11

Disease Mapping

Spatial statistics have been widely applied in epidemiology for the study of the distribution of disease. As we have already shown in Chap. 7, displaying the spatial variation of the incidence of a disease can help us to detect areas where the disease is particularly prevalent, which may lead to the detection of previously unknown risk factors. As a result of the growing interest, Spatial Epidemiology (Elliott et al., 2000) has been established as a new multidisciplinary area of research in recent years.

The importance of this field has been reflected in the appearance of different books and special issues in some scientific journals. To mention a few, recent reviews on the subject can be found in Waller and Gotway (2004), while the special issues of the journal *Statistical Methods in Medical Research* (Lawson, A., 2005) and *Statistics in Medicine* (Lawson et al., 2006) also summarise novel developments in disease mapping and the detection of clusters of disease. Walter and Birnie (1991) compared many different atlases of disease and they compile the main issues to pay attention to when reporting disease maps. Banerjee et al. (2004, pp. 88–97, 158–174) also tackle the problem of disease mapping and develop examples that can be reproduced using `S-PLUS`™ SpatialStats (Kaluzny et al., 1998) and WinBUGS. In addition, some data sets and code with examples are available from the book website.[1] Haining (2003) considers different issues in disease mapping, including a Bayesian approach as well and provides examples, data, and code to reproduce the examples in the book. Schabenberger and Gotway (2005, pp. 394–399) briefly describe the smoothing of disease rates. Finally, Lawson et al. (2003) provide a practical approach to disease mapping, with a number of examples (with full data sets and WinBUGS code) that the reader should be able to reproduce after reading this chapter.

In this chapter we refer to the analysis of data which have been previously aggregated according to a set of administrative areas. The analysis of data available at individual level requires different techniques, which have

[1] `http://www.biostat.umn.edu/~brad/data2.html`

been described in Chap. 7. These kinds of aggregated data are continuously collected by Health Authorities and usually cover mortality and morbidity counts. Special registers have also been set up in several countries to record the incidence of selected diseases, such as cancer or congenital malformations. Spatial Epidemiology often requires the integration of large amounts of data, statistical methods, and geographic information. R offers a unique environment for the development of these types of analysis given its good connectivity to databases and the different statistical methods implemented.

Therefore, the aim of this chapter is not to provide a detailed and comprehensive description of all the methods currently employed in Spatial Epidemiology, but to show those which are widely used. A description as to how they can be computed with R and how to display the results will be provided. From this description, it will be straightforward for the user to adapt the code provided in this chapter to make use of other methods. Other analysis of health data, as well as contents on which this chapter is built, can be found in Chaps. 9 and 10.

The North Carolina SIDS data, which have already been displayed in Chap. 3 (Fig. 3.6), will be used throughout this chapter in the examples that accompany the statistical methodology described here. The SIDS data set records the number of sudden infant deaths in North Carolina for 1974–1978 and 1979–1984 and some other additional information. It is available as `nc.sids` in package **spdep** and further information is available in the associated manual page. Cressie and Read (1985) and Cressie and Chan (1989), for example, provide a description of the data and study whether there is any clustered pattern of the cases.

11.1 Introduction

The aim of disease mapping is to provide a representation of the spatial distribution of the risk of a disease in the study area, which we assume is divided into several non-overlapping smaller regions. The risk may reflect actual deaths due to the disease (mortality) or, if it is not fatal, the number of people who suffer from the disease (morbidity) in a certain period of time for the population at risk.

Hence, basic data must include the population at risk and the number of cases in each area. These data are usually split according to different variables in a number of groups or strata, which can be defined using sex, age, and other important variables. When available, a deprivation index (Carstairs, 2000) is usually employed in the creation of the strata. By considering data in different groups, the importance of each variable can be explored and potential confounding factors can be removed (Elliott and Wakefield, 2000) before doing any other analysis of the data. For example, if the age is divided into 13 groups and sex is also considered, this will lead to 26 strata in the population. Note that depending on the type of study the population at risk may be a reduced subset of the total population. For example, in our examples, it is reduced to the number of children born during the period of study.

Following this structure, we denote by P_{ij} and O_{ij} the population and observed number of cases in region i and stratum j. Summing over all strata j we can get the total population and number of cases per area, which we denote by P_i and O_i. Summing again over all the regions will give the totals, which will be denoted by P_+ and O_+.

Representing the observed number of cases alone gives no information about the risk of the disease given that the cases are mainly distributed according to the underlying population. To obtain an estimate of the risk, the observed number of cases must be compared to an *expected* number of cases.

If P_i and O_i are already available, which is the simplest case, the expected number of cases in region i can be calculated as $E_i = P_i r_+$, where r_+ is the overall incidence ratio equal to $\frac{O_+}{P_+}$. This is an example of the use of *indirect standardisation* (Waller and Gotway, 2004, pp. 12–15) to compute the expected number of cases for each area.

When data are grouped in strata, a similar procedure can be employed to take into account the distribution of the cases and population in the different strata. Instead of computing a global ratio $\frac{O_+}{P_+}$ for all regions, a different ratio is computed for each stratum as $r_j = \frac{\sum_i O_{ij}}{\sum_i P_{ij}}$. In other words, we could compute the ratio between the sum of all cases at stratum j over the population at stratum j. In this situation, the expected number of cases in region i is given by $E_i = \sum_j P_{ij} r_j$.

This standardisation is also called *internal standardisation* because we have used the same data to compute reference rates r_j. Sometimes they are known because another reference population has been used. For example, national data can be used to compute the reference rates to be used later in regional studies.

The following code, based on that available in the `nc.sids` manual page, will read the SIDS data, boundaries of North Carolina, and the adjacency structure of the North Carolina counties (as in Cressie and Read, 1985) in GAL format (see Chap. 9). By using the argument `region.id` we make sure that the order of the list of neighbours `ncCR85` is the same as the areas in the `SpatialPolygonDataFrame` `nc`.

```
> library(maptools)
> library(spdep)
> nc_file <- system.file("shapes/sids.shp", package = "maptools")[1]
> llCRS <- CRS("+proj=longlat +datum=NAD27")
> nc <- readShapePoly(nc_file, ID = "FIPSNO", proj4string = llCRS)
> rn <- sapply(slot(nc, "polygons"), function(x) slot(x,
+     "ID"))
> gal_file <- system.file("etc/weights/ncCR85.gal",
+     package = "spdep")[1]
> ncCR85 <- read.gal(gal_file, region.id = rn)
```

11.2 Statistical Models

A common statistical assumption to model the number of observed number of cases in region i and stratum j is that it is drawn from a Poisson distribution with mean $\theta_i E_{ij}$. Thus, a relative risk of 1 means that the risk is as the average in the reference region (from where the rates r_j are obtained) and it will be of interest in the location of the regions where the relative risk is significantly higher than 1. This basic model is described in Banerjee et al. (2004, pp. 158–159), Haining (2003, pp. 194–199), and Lawson et al. (2003, pp. 2–8).

Note that implicitly we are assuming that there is no interaction between the risk and the population strata, i.e. the relative risk θ_i depends only on the region.

At this point, a basic estimate of the risk in a given region can be computed as $\mathrm{SMR}_i = O_i/E_i$, which is known as the *Standardised Mortality Ratio*. This is why the data involving the cases are often referred to as the *numerator* and the data of the population as the *denominator*, because they are used to compute a ratio that estimates the relative risk. Figure 11.1 shows the SMRs of the SIDS data for the period 1974–1978. Waller and Gotway (2004, pp. 11–18) describe in detail this and other types of standardisation, together with other risk ratios frequently used in practise.

```
> nc$Observed <- nc$SID74
> nc$Population <- nc$BIR74
> r <- sum(nc$Observed)/sum(nc$Population)
> nc$Expected <- nc$Population * r
> nc$SMR <- nc$Observed/nc$Expected
```

Using the fact that O_i is Poisson distributed, we can obtain a confidence interval for each SMR (using function `pois.exact` from package **epitools**). Figure 11.2 displays the 95% confidence interval of the SMR computed for

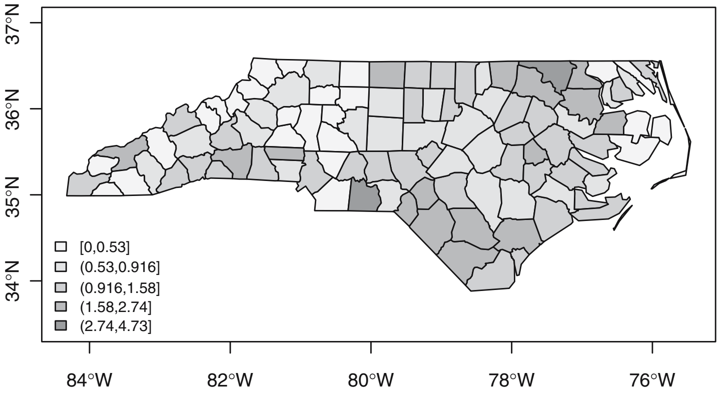

Fig. 11.1. Standardised mortality ratio of the North Carolina SIDS data in the period 1974–1978

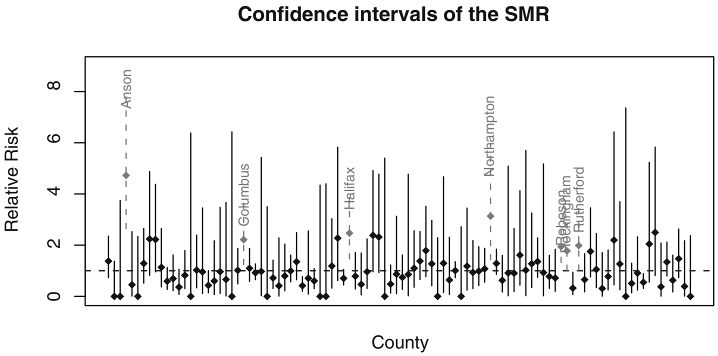

Fig. 11.2. Confidence intervals of the SMR obtained with an exact formula. The black dot represents the SMR of each area. The confidence intervals shown by dashed lines are significantly higher than 1

each area. Highly significant risks (i.e. those whose confidence interval is above one) have been drawn using a dashed line and the name of the county has been added as a label. Anson county, which has been pointed out as a clear extreme value in previous studies (Cressie and Chan, 1989), is the one with the highest confidence interval.

