Spatiotemporal models in environmental epidemiology

Lecture 3 and 4

Disease mapping has a long history in epidemiology, and may be defined as the estimation and presentation of summary measures of health outcomes.

The aims of disease mapping include

- simple description,
- hypothesis generation,
- allocation of health care resources, assessment of inequalities, and
- estimation of background variability in underlying risk in order to place epidemiological studies in context.

We begin by noting a number of non-statistical issues

- In broad-scale studies (in particular international endeavors), data comparability is a major issue. Precise disease definition (via ICD codes) is also extremely important.
- Mortality data tend to be more reliable than incidence data, but the latter are of greater epidemiological interest in general.
- There is a trade-off when a geographical scale is chosen: larger geographical areas providing more stable rates and less problems of migration, but relative risk summaries may be distorted due to the large aggregation of individuals.
- If the relative risk shows marked variation within a particular area this information will be lost – if a particular subregion has a high relative risk then this will be diluted under aggregation; finding such subregions is not possible unless there are data available at a lower level of aggregation.

- The size of the areas chosen also determines the sort of questions that can be posed – larger areas are likely to offer greater contrasts in relative risks and exposures. Localized effects can only be detected with data at a smaller level of aggregation.
- Presentation:
    - Chloropleth (areas shaded) are the most popular kind of maps, but isopleth (contours) and cartograms (size of areas proportional to denominator), have also been used.
    - Choice of color is important – multiple colors can be confusing, shading with a single color can work well.
    - Cut-points should be chosen to be epidemiologically meaningful and convey as much information as possible.

We first consider mapping for area-level data.

Unfortunately there are well-documented difficulties with the mapping of raw estimates since, for small areas and rare diseases in particular, these estimates will be dominated by sampling variability.
For the model

$$Y_i \sim \text{Poisson}(E_i \theta_i)$$

the MLE is

$$\widehat{\theta}_i = \text{SMR}_i = \frac{Y_i}{E_i}$$

with variance

$$\text{var}(\widehat{\theta}_i) = \frac{\theta_i}{E_i}$$

so that areas with small E_i have high associated variance.

Example: Surveillance

We imagine separate monthly surveillance for each of three areas over a 10-year period.

We simulate data from the model

$$Y_i | \theta \sim_{ind} \text{Poisson}(E\theta),$$

$i = 1, ..., 120$, where the relative risk $\theta = 1$ in each case.

Recall that the MLE of the SMR in each time period is $\widehat{\theta}_i = Y_i/E$ with variance proportional to $1/E$ so that areas with small expected numbers have high variability.

The expected numbers differ in the three plots in Figure 1, and the resultant instability in the SMR is apparent.
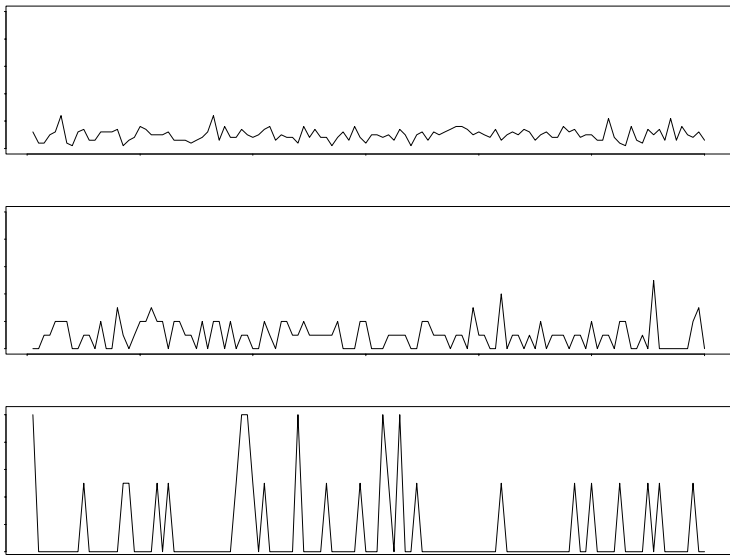
Figure: Simulations from the Poisson distribution under different expected

Example: Scottish Lip Cancer

Figure 2 shows the SMRs for the Scottish lip cancer data, and indicates a large spread with an increasing trend in the south-north direction.

The variance of the estimate is $\text{var}(\text{SMR}_i) = \text{SMR}_i/E_i$, which will be large if E_i is small.

For the Scottish data the expected numbers are highly variable, with range 1.1–88.7. This variability suggests that there is a good chance that the extreme SMRs are based on small expected numbers (many of the large, sparsely-populated rural areas in the north have high SMRs).

