and 1997, aggregated at the level of 43 administrative regions. The choropleth map

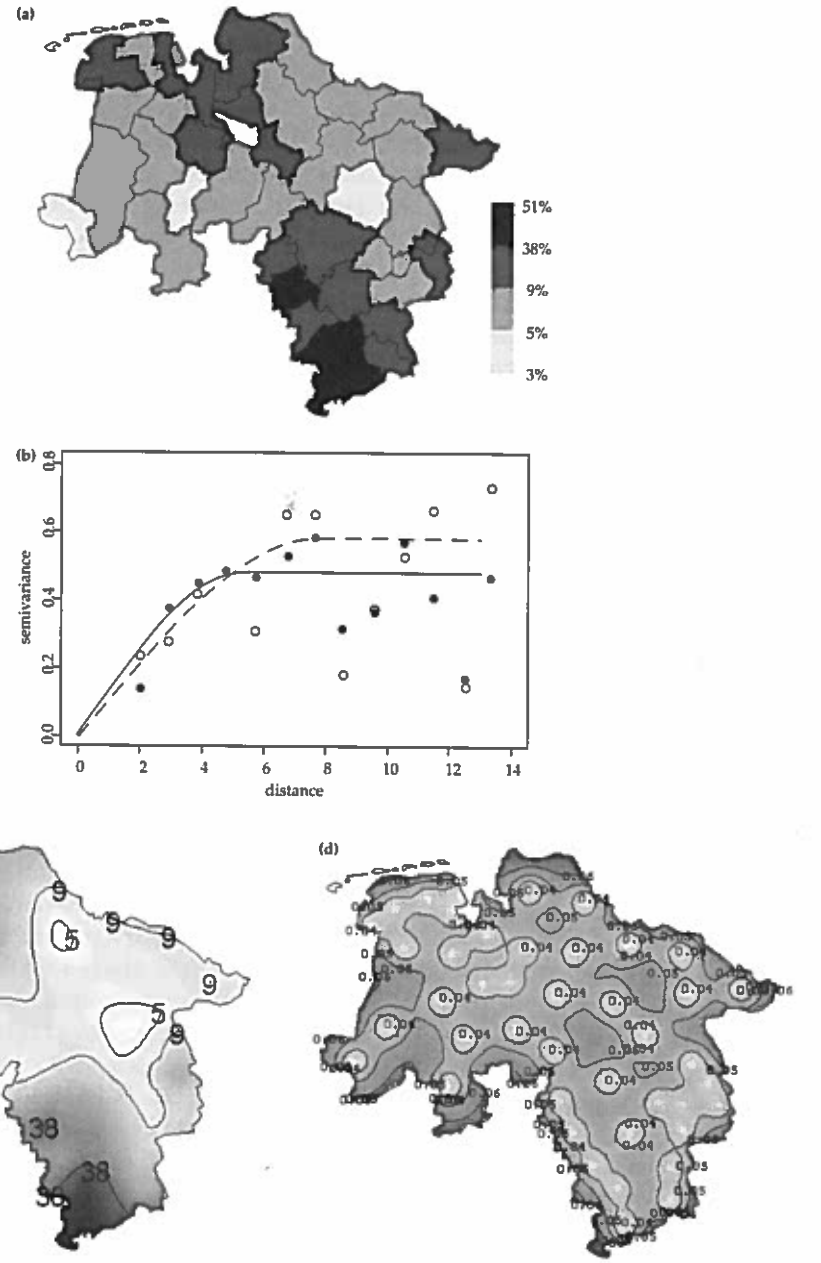


Figure 6.8 Example of universal kriging applied to tapeworm infection prevalence data from foxes surveyed in Lower Saxony, Germany (Reproduced with permission from Berke (2004)). (a) Choropleth map of empirical Bayes smoothed tapeworm prevalences, (b) empirical and theoretical semi-variogram detrended (solid line) and trend-contaminated (broken line) smoothed prevalences, (c) isopleth map from kriging smoothed prevalences, and (d) isopleth error map based on universal kriging standard errors.

shown in Fig. 6.8a indicates a spatial trend with prevalence levels increasing from north to south. Since this reflects the presence of non-stationarity, it is recommended to use universal rather than ordinary kriging. The spherical semivariogram functions shown in Fig. 6.8b demonstrate the benefit obtained from detrending the data, in that the sill is reduced. The resulting kriging surface shown in Fig. 6.8c maintains the key patterns shown in the choropleth map but avoids the jumps between regions. It also provides a prediction for the central Bremen region which had a missing value. The error map shown in Fig. 6.8d suggests the presence of a particular spatial pattern for the errors, which increases confidence in the validity of the model.