11.2.1 Poisson-Gamma Model

Unfortunately, using a Poisson distribution implies further assumptions that may not always hold. One key issue is that for this distribution the mean and the variance of O_i are supposed to be the same. It is often the case that data are 'over-dispersed', so that the variance of the data is higher than their mean and the statistical model needs to be expanded. A simple way to allow for a higher variance is to use a negative binomial distribution instead of the Poisson.

The negative binomial distribution can also be regarded as a mixed model in which a random effect following a Gamma distribution for each region is considered. This formulation is known as the Poisson-Gamma (PG) model, because it can be structured as the following two-level model:

$$\begin{array}{ll} O_i|\theta_i, E_i & \sim \mathrm{Po}(\theta_i E_i), \\ \theta_i & \sim \mathrm{Ga}(\nu, \alpha). \end{array}$$

In this model, we also consider the relative risk θ_i as a random variable, which is drawn from a Gamma distribution with mean ν/α and variance ν/α^2. Note that now the distribution of O_i is conditioned on the value of θ_i. The unconditioned distribution for each O_i is easy to derive and it is a negative binomial with size parameter ν and probability $\frac{\alpha}{\alpha+E_i}$.

In addition, the posterior distribution of θ_i, i.e. its distribution given the observed data $\{O_i\}_{i=1}^n$, can also be derived and it is a Gamma with parameters

$\nu + O_i$ and $\alpha + E_i$. In other words, the information provided by observing the data has *updated* our prior knowledge or assumptions on θ_i. The posterior expectation of θ_i is

$$E[\theta_i|O_i, E_i] = \frac{\nu + O_i}{\alpha + E_i},$$

which can also be expressed as a compromise between the prior mean of the relative risks and SMR_i, so that this is a *shrinkage* estimator:

$$E[\theta_i|O_i, E_i] = \frac{E_i}{\alpha + E_i}\mathrm{SMR}_i + (1 - \frac{E_i}{\alpha + E_i})\frac{\nu}{\alpha}.$$

Two issues should be noted from this estimator. First of all, when E_i is small, as often happens in low populated areas, a small variation in O_i can produce dramatic changes in the value of SMR_i. For this reason, according to the previous expectation, the SMR_i will have a low weight, as compared to that of the prior mean. Secondly, information is borrowed from all the areas in order to construct the posterior estimates given that ν and α are the same for every region. This concept of *borrowing strength* can be modified and extended to take into account a different set of areas or neighbours.

```
> library(DCluster)
> eb <- empbaysmooth(nc$Observed, nc$Expected)
> nc$EBPG <- eb$smthrr
```

Given that ν and α are unknown, we need a procedure to estimate them. They can be easily estimated from the data using the method of moments, following formulae given by Clayton and Kaldor (1987) to produce Empirical Bayes (EB) estimates, implemented in package **DCluster**. In this example, the values are $\nu = 4.6307$ and $\alpha = 4.3956$, which gives a prior mean of the relative risks of 1.0535 (very close to 1).

Probability maps (Choynowski, 1959) are a convenient way of representing the significance of the observed values. These maps show the probability of a value being higher than the observed data according to the assumption we have made about the model. In other words, probability maps show the p-value of the observed number of cases under the current model. Figure 11.3 represents the probability maps for the Poisson and Poisson-Gamma models. The reason to compare both maps is to show how significance varies with the model. We noted that the Poisson-Gamma model was more appropriate in this case due to over-dispersion, and we should try to make inference based on this model. As expected, the p-values for the Poisson-Gamma model are higher because more variability is permitted. Nevertheless, there are still two zones of high risk to the northeast and south.

11.2.2 Log-Normal Model

Clayton and Kaldor (1987) proposed another risk estimator based on assumtion that the logarithm of the relative risks ($\beta_i = \log(\theta_i)$) follows a multivariate

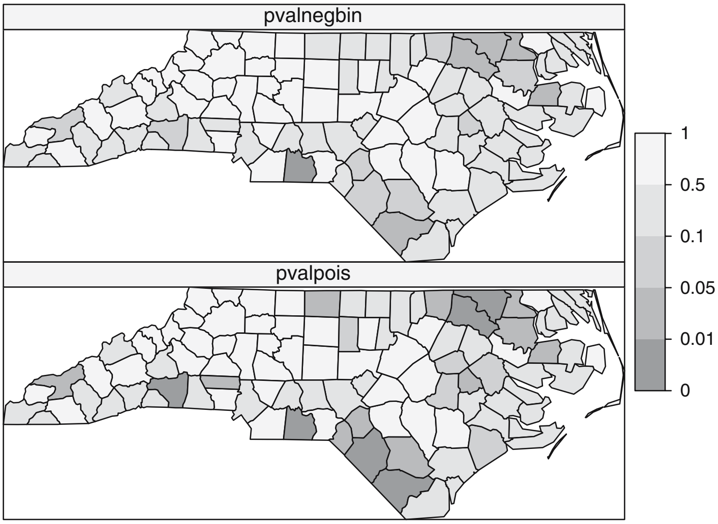

Fig. 11.3. Probability maps for the Poisson and negative binomial models

normal distribution with common mean ϕ and variance σ^2. In this case, the estimate of the log-relative risk is not taken as $\log(O_i/E_i)$ but $\log((O_i{+}1/2)/E_i)$, because the former is not defined if O_i is zero. The EM algorithm is used to obtain estimates of the mean and variance of the model, which can be plugged in to the following Empirical Bayes estimator of β_i:

$$\hat{\beta}_i = b_i = \frac{\hat{\phi} + (O_i + \frac{1}{2})\hat{\sigma}^2 \log[(O_i + \frac{1}{2})/E_i] - \hat{\sigma}^2/2}{1 + (O_i + \frac{1}{2})\hat{\sigma^2}},$$

where $\hat{\phi}$ and $\hat{\sigma}^2$ are the estimates of the prior mean and variance, respectively. These are given by

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} b_i = \overline{b}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\left\{\hat{\sigma}^2 \sum_{i=1}^{n} [1 + \hat{\sigma}^2 (O_i + 1/2)]^{-1} + \sum_{i=1}^{n} (b_i - \hat{\phi})^2\right\}.$$

Estimates b_i are updated successively using previous formulae until convergence. Hence, the estimator for θ_i is $\hat{\theta}_i = \exp\{\hat{\beta}_i\}$. Note that now the way information is borrowed is to estimate the common parameters ϕ and σ^2, and that the resulting estimates are a combination of the local estimate of the log relative risk and ϕ. Unfortunately, the current estimator is more complex than the previous one and it cannot be reduced to a shrinkage expression.

```
> ebln <- lognormalEB(nc$Observed, nc$Expected)
> nc$EBLN <- exp(ebln$smthrr)
```

11.2.3 Marshall's Global EB Estimator

Marshall (1991) developed a new EB estimator assuming that the relative risks θ_i have a common prior mean μ and variance σ^2, but without specifying any distribution. By using the method of moments, he is able to work a new estimator out employing a shrinkage estimator as follows:

$$\hat{\theta}_i = \hat{\mu} + C_i(\mathrm{SMR}_i - \hat{\mu}) = (1 - C_i)\hat{\mu} + C_i\mathrm{SMR}_i,$$

where

$$\hat{\mu} = \frac{\sum_{i=1}^{n} O_i}{\sum_{i=1}^{n} E_i}$$

and

$$C_i = \frac{s^2 - \hat{\mu}/\overline{E}}{s^2 - \hat{\mu}/\overline{E} + \hat{\mu}/E_i}.$$

$\overline{E}$ stands for the mean of the E_i's and s^2 is the usual unbiased estimate of the variance of the SMR_i's. Unfortunately, this estimator can produce negative estimates of the relative risks when $s^2 < \hat{\mu}/\overline{E}$, in which case $\hat{\theta}_i = \hat{\mu}$ is taken.