Figure 3 (left panel) shows the SMRs versus the estimated standard errors and clearly illustrates that the high SMRs have high associated standard error.
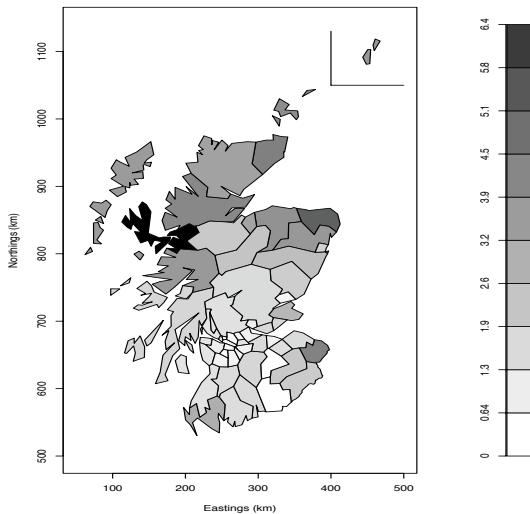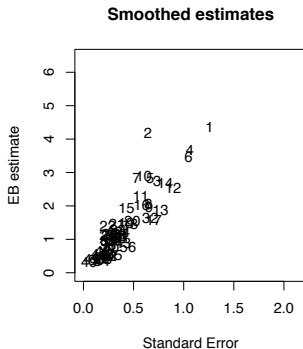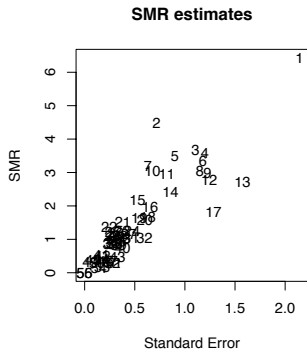
Figure: SMRs in 56 counties of Scotland.

Figure: Estimates versus standard errors for 56 counties of Scotland.

Maps showing p-values of exceedence of 1 are even less informative than maps of SMRs since although they account for sample size they do not show the extent of the risk. Hence areas with large populations may provide statistically significant SMRs, even for small exceedences of 1.

Smoothing Models

The above considerations led to methods being developed to *smooth* the SMRs using hierarchical/random effects models that use the data from the totality of areas to provide more reliable estimates in each of the constituent areas.

We first describe models that do not use spatial information before turning to models that exploit both spatial and non-spatial informtation.

Poisson-Gamma Model Without Covariates

We begin by describing a simple Poisson-Gamma two-stage model that offers analytic tractability and ease of estimation.

We assume there are no covariates and assume the first stage likelihood is given by

$$Y_i|\theta_i, \beta \sim_{ind} \text{Poisson}\left(\mu E_i \theta_i\right),\tag{1}$$

where μ is the overall relative risk, and reflects differences between the reference rates and the rates in the study region.

At the second stage the random effects θ_i are assigned a distribution. We initially assume that across the map the deviations of the relative risks from the mean, μ, are modelled by

$$\theta_i|\alpha \sim_{iid} \text{Ga}(\alpha, \alpha),\tag{2}$$

a gamma distribution with mean 1, and variance $1/\alpha$.

The advantage of this Poisson-gamma formulation is that the marginal distribution of $Y_i|\mu, \alpha$ (obtained by integrating out the random effects θ_i), is negative binomial.

Marginally, the mean and variance are given, respectively, by

$$
\begin{array}{rcl}
E[Y_i|\mu, \alpha] &=& E_i\mu \\
\mathsf{var}(Y_i|\mu, \alpha) &=& \mathsf{E}[Y_i|\mu, \alpha](1 + \mathsf{E}[Y_i|\mu, \alpha]/\alpha), \quad (3)
\end{array}
$$

so that the variance increases as a quadratic function of the mean, and the scale parameter α can accommodate different levels of "overdispersion".

This form is substantively more reasonable than the naive Poisson model; it is important to consider excess-Poisson variability resulting from unmeasured confounders, data anomalies in numerator and denominator, and model misspecification.

Empirical Bayes Estimation without Covariates

If μ and α were known then the posterior distribution of θ_i would be gamma.

Suppose we have estimates $\widehat{\mu}, \widehat{\alpha}$. Then the distribution is given by

$$\theta_i | \mathbf{y}, \widehat{\mu}, \widehat{\alpha} \sim \mathsf{Ga}(\widehat{\alpha} + y_i, \widehat{\alpha} + E_i \widehat{\mu}).$$

The relative risk (relative to the reference rates used in the expected numbers) is given by $\mathsf{RR}_i = \mu \theta_i$, and has mean

$$
\begin{aligned}
\widehat{RR}_i &= \widehat{\mu} \times \mathsf{E}[\theta_i | \mathbf{y}, \widehat{\mu}, \widehat{\alpha}] = \widehat{\mu} \left(\frac{\widehat{\alpha} + Y_i}{\widehat{\alpha} + \widehat{\mu} E_i} \right) \\
&= \mathsf{E}[\mathsf{RR}_i] \times (1 - w_i) + \mathsf{SMR}_i \times w_i,
\end{aligned}
\tag{4}
$$

a weighted combination of the prior estimate $\mathsf{E}[\mathsf{RR}_i] = \mu$, and the SMR in area i.