6.5 Conclusion

The presentation of spatial variation in risk is one of the most important functions of spatial analysis. If data have been collected for point locations or areas, enhanced visualization is possible by producing continuous surfaces using kernel smoothing and kriging interpolation. Both techniques are now available in most GIS software packages. Kriging is subject to computational constraints when using large datasets as it assesses the covariation between location pairs. While both methods can be used to produce a risk surface, there is otherwise only a limited degree of overlap with respect to suitable applications of the two methods. Kernel smoothing is subject to less restrictive statistical assumptions than kriging, and will not generate negative estimates for risk data. With aggregated ratio data, the statistical uncertainty of local risk estimates can elegantly be reduced by taking advantage of spatial dependence through Bayesian methods. If further smoothing is considered desirable, the Bayesian area estimates can be converted into a smooth surface using kriging. All these methods however, result in changes to the original data values. While this should result in improvements with respect to visual interpretation, it is also likely that artefacts and biases will be introduced and particular care needs to be taken in order to minimize these. Rather than being able to rely solely on the statistical methods, in most cases the user is required to make decisions about the parameters needed for the smoothing algorithms, and these subjective decisions may have quite a dramatic effect on the outcome.

CHAPTER 7

Identifying factors associated with the spatial distribution of disease

7.1 Introduction

The concepts and techniques discussed so far have dealt with describing, visualizing, and exploring spatial data. In this chapter the analytical techniques of regression and discrimination are introduced as a means of quantifying the effect of a set of explanatory variables on the spatial distribution of a particular outcome. The material presented here is similar to that which might be presented in standard statistical texts, but includes an overview of the modifications needed to account for the spatial dependency frequently associated with disease data.

Perhaps the first aspect to consider is the type of outcome variable under investigation. In epidemiology interest lies in understanding patterns of disease in populations, so it is often the case that the outcome variable is either a count of disease events for area units (see for example Jarup et al. 2002), or more simply a binary response indicating the presence or absence of disease at a given location (see for example Diggle et al. 2002). Less frequently in epidemiology (compared with other disciplines) outcomes may be measured on a continuous scale, for example, concentrations of carbon monoxide in an urban environment (Best et al. 2000). Knowledge of the type of outcome variable is important since it determines the regression technique to be used and the options available to account for spatial dependence.

This chapter is divided into four sections. The first outlines the principles of linear, Poisson, and logistic regression in order to provide a background to the material presented later in the chapter. The second section discusses the options available to identify and account for spatial dependency in data when modelling. The third section reviews the common analytical techniques available for dealing with the three major spatial data types (area, point, and continuous data), and the fourth deals with discriminant analysis. The aim is to provide a broad overview of the available methodologies, highlighting the additional information to be gained by accounting for spatial dependence, and to suggest ways in which this information might be used to better understand the determinants of disease in human and animal populations. For a more in-depth discussion the reader is directed to the references cited in each section of the text. Throughout, the British cattle TB data are used to illustrate the concepts discussed.

7.2 Principles of regression modelling

7.2.1 Linear regression

Linear regression allows the mean value of a continuous outcome variable (also known as a response or dependent variable) μ_i to be represented as a function of m explanatory (also known as covariate, predictor, or independent) variables:

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_m X_{mi} + \varepsilon_i \qquad (7.1)$$

If the term X is used to represent the $(m \times i)$ matrix of explanatory variables, the term β_0 in (7.1) is a constant indicating the value of μ_i when $X = 0$. The constants $\beta_1, \ldots, \beta_m$ determine how much the outcome variable changes in response to unit changes in each of the m explanatory variables.

Linear regression is a technique that can be used to model the broad-scale (first-order) spatial trend