The shrinkage of this estimator highly depends (again) on the value of E_i. If it is high, which means that the SMR_i is a reliable estimate, C_i will be close to 1 and the estimator will give more weight to the SMR_i. On the other hand, if E_i is small, more weight is given to the estimate of the prior mean $\hat{\mu}$ because the SMR_i is less reliable and so it borrows more information from other areas.

```
> library(spdep)
> EBMarshall <- EBest(nc$Observed, nc$Expected)
> nc$EBMarshall <- EBMarshall[, 2]
```

Figure 11.4 represents the different estimates obtained by the different estimators described so far. All EB estimators seem to produce very similar estimates in all the areas. By comparing those maps to the map that shows the SMR, it is possible to see how very extreme values (either high or low) have been shifted towards the global mean. In other words, these values have been *smoothed* by taking into account global information in the computation of the estimate.

To compare the variability of the estimates produced by each method, we have created a boxplot of each set of values, which appear in Fig. 11.6. From the plot it is clear that the SMR is the most variable and that the other three have been shrunk towards the global mean, which is approximately 1. Hence, we might expect similar results when using any of the EB estimators. As pointed by Marshall (1991), the estimation procedure based on the Poisson-Gamma model proposed by Clayton and Kaldor (1987) may not converge in

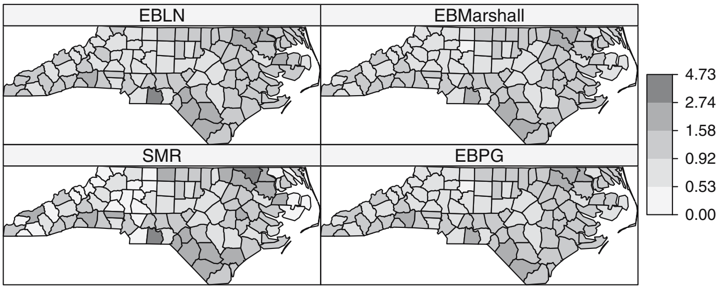

Fig. 11.4. Comparison of different risk estimators. SMR displays the standardised mortality ratio, whilst EBPG, EBLN, and EBMarshall show different empirical Bayes estimates using the Poisson-Gamma model, the log-normal model, and Marshall's global estimator

some circumstances and another estimator should be used. The EB proposed by Marshall (1991) can also be unfeasible in similar circumstances. Hence, the EB estimator based on the log-Normal model seems to be the most computationally stable and reliable.

All these EB estimators produce smoothed estimates of the risk rates *borrowing information* from the global area but, depending on the size and extension of the total area under study, it could be more reasonable to consider only a small set of areas that are close to each other. A common example is to use only the areas that share a boundary with the current region to compute its risk estimate. Unfortunately, this procedure involves the use of more complex models that require the use of additional software and will be discussed in the following sections.

11.3 Spatially Structured Statistical Models

Although borrowing strength globally can make sense in some cases, it is usually better to consider a reduced set of areas to borrow information from. A sensible choice is to take only neighbouring areas or areas which are within a certain distance from the current area.

Marshall (1991) proposed another estimator that requires only local information to be computed. For each region, a set of neighbours is defined and local means, variances, and shrinkage factors are defined in a similar way as in the global estimator, but considering only the areas in the neighbourhood. This produces a local shrinkage for each area, instead of the global shrinkage provided by the previous estimator.

```
> nc$EBMrshloc <- EBlocal(nc$Observed, nc$Expected, ncCR85)$est
```

The way this estimator is computed raises a new question about how areas are related to each other. In the previous models, no account for how areas were distributed in the study region was considered, so that the influence of a region did not depend on its location at all. That is, we would obtain the same estimates if the distribution of the regions were permutated at random. With the new estimator the exact location of the areas is crucial, and different locations of the regions will give different estimates as a result. The way regions are placed in a map can be described by means of its *topology*, which accounts for the neighbours of a given region. See Chap. 9 for more details on this and how to obtain it.

Although neighbours are usually defined as two regions that share a common boundary, Cressie and Chan (1989) define two regions as neighbours if the distance between their centroids is within 30 miles. This is not a trivial issue since different definitions of neighbourhood will produce different results.

The two estimators proposed by Marshall have been displayed in Fig. 11.5. The version that uses only local information produces smoothed estimates of the relative risks that are shrunk towards the local mean that turned out to be less shrunk towards the global mean. In addition, the shrinkage produced by the local estimator is in general lower than that for the global estimator.

The boxplot presented in Fig. 11.6 compares the different EB estimators discussed so far. Marshall's local estimator also shows a general shift towards the global mean, but it is less severe than for the others because only local information is employed. In general, EB smoothed estimators have been criticised because they fail to cope with the uncertainty of the parameters of the model (Bernardinelli and Montomoli, 1992) and to produce an overshrinkage since the parameters of the prior distributions are estimated from the data

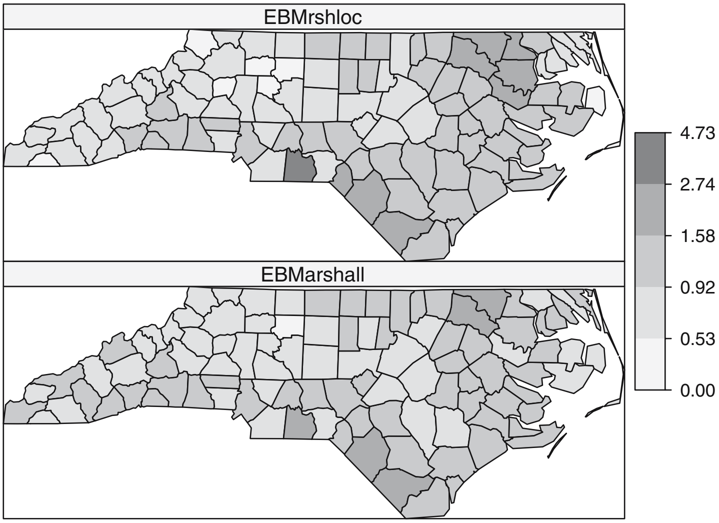

Fig. 11.5. Marshall's EB estimator using local (*top*) and global (*bottom*) information

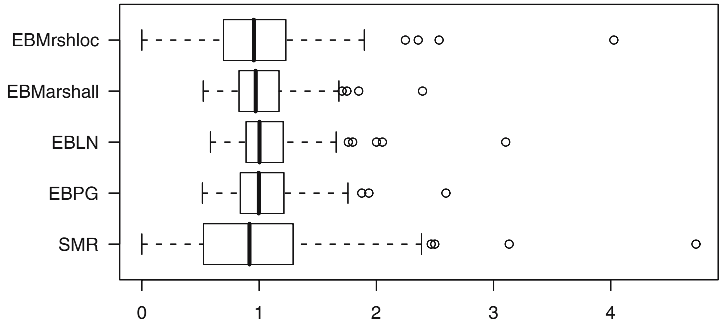

Fig. 11.6. Comparison of raw and EB estimators of the relative risk

and remain fixed. To solve this problem several *constrained* EB estimators have been proposed to force the posterior distribution of the smoothed estimates to resemble that of the raw data (Louis, 1984; Devine and Louis, 1994; Devine et al., 1994).

Full Bayes methods allow setting the prior distributions for these parameters and, hence, permit a greater variability and produce more suitable smoothed estimates. More standard smoothed risk estimators that borrow information locally can be developed by resorting to Spatial Autoregressive and Conditional Autoregressive specifications (Waller and Gotway, 2004). Basically, these models condition the relative risk in an area to be similar to the values of the neighbouring areas. More details are given in the next sections of this chapter and in Chap. 10 for non-Bayesian models.

11.4 Bayesian Hierarchical Models

Bayesian Hierarchical Models make an appropriate framework for the development of spatially structured models. The model is specified in different layers, so that each one accounts for different sources of variation. For example, they can cope with covariates at the same time as borrowing strength from neighbours to improve the quality of estimates. The use of these models in disease mapping is considered in Haining (2003, pp. 307–311, 367–376), Waller and Gotway (2004, pp. 409–429), Banerjee et al. (2004, pp. 159–169), and Schabenberger and Gotway (2005, pp. 394–399). Lawson et al. (2003) offer a specific volume on the subject, with reproducible examples.

Besag et al. (1991), BYM henceforth, introduced in their seminal paper a type of models that split the variability in a region as the sum of a spatially correlated variable (which depends on the values of its neighbours) plus an area-independent effect (which reflects local heterogeneity). Although direct estimates of the variables in the model can seldom be obtained when using Bayesian Hierarchical Models, their posterior distributions can be obtained by means of Markov Chain Monte Carlo (MCMC) techniques. Basically, MCMC methods generate simulations of the parameters of the model which, after a suitable burn-in period, become realisations of their posterior distributions. An introduction to MCMC and its main applications, including disease mapping, can be found in Gilks et al. (1996).

WinBUGS (Spiegelhalter et al., 2003) is software that uses MCMC methods (in particular, Gibbs Sampling; Gelman et al., 2003) to simulate from the posterior distributions of the parameters in the model. Starting from a set of initial values, one sample of each variable is simulated at the time using the full conditional distribution of the parameter given the other parameters. After a suitable burn-in period, the simulations generated correspond to the joint posterior distribution.