The *weight*

$$w_i = \frac{E_i \widehat{\mu}}{\widehat{\alpha} + E_i \widehat{\mu}}. \tag{5}$$

on the observed SMR increases as E_i increases so that for areas with large populations the estimate is dominated by the data.
If α is large then the random effects have a tight spread, and there is more shrinkage since SMRs that are far from unity are inconsistent with the total collection of estimates – the weight is small in this case. This behavior illustrates both the potential benefits and hazards of smoothing; the estimates will be less variable than the SMRs, but an outlying estimate that is not based on a large expected number, will be shrunk, and we may miss an important excess.

Poisson-Gamma Model with Covariates

With area-level covariates we have the model

$$Y_i|\theta_i, \beta \sim_{\mathbf{ind}} \text{Poisson}\left(\mu_{\mathbf{i}}\mathbf{E_i}\theta_{\mathbf{i}}\right),$$

At the second stage the random effects θ_i are assigned a distribution. We assume that across the map the deviations of the relative risks from the mean, μ_i, are modelled by

$$\theta_i|\alpha \sim_{iid} \text{Ga}(\alpha, \alpha),$$

a gamma distribution with mean 1, and variance $1/\alpha$.

Empirical Bayes Estimation with Covariates

Suppose we have estimates $\widehat{\beta}$, $\widehat{\alpha}$. Then the distribution is given by

$$\theta_i | \mathbf{y}, \widehat{\beta}, \widehat{\alpha} \sim \mathsf{Ga}(\widehat{\alpha} + y_i, \widehat{\alpha} + E_i \widehat{\mu}_i),$$

which has mean

$$\widehat{RR_i} = \mathsf{E}[\mathsf{RR}_i] \times (1 - w_i) + \mathsf{SMR}_i \times w_i, \tag{6}$$

a weighted combination of the prior estimate $\mathsf{E}[\mathsf{RR}_i] = \mu_i$, and the SMR in area i.

The *weight* is given by

$$w_i = \frac{E_i \widehat{\mu}_i}{\widehat{\alpha} + E_i \widehat{\mu}_i}. \tag{7}$$

In the right hand panel of Figure 3 we plot empirical Bayes estimates versus standard errors using a log-linear model in AFF.

One possibility for obtaining estimates $\widehat{\beta}$, $\widehat{\alpha}$ is to use maximum likelihood estimation over the marginal likelihood $\prod_{i=1}^{n} \Pr(Y_i | \beta, \alpha)$ (each term is a negative binomial).

Example: Scottish Lip Cancer

Figure 5 shows relative risk estimates from a variety of models, with the SMRs on the left (referenced as position 0).

At position 1 the empirical Bayes estimates obtained without the use of the covariate AFF are displayed.

The weights on the SMR, (5), range between 0.45 and 0.99, with median 0.83. For these data the residual variability is large.

The standard deviation of the random effects is $1/\sqrt{\alpha}$, and is estimated as 0.73, with 90% interval for residual relative risks (0.16,2.4).
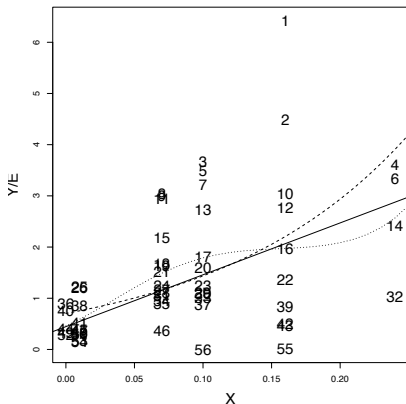
Figure: Plot of Y/E versus proportion in AFF, x. Solid line corresponds to a linear in x model; dashed line to a log link, linear in x model; and dotted line to log link, cubic in x model.
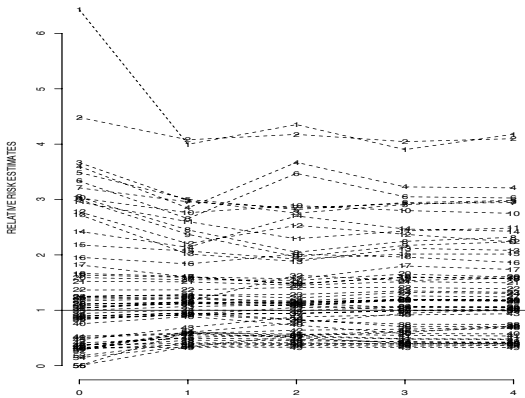
Figure:

Figure details: Relative risk estimates for Scottish lip cancer data:

0   denote the SMRs;
1   the empirical Bayes estimates without the use of AFF;
2   the empirical Bayes estimates with log link and a linear model in AFF;
3   the empirical Bayes estimates with a log-linear cubic model in AFF;
4   the fully Bayes estimates with a log-linear cubic model in AFF.

In position 2 gives EB estimates using a log-linear model in AFF,

$$\log \mu_i = \beta_0 + \beta_1 x_i.$$

The standard errors of the estimates are shown in the right hand panel of Figure 3. Four of the counties (4, 6, 14 and 32) have proportion in AFF equal to 0.24 (the highest value) and we see that the estimates for these counties are all moved upwards relative to the no covariate model (position 1) when the covariate is added to the model.
The latter is worrying, and we see the reason in Figure 4; the log-linear model (dashed line) does not fit the data well for large values of AFF. This suggests that we use a more flexible model; after some exploratory work we choose the cubic form

$$\log \mu_i = \beta_0 + \beta_1(x_i - \overline{x}) + \beta_2(x_i - \overline{x})^2 + \beta_3(x_i - \overline{x})^3 \tag{8}$$

Figure 4 shows that this cubic model provides a better fit to the data (dotted line), and in particular flattens off for larger values of x. With linear and cubic models the sd of the random effects are 0.58 and 0.53.

We might expect the standard deviation to be reduced in size when we add an important covariate but this does not have to happen.

In position 3 of Figure **??** the cubic estimates are plotted and we see that for counties 4 and 6 in particular the estimates are more reasonable.

In this study there are only six distinct AFF values and so one could treat AFF as a factor and smooth using only those counties with identical covariate values.

This example illustrates how smoothing is carried out via the covariates, and the importance of deciding how much local smoothing is appropriate. A similar issue is relevant to the extent and nature of spatial smoothing.

Review

- The aim is the provide stable relative risk estimates for area-level data.
- We have assumed that the relative risks arise from a common gamma distribution, which allows smoothing towards a common value.
- An empirical Bayes approach estimates the parameters of the negative binomial model (β and α) and then combines the gamma distribution with the data to obtain the empirical Bayes posterior distribution for the relative risks.

Drawing Maps in R

Stages for mapping:

1. Require a set of polygons for each of the constituent areas in the study region, each polygon defined by a set of $x - y$ coordinates.
2. Need to be able to draw a map using these polygons.
3. Data to be mapped needs to be spatially-referenced with a common set of labels/order as the polygons.
4. Need to be able to fill in the polygons of the map using the data.

Usually the number of polygons will be greater than the number of areas, because some areas will be made up of disjoint sub-areas (for example, islands).

The maps Library

```
> library(maps)
> map.text("county", "ohio")
> testdat <- runif(88) # need to read in the OhioMap function
> OhioMap(testdat,ncol=8,type="e",figmain="Ohio",lower=0,upper=2)
> OhioMap(testdat,ncol=8,type="e",figmain="Ohio random numbers",lower=0,upper=2)
> temp <- map("county","ohio")
> temp$names
 [1] "ohio,adams"      "ohio,allen"      "ohio,ashland"    "ohio,ashtabula"
 [5] "ohio,athens"     "ohio,auglaize"   "ohio,belmont"    "ohio,brown"
 [9] "ohio,butler"     "ohio,carroll"    "ohio,champaign"  "ohio,clark"
...
[77] "ohio,summit"     "ohio,trumbull"   "ohio,tuscarawas" "ohio,union"
[81] "ohio,van wert"   "ohio,vinton"     "ohio,warren"     "ohio,washington"
[85] "ohio,wayne"      "ohio,williams"   "ohio,wood"       "ohio,wyandot"
```

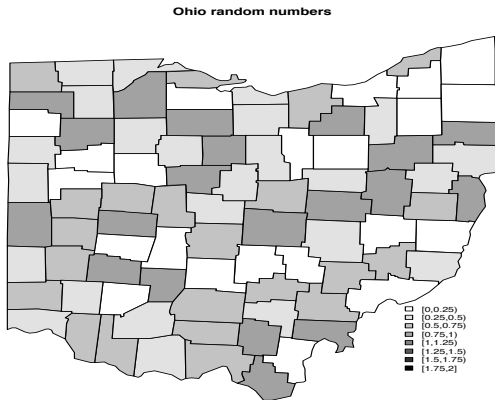Figure: Map of Ohio counties, with names.

Ohio random numbers

[0,0.25)
[0.25,0.5)
[0.5,0.75)
[0.75,1)
[1,1.25)
[1.25,1.5)
[1.5,1.75)
[1.75,2)

Figure: Map of Ohio random numbers, using the ÔhioMap function.

Example: Disease Mapping for Scotland

We make use of a mapping function that is on the class website:

```
PrettyPoly <- function(y, poly, nrepeats, ncut=1000,
nlevels=10, lower=NULL, upper=NULL )
```

with arguments:

- y the variable to be mapped
- poly the $x - y$ coordinates of the polygons, with different polygons separated by NAs.
- nrepeats a vector of the same length as y with each entry containing the number of repeats of the appropriate entry in y.
- ncut The number of grey-scale levels to convert y to.
- nlevels The number of grey levels to plot.
- lower The value (on the same scale as y) that white is assigned to.
- upper The value (on the same scale as y) that black is assigned to.