Although WinBUGS is the main software package, it was previously known as BUGS and currently it comes in different flavours. OpenBUGS, for example is the open source alternative to WinBUGS and it is actually a fork of the main WinBUGS software. Apart from the advantage of coming with the source code, OpenBUGS can be called from R using package **BRugs**. In addition, some specific plug-ins have been developed for WinBUGS to deal with certain applications. It is worth mentioning GeoBUGS, which provides a graphical interface to the management of maps and compute adjacency relationships within WinBUGS and OpenBUGS, and it can create maps with the results. Lawson et al. (2003) have described extensively how to do a disease mapping using Multilevel Models with WinBUGS (and MLwiN), and is a complete reference for those readers willing to go deeper in this subject.

Another package to use WinBUGS from R is **R2WinBUGS** (Sturtz et al., 2005). This package calls WinBUGS using its scripting facilities so that the resulting log file containing all the results can be loaded into R after the computations have finished. **R2WinBUGS** will be the package used in this book. The main reason is that at the time of writing **BRugs** only works on Windows (although the authors claim that it should also work on Linux with minor modifications) whilst **R2WinBUGS** can be used on several platforms with minor adjustments. Under Linux, for example it can be run using the Wine programme. Finally, it is worth noting that Gelman and Hill (2007) provide a good and accessible text on data analysis using Bayesian hierarchical models and describe the use of R and WinBUGS via **R2WinBUGS** in Chaps. 16 and 17.

11.4.1 The Poisson-Gamma Model Revisited

The following example shows a full Bayesian Poisson-Gamma formulation (i.e. assigning priors to the parameters ν and α) to produce smoothed estimates of the relative risks that can be run from R using **R2WinBUGS**. In this model, ν and α have been assigned vague gamma priors so that as little prior information as possible is introduced. The WinBUGS code needed to run the Poisson-Gamma model is shown in Fig. 11.7.

```
model
{

  for(i in 1:N)
  {
     observed[i]~dpois(mu[i])
     mu[i]<-theta[i]*expected[i]
     theta[i]~dgamma(nu, alpha)
  }

  nu~dgamma(.01, .01)
  alpha~dgamma(.01, .01)
}
```

Fig. 11.7. Code of the Poisson-Gamma model for WinBugs

The next chunk of code shows how to convert all the necessary data into the structure used by WinBUGS. In addition, we need to set up the initial values for some of the parameters of the model. Data and initial values must be saved into a separated file.

```
> library(R2WinBUGS)
> N <- length(nc$Observed)
> d <- list(N = N, observed = nc$Observed, expected = nc$Expected)

> pgmodelfile <- paste(getwd(), "/PG-model.txt", sep = "")
> wdir <- paste(getwd(), "/PG", sep = "")
> if (!file.exists(wdir)) {
+     dir.create(wdir)
+ }
> BugsDir <- "/home/asdar/.wine/dosdevices/c:/Program Files/WinBUGS14"
> MCMCres <- bugs(data = d, inits = list(list(nu = 1, alpha = 1)),
+     working.directory = wdir, parameters.to.save = c("theta",
+         "nu", "alpha"), n.chains = 1, n.iter = 20000,
+     n.burnin = 10000, n.thin = 10, model.file = pgmodelfile,
+     bugs.directory = BugsDir, WINEPATH = "/usr/bin/winepath")
```

Briefly explained, the `bugs` function will take data, initial values, model file, and other information required and it will create a script that will be run with WinBUGS.[2] `bugs` will create the necessary files (data, initial values, and script) that will be placed under `working.directory`. After running the model, the output will be stored here as well. The WinBUGS script will basically check the syntax of the model, load the data, and compile the model. The following step is to read (or generate from the priors) the initial values for the parameters of the model and 10,000 simulations of the Markov Chain

[2] Windows users must modify the paths in `working.directory`, `model.file`, and `bugs.directory` accordingly, and remove the argument `WINEPATH`, which is not needed.

are generated (keeping just 1 every 10). Note that, since we need a burn-in period, these are not saved. Then, we set that variables ‘nu’, ‘alpha’, and ‘theta’ will be saved and 10,000 more simulations are generated, of which only 1 of every 10 are saved to avoid autocorrelation and improve mixing and convergence. Finally, the summary statistics and plots are saved into the log files under the working directory. Two such files are created: an ODC file (WinBUGS format) with summary statistics and plots, and an ASCII file with the summary statistics. In addition, a summary of the output is stored as a series of lists in `MCMCres`. The posterior mean and median of the relative risks can be extracted as follows:

```
> nc$PGmean <- MCMCres$mean$theta
> nc$PGmedian <- MCMCres$median$theta
```

Although it will not be described here in detail, it is essential to check that the Markov Chain has converged so that the values that we are using have been drawn from the posterior distribution of the parameters. A example using package **coda** (Best et al., 1995) is shown later in a more complex model.

As we have obtained samples from the posterior distributions of ν and α, it is possible to compute pointwise estimates and probability intervals for both parameters. For the sake of simplicity and to be able to compare the values obtained with those from the EB approach, the pointwise estimates (posterior means) of these values were $\hat{\nu} = 6.253$ and $\hat{\alpha} = 5.967$, which are slightly higher than the ones obtained with the EB estimator. Similar estimates can be obtained for the relative risks, but note that now they are not based on single values of ν and α, but that the relative risk estimates are *averaged* over different values of those parameters.

Even though point estimates of the relative risks are usually very useful, for most applications it is better to give a credible interval, for it can be used to detect areas of significantly high risk, if the interval is over 1. Figure 11.8 summarises the 95% credible intervals for each region. The median has been

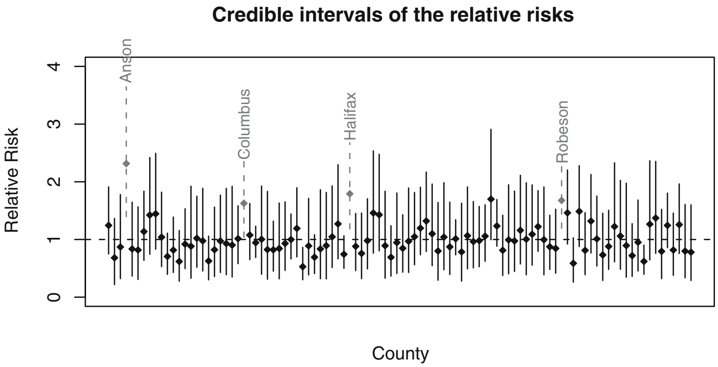

Fig. 11.8. 95% credible intervals of the relative risks obtained with WinBUGS using a full Bayes Poisson-Gamma model

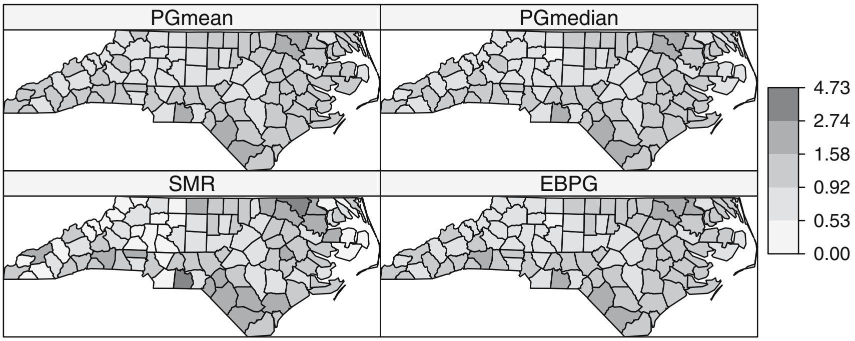

Fig. 11.9. Comparison of empirical Bayes and full Bayes estimates of the relative risks using a Poisson-Gamma model

included (black dot), and the areas whose credible intervals are above 1 have been highlighted using a dashed line and the county name displayed. As we mentioned before, Anson county is of special interest because it shows the highest risk.

In Fig. 11.9 we have compared the estimates of the relative risks provided by the Poisson-Gamma model using both Empirical Bayes and Full Bayes approaches. Both estimation procedures lead to very similar estimates and they only differ in a few areas. Note how they all provide smoothed estimates of the relative risks, as compared to the raw SMRs.

11.4.2 Spatial Models

Additional spatial structure can be included by considering a CAR model and covariates can be used to explain part of the variability of the relative risks. Cressie and Chan (1989) considered the proportion of non-white births as an important factor related to the incidence of SIDS. A full description of these models can be found in Banerjee et al. (2004, Chap. 5). In general, these models are far more complex than the Poisson-Gamma described before, and they should be used with extreme caution because of the high number of parameters and possible interactions between them.

As described in Sect. 10.2.1, the CAR specification for a set of random variables $\{v_i\}_{i=1}^n$ can be written as follows:

$$v_i | v_{-i} \sim N\left(\sum_{j\sim i} \frac{w_{ij} v_j}{\sum_j w_{ij}}, \sigma_v^2 / \sum_j w_{ij}\right),$$

where w_{ij} is a weight that measures the strength of the relationship between (neighbour) regions i and j and σ_v^2 indicates the conditional variance of the CAR specification.