The following code produces Figure 8.

```
> library(MASS)
> source("PolyMap.R")
> source("scotdat.txt")
> SMR <- z$Y/z$E
> zp <- read.table("SMRSplusmap.txt")
> x <- zp[,2]/1000
> y <- zp[,3]/1000
> poly <- matrix(c(x,y),ncol=2)
> nrepeats <- c(3,1,1,1,1,3,1,2,2,1,5,1,1,1,1,1,1,1,1,1,1,2,8,1,1,1,1,1,1,1,
              1,1,1,1,1,1,2,1,2,1,1,1,1,1,1,1,1,1,2,2,1,1,1,1,1,1)
> PrettyPoly(SMR,poly,nrepeats=nrepeats)#,lower=0,upper=10)
```
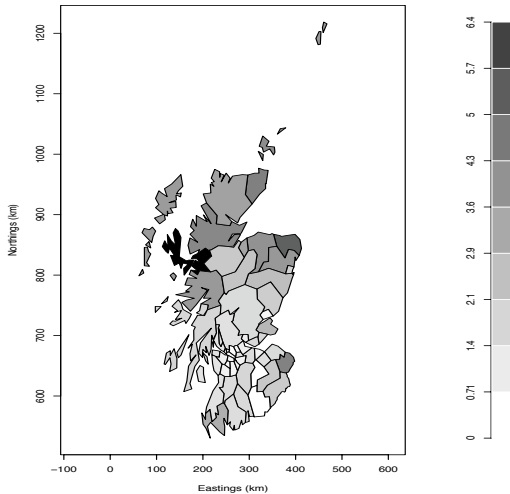
Figure: SMRs for Scottish counties.

The code below produces EB posterior mean estimates, the results are plotted in Figure 9. Notice that the extremes have been smoothed. It makes use of the function

```
eBayes <- function(Y,E,Xmat=NULL)
```

which takes as input, in addition to Y and E, an $n \times p$ matrix of ecological covariates where n is the number of areas and p is the number of covariates.

The outputs are:

- `RR` the ecological relative risk posterior mean estimates
- `RRmed` the ecological relative risk posterior median estimates
- `beta` the MLEs of the regression coefficients
- `alpha` the MLE of negative binomial dispersion parameter
- `SMR` the standardized moratility/morbifity ratio, Y/E.

```
> emp <- eBayes(z$Y,z$E,Xmat=cbind(z$X,z$X^2,z$X^3))
> postscript("Scotland_gamma.ps", horizontal=FALSE)
> PrettyPoly(emp$RR, poly, nrepeats=nrepeats,lower=0,upper=max(SMR))
```
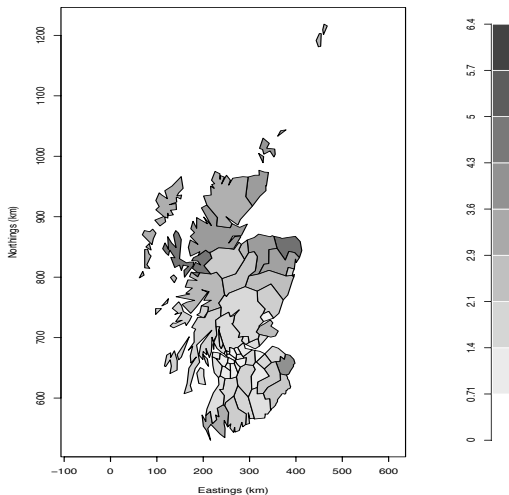
Figure: Empirical Bayes posterior mean estimates for Scottish counties.

As an alternative summary measure we plot the posterior medians in Figure 10. Produced using:

```
> PrettyPoly(emp$RRmed, poly, nrepeats=nrepeats,lower=0,upper=max(SMR))
```

Compared with 9, the estimates are higher, because the median exceeds the mean for a gamma distribution (for examples see Figure 11). So, for example, the areas with SMRs greater than 1 tend to have median estimates closer to the SMRs than the posterior means. Figure 11 gives the empirical Bayes posterior densities for areas 1, 2, 55 and 56 – the vertical line denotes the SMR. Arguments to the function EBpostdens are reasonably self-explanatory.

```
> xvals <- seq(0,15,.01)
> beta <- matrix(c(emp$beta),nrow=4,ncol=1)
> Xrow1 <- matrix(c(1,z$X[1],z$X[1]^2,z$X[1]^3),nrow=1,ncol=4)
> EBpostdens(z$Y[1],z$E[1],emp$alpha,emp$beta,Xrow1,
lower=0,upper=15,main="Area 1")
```
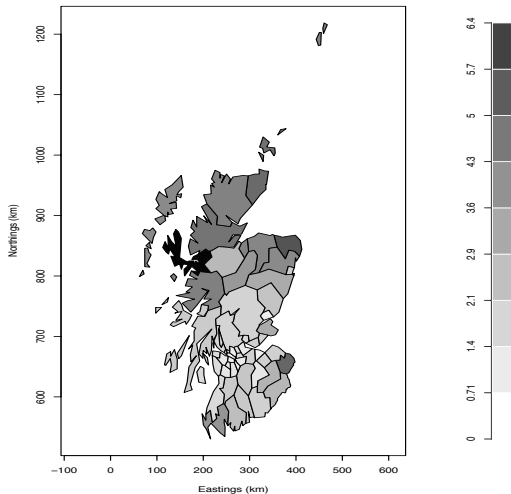
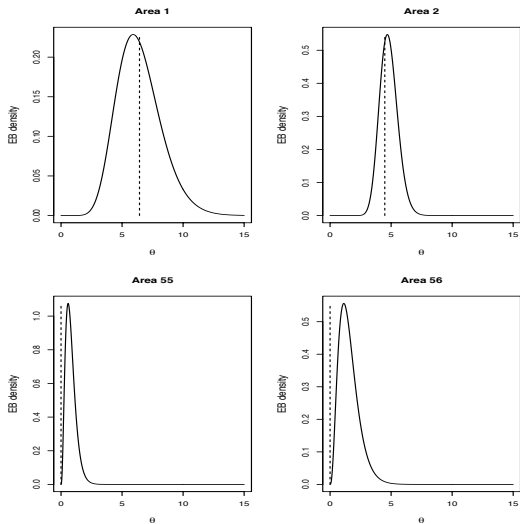Figure: Empirical Bayes posterior median estimates for Scottish counties.

Figure: Empirical Bayes densities for 4 counties.

Figure 12 gives the posterior probability that the relative risk exceeds the value 3 in each area, given the gamma global smoothing model. Arguments to the function `EBpostthresh` are reasonably self-explanatory.

```
> Xrow <- matrix(cbind(1,z$X,z$X^2,z$X^3),nrow=56,ncol=4)
> thresh3 <- EBpostthresh(z$Y,z$E,emp$alpha,emp$beta,Xrow,rrthresh=3)
> PrettyPoly(thresh3,poly,nrepeats,nlevels=11,lower=0,upper=1)
```

We see that the probabilities cover the full range of (0,1). Plots such as these are a very useful accompaniment to the raw SMRs and the smoothed estimates, since they reflect the uncertainty in the estimates. Code for posterior probability threshold function:

```
EBpostthresh <- function(Y,E,alpha,beta,Xrow=NULL,rrthresh){
if (is.null(Xrow)) Xrow <- matrix(rep(1,length(Y)),
nrow=length(Y),ncol=1)
mu <- as.numeric(exp(Xrow %*% beta))
thresh <- 1-pgamma(rrthresh,alpha+Y,(alpha+E*mu)/mu)
return(thresh=thresh)
}
```
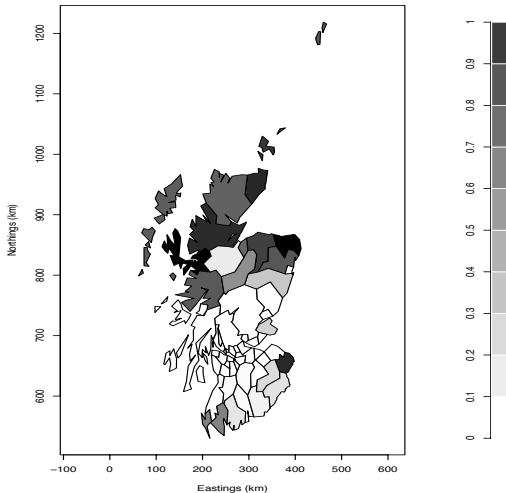
Figure: Empirical Bayes posterior probabilities that relative risk exceeds 3 in each county.

Summary of Smoothing Models for Disease Mapping

- The aim is to reduce the instability inherent in SMRs based on small expected numbers.

- This is achieved by fitting a random effects model which assumes that area-level deviations from the regression model arise from a probability distribution (e.g. gamma or lognormal).

- Care should be exercised in the regression model that is used, to make sure appropriate smoothing is being carried out (recall the inadequate log-linear model in the Scottish data).

- Comparing SMRs with smoothed estimates is important – if there are big changes, are they appropriate? i.e. Were the expected numbers small? Did the regression model fit this area well?

- Fitting can be carried out using empirical Bayes or full Bayes.

- *Empirical Bayes* – MLE used for regression parameters and variance parameters of random effects distribution (α in the gamma model).
  - Advantage: ease of fitting.
  - Disadvantages: cannot do spatial smoothing, not quite right statistically.
- *Full Bayes* – requires a prior distribution on regression parameters and variance parameters of random effects distribution.
  - Advantages: all uncertainties correctly accounted for, extends to spatial models.
  - Disadvantages: computation must be carried out with Markov chain Monte Carlo, which requires some experience.
- Posterior distributions for each area relative risk can be summarized in a number of ways, e.g. posterior mean, posterior median, posterior quantiles, posterior probability of exceedence of a threshold.