Although the conditional distributions are proper, it is not the case for the joint distribution. Nevertheless, this CAR specification is often used as a prior distribution of the spatial random effects and it can lead to a proper posterior under some constraints (Ghosh et al., 1998).

Given the structure of the CAR specification, it is necessary to know the neighbours of each region. They can be defined in different ways, depending on the type of relationship that exists between the areas. In our example, we use the same neighbourhood structure as in Cressie and Read (1985), which can be found in package `spdep`. In addition, it is necessary to assign a weight to each pair of neighbours, which measure the strength of the interaction. Following Besag et al. (1991), we set all the weights to 1 if regions are neighbours and 0 otherwise.

The flexibility of the Bayesian Hierarchical Models allows us to perform an Ecologic Regression (English, 1992) at the same time as we consider independent and spatial random effects. By including covariates in our model we aim to assess and remove the effect of potential confounders or risk factors. The assessment of the importance of a covariate is indicated by the estimated value of its coefficient and its associated probability interval. If, for example, the 95% credible interval does not contain the value 0, we may assume that the coefficient is significant and, if greater than zero, it will indicate a positive relationship between the risk and the variable.

The results of an Ecologic Regression can be potentially misleading if we try to make inference at the individual level, since the effects that operate at that level may not be the same as those reflected at the area level. In the extreme case, the effects might even be reversed. A solution to this is to combine the aggregated data with some individual data from a specific survey, which can be also used to improve the estimation of the effects of the covariates (Jackson et al., 2006).

In our example, we have the available number of non-white births in each county. The variable ethnicity is often used in the United States as a surrogate of the deprivation index (Krieger et al., 1997). Considering this variable in our model may help to explain part of the spatial variability of the risk of SIDS. To account for the ethnicity, we use the proportion of non-white births in the area. This also allow us to compare the values for different counties. Figure 11.10 shows the spatial variation of the proportion of non-white births. Notice how there exists a similar pattern to that shown by the spatial distribution of the SMR and the different EB estimates. Finally, the WinBUGS model used in this case can be found in Fig. 11.11. We have used the priors suggested in Best et al. (1999) to allow a better identifiability of the random effects u_i and v_i.

The chunk of code shown below converts the neighbours of each county as specified in Cressie and Read (1985) into the format required by WinBUGS. Note that these are already available in an R object and that they have been matched so that the list of neighbours is in the right order. When this is not the case, proper matching must be done. Function `nb2WB` can be used to

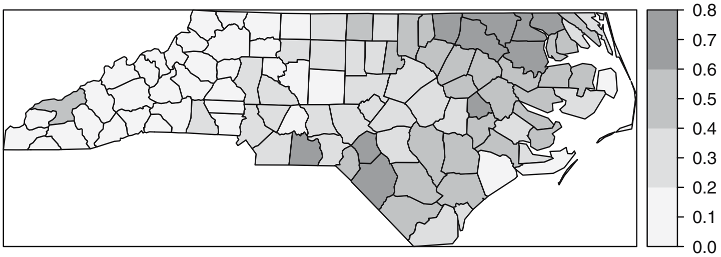

Fig. 11.10. Proportion of non-white births in Carolina, 1974–1978. Notice the similar pattern to the relative risk estimates

```
model
{

  for(i in 1:N)
  {
     observed[i] ~ dpois(mu[i])
     log(theta[i]) <-  alpha + beta*nonwhite[i] + u[i] + v[i]
     mu[i] <- expected[i]*theta[i]

     u[i] ~ dnorm(0, precu)
  }

  v[1:N] ~ car.normal(adj[], weights[], num[], precv)

  alpha ~ dflat()
  beta ~ dnorm(0,1.0E-5)
  precu ~ dgamma(0.001, 0.001)
  precv ~ dgamma(0.1, 0.1)

  sigmau<-1/precu
  sigmav<-1/precv
}
```

Fig. 11.11. Code of the Besag-York-Mollié model for WinBugs

convert an `nb` object into a list containing the three elements (`adj`, `weights`, and `num`) required to use a CAR specification in WinBUGS. Similarly, the function `listw2WB` can be used for a `listw` object. The main difference is that `nb2WB` sets all the weights to 1, whilst `listw2WB` keeps the values of the weights as in the `listw` object.

```
> nc.nb <- nb2WB(ncCR85)
```

The last step is to compute the proportion of non-white births in each county and create the R lists with the data and initial values.

```
> nc$nwprop <- nc$NWBIR74/nc$BIR74
> d <- list(N = N, observed = nc$Observed, expected = nc$Expected,
+     nonwhite = nc$nwprop, adj = nc.nb$adj, weights = nc.nb$weights,
+     num = nc.nb$num)
> dwoutcov <- list(N = N, observed = nc$Observed,
+     expected = nc$Expected, adj = nc.nb$adj, weights = nc.nb$weights,
+     num = nc.nb$num)
> inits <- list(u = rep(0, N), v = rep(0, N), alpha = 0,
+     beta = 0, precu = 0.001, precv = 0.001)
```

The procedure to run this model is very similar to the previous one. We only need to change the file names of the model, data, and initial values. Notice that not all initial values must be provided and that some can be generated randomly. In this model, we are going to keep the summary statistics for a wide range of variables. In addition to the relative risks θ_i, we want to summarise the values of the intercept (α), the coefficient of the covariate (β), and the values of the non-spatial (u_i) and spatial (v_i) random effects.

```
> bymmodelfile <- paste(getwd(), "/BYM-model.txt", sep = "")
> wdir <- paste(getwd(), "/BYM", sep = "")
> if (!file.exists(wdir)) {
+     dir.create(wdir)
+ }
> BugsDir <- "/home/asdar/.wine/dosdevices/c:/Program Files/WinBUGS14"

> MCMCres <- bugs(data = d, inits = list(inits),
+     working.directory = wdir, parameters.to.save = c("theta",
+         "alpha", "beta", "u", "v", "sigmau", "sigmav"),
+     n.chains = 1, n.iter = 30000, n.burnin = 20000,
+     n.thin = 10, model.file = bymmodelfile, bugs.directory = BugsDir,
+     WINEPATH = "/usr/bin/winepath")
```

After running the model, the summary statistics are added to the spatial object that contains all the information about the North Carolina SIDS data so that it can be displayed easily.

```
> nc$BYMmean <- MCMCres$mean$theta
> nc$BYMumean <- MCMCres$mean$u
> nc$BYMvmean <- MCMCres$mean$v
```

Convergence of the Markov Chain must be assessed before attempting any valid inference from the results. Cowles and Carlin (1996) provide a summary of several methods and a useful discussion. They state the difficulty to assess convergence in practise. Some of the criteria discussed in the paper are implemented in package **coda**. These criteria can be applied to the *deviance* of the model to monitor convergence of the joint posterior. Ideally, several chains (each one starting at a sufficiently different point) can be run in parallel so that the traces can be compared (Gelman and Rubin, 1992).

WinBUGS can produce the output in the format required by **coda**. Basically, it will produce an index file (`codaIndex.txt`) plus another file with the values of the variables (`coda1.txt`) that can be read using function `read.coda`.

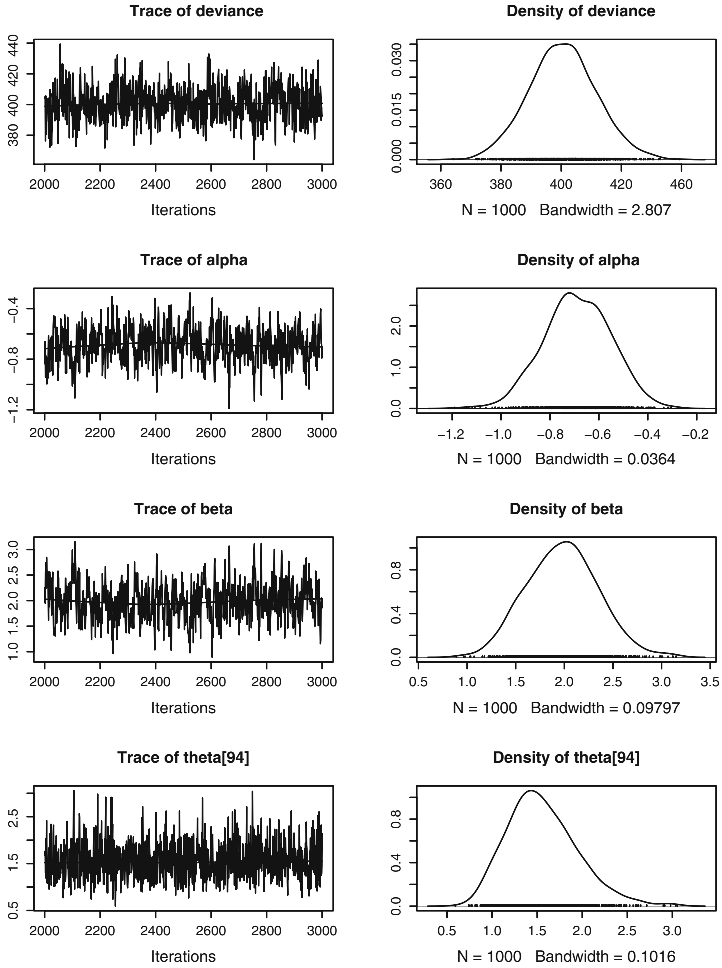

Fig. 11.12. Plots of the posterior distributions of α, β, and the deviance of the model

This will create an object of type `mcmc`, which contains the simulations from all the variables saved in WinBUGS. Figure 11.12 shows the trace and density of the posterior distribution of the deviance and the parameters α, β and the relative risk of Robeson county (area number 94 and cluster centre in Fig. 11.18).

For a single chain, Geweke's criterion (Geweke, 1992) can be computed to assess convergence. It is a score test based on comparing the means of the first and the last part of the Markov Chain (by default, the 10% initial values to the 50% last values). If the chain has converged, both means should be equal. Given that it is a score test, values of the test statistics between -1.96 and 1.96 indicate convergence, whilst more extreme value will denote lack of convergence. For the selected parameters, it seems that convergence has been reached:

```
> geweke.diag(ncoutput[, c("deviance", "alpha", "beta",
+     "theta[94]")])

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

 deviance     alpha      beta theta[94]
   0.1216   -0.8550    0.4239   -1.0669
```

Figure 11.13 shows the SMR and the smoothed estimate of the relative risks obtained. When the posterior distribution is very skewed, the posterior median can be a better summary statistic, but it is not the case now.

According to the posterior density of β shown in Fig. 11.12, the coefficient of the covariate can be considered as significantly positive given that its posterior mean is greater than 0 and its 95% credible interval is likely not to contain the value 0. This means that there is an actual risk increase in those regions with a high proportion of non-white births. Point posterior estimates (mean) of the random effects u_i and v_i are shown in Figs. 11.14 and 11.15, respectively. They seem to have a very small variation, specially the former, but this is not so because they are in the log-scale.

It should be noted that if the spatial pattern is weak or appropriate covariates are included in the model, the random effects u_i and v_i may become unidentifiable. However, following Besag et al. (1995), valid inference could

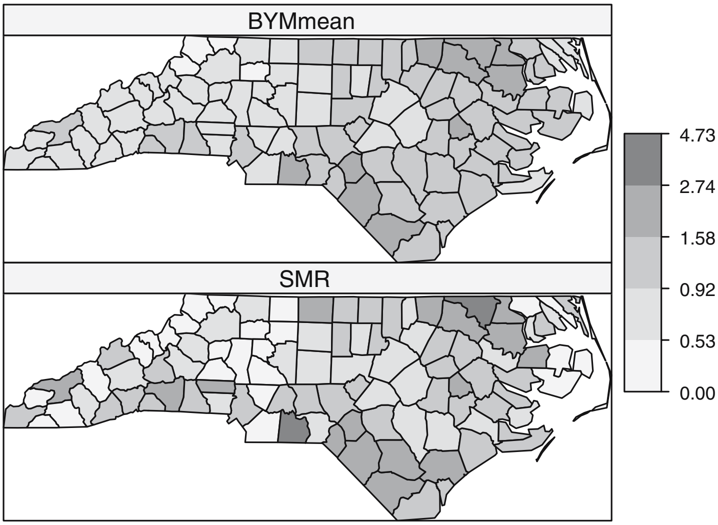

Fig. 11.13. Standardised Mortality Ratio and posterior means of the relative risks obtained with the BYM model

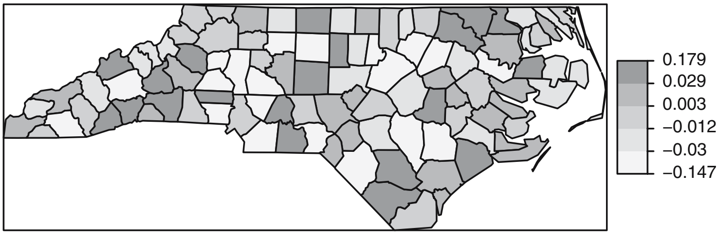

Fig. 11.14. Posterior means of the non-spatial random effects (u_i) estimated with the BYM model

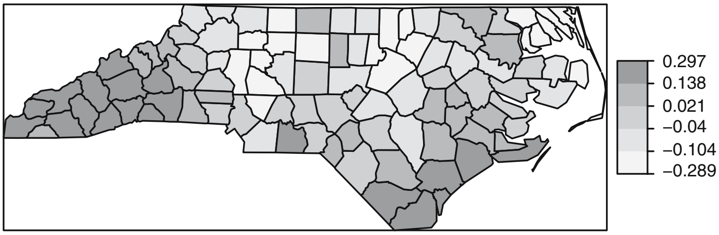

Fig. 11.15. Posterior means of the spatial random effects (v_i) estimated with the BYM model

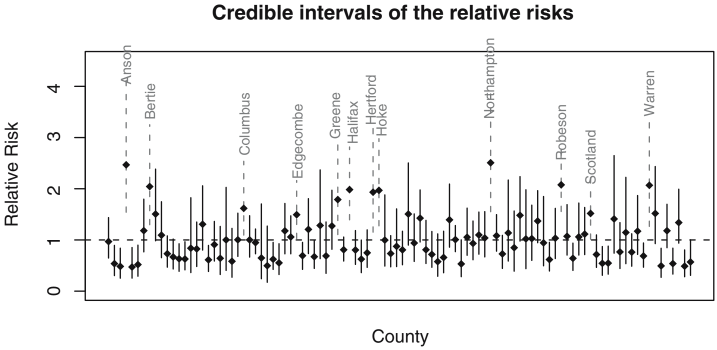

Fig. 11.16. 95% credible intervals of the relatives risks obtained with the BYM model

still be done for the relative risks, but care should be taken to avoid having an improper posterior. For this reason, we can monitor $u_i + v_i$ to assess that these values are stable and that they do not have an erratic behaviour that could have an impact on the posterior estimates of the relative risks and the coefficients of the covariates.

The credible intervals of the relative risks have been plotted in Fig. 11.16. The intervals in dashed line show the counties where the relative risk is significantly higher than one. All these regions are among the ones that appear in the two zones of high risk, plus Anson county.

A few words must be said as to how we have selected the intervals and colours to display the relative risks in the maps. The intervals have been chosen by taking the cut points equally spaced in the range of the relative risks in the log scale (N. Best, personal communication). As discussed in Chap. 3, the colours used to produce the maps can be based on the palettes developed by Brewer et al. (2003), which are available in package **RColorBrewer**. The research was initiated by Brewer et al. (1997) to produce an atlas of disease in the United States. Brewer and Pickle (2002) study how the variable intervals and colours affect how maps are perceived, and Olson and Brewer (1997) developed a useful set of palettes to be used in disease mapping and that are suitable for colour-blind people.

11.5 Detection of Clusters of Disease

Disease mapping provides a first insight to the spatial distribution of the disease, but it may be required to locate the presence of zones where the risk tends to be unusually higher than expected. Besag and Newell (1991) distinguish between methods for clustering and the assessment of risk around putative pollution sources. The former tackle the problem of assessing the presence of clusters, whilst the latter evaluate the risk around a pre-specified source. A third type of method is related to the location of the clusters themselves, which usually involve the examination of small portions of the whole study area each at a time.

Wakefield et al. (2000) provide a review of some classic methods for the detection of clusters of disease. Haining (2003, pp. 237–264) summarises a good number of well-known methods. Waller and Gotway (2004) also cover in detail most of the methods described in this chapter and many others, providing a discussion on the statistical performance of the tests (pp. 259–263). Lawson et al. (2003, Chap. 7) describe the use of Hierarchical Bayesian models for the analysis of risk around pollution sources.

Some of these methods have been implemented in package **DCluster** (Gómez-Rubio et al., 2005), which uses different models and bootstrap (Davison and Hinkley, 1997) to compute the significance of the observed values. This can be done in a general way by resampling the observed number of cases in each area and re-computing the value of the test statistic for each of the simulated data sets. Then, a p-value can be computed by ranking the observed value of the test statistic among the values obtained from the simulations.

Under the usual assumption that O_i is drawn from a Poisson with mean $\theta_i E_i$ and conditioning on the total number of cases, the distribution of $(O_1, \ldots, O_n)$ is Multinomial with probabilities $(E_1/E_+, \ldots, E_n/E_+)$. In addition to the multinomial model, **DCluster** offers the possibility of sampling using a non-parametric bootstrap, or from a Poisson (thus, not conditioning on O_+) or Negative Binomial distribution, to account for over-dispersion in the data. As discussed below, over-dispersion may affect the p-value of the test and when data are highly over-dispersed it may be worth re-running the test sampling from a Negative Binomial distribution.

11.5.1 Testing the Homogeneity of the Relative Risks

Before conducting any analysis of the presence of clusters, the heterogeneity of the relative risks must be assessed. In this way, we can test whether there are actual differences among the different relative risks. The reasons for this heterogeneity may be related to many different factors, such as the presence of a pollution source in the area, which may lead to an increase in the risk around it. Other times the heterogeneity is due to a spatially varying risk factor, and higher risks are related to a higher exposure to this risk factor.