Poisson-Gamma Model

We now carry out a fully Bayesian analysis of the model for which empirical Bayes was used previously:

$$Y_i | \theta_i, \beta_0 \sim \text{Poisson}(E_i e^{\beta_0} \theta_i)$$
$$\theta_i \sim \text{Ga}(\alpha, \alpha)$$

We require priors for β_0 and α. For example:

$$\beta_0 \sim \text{N}(m, v)$$
$$\alpha \sim \text{Ga}(a, b)$$

with m, v, a, b picked to reflect beliefs about β_0 and α.

Empirical Bayes for Scotland

We recap on the previous analyses – this involved maximum likelihood estimation for β_0 and α in a negative binomial model and produced:

```
> emp0 <- eBayes(z$Y,z$E)
> emp0$beta
 0.3521065
> emp0$alpha
[1] 1.87949
> emp0$RR
 [1] 3.9973624 4.0791107 2.9802133 2.8467916 3.0025773 2.6545872 2.9590825
 [8] 2.4517687 2.3721492 2.7619805 2.6005515 2.2037872 2.0149301 2.1376464
...
[43] 0.6900960 0.4948910 0.4013614 0.5124617 0.5604849 0.4593902 0.3319144
[50] 0.3766186 0.6098460 0.5850639 0.4100864 0.3460232 0.3403845 0.6020789
> emp0$RRmed
 [1] 3.8755781 4.0458981 2.9034476 2.7600608 2.9434956 2.5655788 2.9237792
 [8] 2.3603697 2.2725880 2.7200177 2.5425312 2.0979757 1.8790820 2.0659710
...
[43] 0.6317935 0.4741200 0.3949723 0.4779112 0.5131326 0.4284178 0.3282190
[50] 0.3608116 0.5408883 0.5189084 0.3637163 0.3068970 0.2822885 0.4993176
```

Bayesian Analysis and WinBUGS

Inference for models with a spatial component is often not reliable using likelihood, and so Bayesian methods are commonly used. Unfortunately, most Bayesian models are not conducive to analytical analysis, and so are not available in standard software packages. WinBUGS is a package that allows very general Bayesian modeling; the GeoBUGS module contains a number of useful spatial models, and mapping facilities.

Dependent samples that are approximate draws from the posterior distribution are produced and summarized within WinBUGS.

The algorithm that produces these samples require (a Markov chain) a starting point for initialization. To lose dependence on this starting point, initial iterations of the algorithm are discarded (and not used for inference) – this is known as the *burn in*.

WinBUGS analysis of the Poisson-Gamma model

In the example that follows we specify a flat prior for β_0, and a Ga(1,1) prior for α.

The iterative algorithm is run for 10,000 iterations, with the first 4,000 discarded as "burn-in".

We summarize the posteriors for the relative risks:

$$\mathsf{RR}_i = \exp(\beta_0)\theta_i$$

and for β_0 and α. The posterior mean for β_0 is 0.36, compared to 0.35 under empirical Bayes, and the posterior mean for α is 1.79, compared to 1.88 under empirical Bayes.