Given that for each area we have computed its expected and observed number of cases, a chi-square test can be carried out to test for (global) significant differences between these two quantities. The statistic is defined by the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - \theta E_i)^2}{\theta E_i},$$

where θ is the global SMR $= \sum_i O_i / \sum_i E_i$ and, asymptotically, it follows a chi-square distribution with n degrees of freedom. If internal standardisation has been used to obtain E_i, then θ is equal to one and the number of degrees of freedom are reduced to $n-1$ because the additional constraint $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i$ holds (Wakefield et al., 2000).

```
> chtest <- achisq.test(Observed ~ offset(log(Expected)),
+     as(nc, "data.frame"), "multinom", 999)
> chtest

Chi-square test for overdispersion

    Type of boots.: parametric
    Model used when sampling: Multinomial
    Number of simulations: 999
    Statistic:  225.5723
    p-value :  0.001
```

Note that in this case we know that the asymptotic distribution of the test statistic is a chi-square with $n-1$ degrees of freedom and that an exact test can be done instead of re-sampling (however, it may still be useful for small samples and recall that we may be interested in a Monte Carlo Test using a Negative Binomial).

```
> 1 - pchisq(chtest$t0, 100 - 1)

[1] 7.135514e-12
```

Potthoff and Whittinghill (1966) proposed another test of homogeneity of the means of different Poisson distributed variables, which can be used to test the homogeneity of the relative risks (Wakefield et al., 2000). The alternative hypothesis is that the relative risks are drawn from a gamma distribution with mean λ and variance σ^2:

$$
\begin{aligned}
H_0 &: \theta_1 = \ldots = \theta_n = \lambda, \\
H_1 &: \theta_i \sim Ga(\lambda^2/\sigma^2, \lambda/\sigma^2).
\end{aligned}
$$

The test statistic is given by

$$
\mathrm{PW} = E_+ \sum \frac{O_i(O_i - 1)}{E_i}. \tag{11.1}
$$

The alternative hypothesis of this test is that the O_i are distributed following a Negative Binomial distribution, as explained before and, therefore, this test can also be considered as a test of over-dispersion.

```
> pwtest <- pottwhitt.test(Observed ~ offset(log(Expected)),
+     as(nc, "data.frame"), "multinom", 999)
```

The asymptotic distribution of this statistic is Normal with mean $O_+(O_+-1)$ and variance $2nO_+(O_+ - 1)$, so a one-side test can be done as follows:

```
> Oplus <- sum(nc$Observed)
> 1 - pnorm(pwtest$t0, Oplus * (Oplus - 1), sqrt(2 * 100 *
+     Oplus * (Oplus - 1)))

[1] 0
```

Other tests for over-dispersion included in **DCluster** are the likelihood ratio test and some of the score tests proposed in Dean (1992). Although they are not described here, all these tests agree with the previous results obtained before and support the fact that the relative risks are not homogeneous and the observed cases are over-dispersed. Therefore, we have preferred to use a Negative Binomial to produce the simulations needed to assess the significance of some of the methods described in the remainder of this section. McMillen (2003) has addressed the importance of choosing the right model in a statistical analysis, and how autocorrelation can appear as a result of a wrong specification of the model.

In addition, Loh and Zhou (2007) discuss the effect of not accounting for extra-Poisson variation and sampling from the wrong distribution when detection of clusters of disease employs the spatial scan statistic (see Sect. 11.5.6). Loh and Zhou (2007) propose a correction based on estimating the distribution of the test statistics by sampling from a distribution that accounts for spatial correlation and other factors (for example, covariates). This approach produces more reliable p-values than the original test. Cressie and Read (1989) already mentioned that the Poisson model was not appropriate for the SIDS data due to the presence of over-dispersion and that other models that take it into account would be more appropriate.

In case of doubt, the reader is advised to assess the significance of a given test by using the Multinomial distribution. This is the standard procedure to assess the significance of the test statistic by Monte Carlo in this scenario. See Waller and Gotway (2004, pp. 202–203) for a discussion on this issue.

A first evaluation of the presence of clusters in the study region can be obtained by checking the spatial autocorrelation. Note that using the chi-square test, for example we can only detect that there are clear differences among the relative risks but not if there is any spatial structure in these differences. In other words, if neighbours tend to have similar (and higher) values. Note that a possible scenario is that of regions having significantly different (low and high) relative risks but with no spatial structure, in which the chi-square test will be significant but there will not be any spatial autocorrelation. This can happen if the scale of aggregation of the data is not taken properly into account or the scale of the risk factors does not exceed the scale of aggregation.

11.5.2 Moran's *I* Test of Spatial Autocorrelation

We have already discussed the use of Moran's I statistic to assess the presence of spatial autocorrelation. Here we apply Moran's I statistic to the SMR to account for the spatial distribution of the population. If we computed Moran's statistic for the O_i, we could find spatial autocorrelation only due to the spatial distribution of the underlying population, because it is well known that the higher the population, the higher the number of cases. Binary weights are used depending on whether two regions share a common boundary or not. Spatial autocorrelation is still found even after accounting for over-dispersion.

```
> col.W <- nb2listw(ncCR85, zero.policy = TRUE)
> moranI.test(Observed ~ offset(log(Expected)), as(nc,
+     "data.frame"), "negbin", 999, listw = col.W, n = length(ncCR85),
+     S0 = Szero(col.W))
```

```
Moran's I test of spatial autocorrelation

    Type of boots.: parametric
    Model used when sampling: Negative Binomial
    Number of simulations: 999
    Statistic:  0.2385172
    p-value :  0.001
```

11.5.3 Tango's Test of General Clustering

Tango (1995) proposed a similar test of global clustering by comparing the observed and expected number of cases in each region. He points out that different types of interactions between neighbouring regions can be considered and he proposes a measure of strength based on a decaying function of the distance between two regions.

Briefly, the statistic proposed by Tango is

$$T = (r - p)^{\mathrm{T}} A (r - p) \begin{cases} r^{\mathrm{T}} = [O_1/O_+, \ldots, O_n/O_+], \\ p^{\mathrm{T}} = [E_1/E_+, \ldots, E_n/E_+], \\ A = (a_{ij}) \text{ closeness matrix}, \end{cases} \tag{11.2}$$

where $a_{ij} = \exp\{-d_{ij}/\phi\}$ and d_{ij} is the distance between regions i and j, measured as the distance between their centroids. ϕ is a (positive) constant that reflects the strength of the dependence between areas and the scale at which the interaction occurs.

In our example, we construct the dependence matrix as suggested by Tango and, in addition, we take $\phi = 100$ to simulate a smooth decrease of the relationship between two areas as their relative distance increases. It is advisable to try different values of ϕ because this can have an important impact on the results and the significance of the test. Constructing this matrix in R is straightforward using some functions from package **spdep**, as shown in the following code below. In the computations the weights are globally re-scaled, but this does not affect the significance of the test since they all have simply been divided by the same constant. Furthermore, we have taken the approximate location of the county seats from `nc.sids` (columns `x` and `y`), which are in UTM (zone 18) projection. Note that using the centroids as the county seats – as computed by `coordinates(nc)` – may lead to slightly different coordinates and this may have an impact on the results of this and other tests.

```
> data(nc.sids)
> idx <- match(nc$NAME, rownames(nc.sids))
> nc$x <- nc.sids$x[idx]
> nc$y <- nc.sids$y[idx]
> coords <- cbind(nc$x, nc$y)
> dlist <- dnearneigh(coords, 0, Inf)
> dlist <- include.self(dlist)
> dlist.d <- nbdists(dlist, coords)
> phi <- 100
> col.W.tango <- nb2listw(dlist, glist = lapply(dlist.d,
+     function(x, phi) {
+         exp(-x/phi)
+     }, phi = phi), style = "C")
```

After computing the adjacency matrix we are ready to compute Tango's test of general clustering, which points out the presence of global clustering.

```
> tango.test(Observed ~ offset(log(Expected)), as(nc, "data.frame"),
+     "negbin", 999, listw = col.W.tango, zero.policy = TRUE)
```

```
Tango's test of global clustering
    Type of boots.: parametric
    Model used when sampling: Negative Binomial
    Number of simulations: 999
    Statistic:  0.000483898
    p-value :  0.049
```

11.5.4 Detection of the Location of a Cluster

So far we have considered methods that assess only the presence of heterogeneity of risks in the study area and give a general evaluation of the presence of clusters. To detect the actual location of the clusters present in the area a different approach must be followed. A useful family of methods that can help in this purpose are *scan statistics* (Hjalmars et al., 1996). These methods are based on a moving window that covers only a few areas each time and for which a test of clustering is carried out locally. By repeating this procedure throughout the study area, it will be possible to detect the locations of clusters of disease.

Scan methods usually differ in the way the window is defined, how it is moved over the area, and how the local test of clustering is carried. A recent review of these methods has appeared in *Statistics in Medicine* (Lawson et al., 2006). In this section we only refer to Openshaw's Geographical Analysis Machine (Openshaw et al., 1987) and Kulldorff's statistic (Kulldorff and Nagarwalla, 1995), because the latter is probably the first scan method proposed and the former is a widely established (and used) methodology.

11.5.5 Geographical Analysis Machine

Openshaw's Geographical Analysis Machine considers a regular grid of points $\{(x_i, y_i)\}_{k=1}^{p}$ over the study region at which a circular window is placed in turn. The test only considers the regions whose centroids are inside the window and it is based on comparing the total number of observed cases in the window (O_{k+}) to the total of expected cases in the window (E_{k+}) to assess if the latter is significantly high. Openshaw et al. (1987) define this test as the (one tailed) p-value of O_{k+}, assuming that it follows a Poisson distribution with mean E_{k+}. This procedure can be generalised and, if we have signs that the observed number of cases does not follow a Poisson distribution, the p-value can be obtained by simulation (Gómez-Rubio et al., 2005). Finally, if the current test is significant, the circle is plotted on the map. Alternatively, only the centre of each significant cluster can be plotted for the sake of simplicity and visualisation. Note also that we need to project the cluster centres back to longitude/latitude to be able to plot them on the map of North Carolina.

```
> sidsgam <- opgam(data = as(nc, "data.frame"), radius = 30,
+     step = 10, alpha = 0.002)
> gampoints <- SpatialPoints(sidsgam[, c("x", "y")] * 1000,
+     CRS("+proj=utm +zone=18 +datum=NAD27"))
```

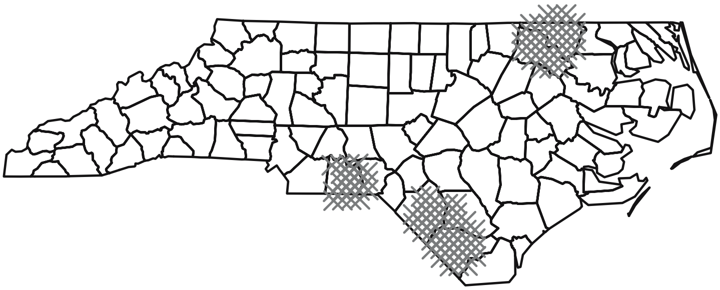

Fig. 11.17. Results of Openshaw's GAM. The dots represent the centre of the clusters

```
> library(rgdal)
> ll <- CRS("+proj=longlat +datum=NAD27")
> gampoints <- spTransform(gampoints, ll)
> gam.layout <- list("sp.points", gampoints)
```

When the complete area has been screened, we will probably have found several places where many overlapping clusters have been found, as shown in Fig. 11.17, where the centres of the clusters found have been plotted. This is due to the fact that the tests performed are not independent and, hence, very similar clusters (i.e. most of their regions are the same) are tested. That is the reason why Openshaw's GAM has been highly criticised by the statistical community and why, in order to maintain global significance, the significance level of the local tests should be corrected. Despite this, the GAM is still helpful as an exploratory method and to generate epidemiological hypotheses (Cromley and McLafferty, 2002).

11.5.6 Kulldorff's Statistic

To overcome this and other problems, Kulldorff and Nagarwalla (1995) developed a new test for the detection of clusters based on a window of variable size that considers only the most likely cluster around a given region. Kulldorf's statistic works with the regions within a given circular window and the overall relative risk in the regions inside the window is compared to that of the regions outside the window. This scan method is available in the SatScan™ software (http://www.satscan.org/), which includes enhancements to handle covariates, detect space-time clusters, and some other functionalities.

The null hypothesis, of no clustering, is that the two relative risks are equal, while the alternative hypothesis (clustering) is that the relative risk inside the window is higher. This is resolved by means of a likelihood ratio test, which has two main advantages. First, the most likely cluster can be detected as the window with the highest value of the likelihood ratio and, second, there is no need to correct the p-value because the simulations for different centres are independent (Waller and Gotway, 2004, p. 220). For a Poisson model, the expression of the test statistic is as follows:

$$\max_{z \in Z_i} \left(\frac{O_z}{E_z} \right)^{O_z} \left(\frac{O_+ - O_z}{E_+ - E_z} \right)^{O_+ - O_z} , \qquad (11.3)$$

where z is an element of Z_i, the set of all circles centred at region i. These circles are constructed so that only those that contain up to a fixed proportion of the total population are considered.

Note that, even though we select the most likely cluster around each region, it might not be significant. On the other hand, we may have more than one significant cluster, around two or more different regions, and that some clusters may overlap each other. When more than one cluster is found, we can consider the cluster with the lowest p-value as the *primary* or most prominent in the study region. *Secondary* clusters, that do not overlap with the former, may be considered too.

Loh and Zhou (2007) show that when data are over-dispersed, the *classical* spatial scan statistic will produce more false positives than the nominal significance level. To correct for this, they propose sampling from a different distribution that accounts for spatial correlation. The Negative Binomial can be used to account for the extra-variability, which may be caused by spatial autocorrelation coming from unmeasured covariates, and estimate the distribution of the test statistic under over-dispersion.

```
> mle <- calculate.mle(as(nc, "data.frame"), model = "negbin")
> thegrid <- as(nc, "data.frame")[, c("x", "y")]
> knresults <- opgam(data = as(nc, "data.frame"),
+     thegrid = thegrid, alpha = 0.05, iscluster = kn.iscluster,
+     fractpop = 0.15, R = 99, model = "negbin",
+     mle = mle)
```

The most likely cluster for the SIDS data set is shown in Fig. 11.18. The p-value is 0.04, which means that the cluster is significant.

The general procedure of application of this method includes testing each area as the centre of a possible cluster, although it can only be used on a single

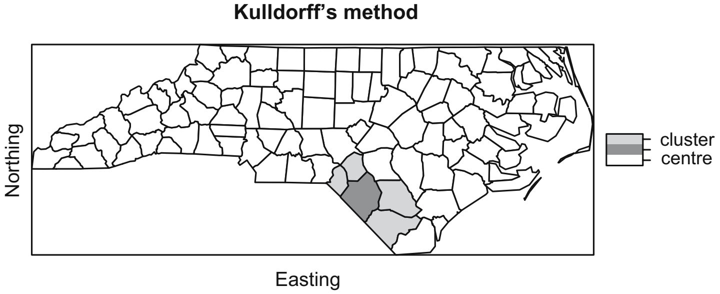

Fig. 11.18. Results of Kulldorff's test. The circles show the most likely cluster

point to test whether it is the centre of a cluster. This is specially helpful to assess the risk around putative pollution sources. Note that no assumption about the variation of the risk around the source is made. This is discussed in Sect. 11.5.7.

11.5.7 Stone's Test for Localised Clusters

As an alternative to the detection of clusters of disease, we may have already identified a putative pollution source and wish to investigate whether there is an increased risk around it. Stone (1988) developed a test that considers the alternative hypothesis of a descending trend around the pollution source. Basically, if we consider $\theta_{(1)}, \ldots, \theta_{(n)}$, the ordered relative risks of the regions according to their distances to the source, the test is as follows:

$$\begin{aligned} H_0 &: \theta_{(1)} = \ldots = \theta_{(n)} = \lambda, \\ H_1 &: \theta_{(1)} \geq \ldots \geq \theta_{(n)}. \end{aligned}$$

λ is the overall relative risk, which may be one if internal standardisation has been used. The test statistic proposed by Stone is the maximum accumulated risk up to a certain region:

$$\max_i \frac{\sum_{j=1}^i O_j}{\sum_{j=1}^i E_j}.$$

A word of caution must be given here because, as already discussed by many authors (Hills and Alexander, 1989, for example), focused tests should be employed before checking the data, because a bias is introduced when we try to use these tests on regions where an actual increased risk has been observed. In those cases, it will be more likely to detect a cluster than usual.

As an example, we try to assess whether there is an increased risk around Anson county, which has been spotted as an area of high risk. A call to `stone.stat` will give us the value of the test statistic and the number of regions for which the maximum accumulated risk is achieved. Later, we can use `stone.test` to compute the significance of this value.

```
> stone.stat(as(nc, "data.frame"), region = which(nc$NAME ==
+     "Anson"))

           region
4.726392 1.000000

> st <- stone.test(Observed ~ offset(log(Expected)), as(nc,
+     "data.frame"), model = "negbin", 99, region = which(nc$NAME ==
+     "Anson"))
> st
```

```
Stone's Test for raised incidence around locations

    Type of boots.: parametric
    Model used when sampling: Negative Binomial
    Number of simulations: 99
    Statistic:  4.726392
    p-value :  0.01
```

As the results show, the size of the cluster is 1 (just Anson county), which turns out to be highly significant.

11.6 Other Topics in Disease Mapping

Although we have tried to cover a wide range of analyses in this chapter, we have not been able to include other important topics, such as the detection of non-circular clusters (see, for example Tango and Takahashi, 2005), spatio-temporal disease mapping (see, for example Martínez-Beneito et al., 2008, and the references therein), or the joint modelling of several diseases (Held et al., 2005). Other data sets and models could be used by making the corresponding modifications to the R and WinBUGS code shown here. Some examples are availabe in Lawson et al. (2003). Furthermore, Banerjee et al. (2004) describe a number of other possible Bayesian analyses of spatial data and provide data and WinBUGS code in the associated website, which the reader should be able to reproduce using the guidelines provided in this chapter.