Similarly the posterior means and posterior medians agree very closely.

```
model
{
  for (i in 1 : N) {
      Y[i]   ~ dpois(mu[i])
      mu[i] <- E[i]*exp(beta0)*theta[i]
      RR[i] <- exp(beta0)*theta[i]
      theta[i] ~ dgamma(alpha,alpha)
  }
# Priors
  alpha ~ dgamma(1,1)
  beta0 ~ dflat()
# Functions of interest:
  sigma.theta <- sqrt(1/alpha)    # standard deviation of non-spatial
  base <- exp(beta0)
}
```

```
DATA
list(N = 56,
  Y = c( 9, 39, 11, 9, 15, 8, 26, 7, 6, 20, 13, 5, 3, 8, 17, 9, 2, 7,
  9, 7, 16, 31, 11, 7, 19, 15, 7, 10, 16, 11, 5, 3, 7, 8, 11, 9, 11,
  8, 6, 4, 10, 8, 2, 6, 19, 3, 2, 3, 28, 6, 1, 1, 1, 1, 0, 0), E = c(
  1.4, 8.7, 3.0, 2.5, 4.3, 2.4, 8.1, 2.3, 2.0, 6.6, 4.4, 1.8, 1.1,
  3.3, 7.8, 4.6, 1.1, 4.2, 5.5, 4.4, 10.5,22.7, 8.8, 5.6,15.5,12.5,
  6.0, 9.0,14.4,10.2, 4.8, 2.9, 7.0, 8.5,12.3,10.1,12.7, 9.4, 7.2,
  5.3, 18.8,15.8, 4.3,14.6,50.7, 8.2, 5.6, 9.3,88.7,19.6, 3.4, 3.6,
  5.7, 7.0, 4.2, 1.8))

INTIAL ESTIMATES
list(alpha = 1, beta0 = 0,
  theta=c(1,1,1,1,1,1,1,1,1,1,1,1,1,
  1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
  1,1,1,1,1,1,1,1,1,1,1,1))
```

```
node     mean    sd       MC error  2.5%     median   97.5%    start    sample
RR[1]    4.07    1.297    0.01877   1.959    3.92     7.001    4000     6001
RR[2]    4.105   0.6469   0.00864   2.938    4.068    5.48     4000     6001
RR[3]    3.006   0.858    0.01159   1.607    2.915    4.937    4000     6001
RR[4]    2.875   0.8995   0.01019   1.391    2.773    4.886    4000     6001
RR[5]    3.016   0.7406   0.01114   1.754    2.955    4.668    4000     6001
RR[6]    2.68    0.8865   0.01325   1.227    2.568    4.696    4000     6001
RR[7]    2.975   0.5666   0.00830   1.994    2.929    4.236    4000     6001
RR[8]    2.476   0.8492   0.01224   1.082    2.379    4.412    4000     6001
....
RR[49]   0.3321  0.06051  7.88E-4   0.2261   0.3286   0.4612   4000     6001
RR[50]   0.3685  0.1334   0.00162   0.1603   0.3522   0.6725   4000     6001
RR[51]   0.6     0.3539   0.00424   0.1112   0.5327   1.45     4000     6001
RR[52]   0.5702  0.3425   0.00519   0.1034   0.5017   1.4      4000     6001
RR[53]   0.4021  0.2446   0.00316   0.07137  0.3546   0.9934   4000     6001
RR[54]   0.3327  0.2042   0.00227   0.05706  0.2924   0.8143   4000     6001
RR[55]   0.3259  0.2533   0.00345   0.02491  0.2646   0.9605   4000     6001
RR[56]   0.5814  0.4538   0.00636   0.04737  0.4723   1.745    4000     6001
alpha    1.79    0.3985   0.00792   1.129    1.753    2.682    4001     6000
beta0    0.3567  0.1188   0.00591   0.1315   0.353    0.5966   4000     6001
```

Summary of Smoothing Models for Disease Mapping

- The aim is to reduce the instability inherent in SMRs based on small expected numbers.

- This is achieved by fitting a random effects model which assumes that area-level deviations from the regression model arise from a probability distribution (e.g. gamma or lognormal).

- Care should be exercised in the regression model that is used, to make sure appropriate smoothing is being carried out (recall the inadequate log-linear model in the Scottish data).

- Comparing SMRs with smoothed estimates is important – if there are big changes, are they appropriate? i.e. Were the expected numbers small? Did the regression model fit this area well?

- Fitting can be carried out using empirical Bayes or full Bayes.

- *Empirical Bayes* – MLE used for regression parameters and variance parameters of random effects distribution (α in the gamma model).
  - Advantage: ease of fitting.
  - Disadvantages: cannot do spatial smoothing, not quite right statistically.
- *Full Bayes* – requires a prior distribution on regression parameters and variance parameters of random effects distribution.
  - Advantages: all uncertainties correctly accounted for, extends to spatial models.
  - Disadvantages: computation must be carried out with Markov chain Monte Carlo, which requires some experience.
- Posterior distributions for each area relative risk can be summarized in a number of ways, e.g. posterior mean, posterior median, posterior quantiles, posterior probability of exceedence of a threshold.

Poisson-Lognormal Model

The Poisson-gamma model offers analytic tractability, but does not easily allow the incorporation of spatial random effects.

A Poisson-lognormal non-spatial random effect model is given by:

$$Y_i | \beta, V_i \sim_{ind} \text{Poisson}(E_i \mu_i e^{V_i}) \quad V_i \sim_{iid} N(0, \sigma_v^2) \tag{9}$$

where V_i are area-specific random effects that capture the residual or unexplained (log) relative risk of disease in area i, $i = 1, ..., n$.

Whereas in the Poisson-Gamma model we have $\theta \sim \text{Ga}(\alpha, \alpha)$, here we have $\theta = e^{V_i} \sim \text{LogNormal}(0, \sigma^2)$.

Model (9) does not give a marginal distribution of known form, but does naturally lead to the addition of spatial random effects.

The marginal variance is of the same quadratic form as (3). Empirical Bayes is not so convenient for this model, and so we resort to a fully Bayesian approach for which we need to specify prior distributions.

Prior Choice for Non-Spatial Model

We need to specify priors for:

- The regression coefficients β.
- The variance of the random effets σ_v^2.

For a rare disease, a log-linear link is a natural choice:

$$\log \mu(\mathbf{x}_i, \beta) = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij},$$

where x_{ij} is the value of the j-th covariate in area i.

For regression parameters $\beta = (\beta_0, \beta_1, ..., \beta_J)$, an improper prior

$$p(\beta) \propto 1$$

may often be used, but in very circumstances such a choice may lead to an improper posterior.

If there are a large numbers of covariates, or high dependence amongst the elements of $\mathbf{x}$, then more informative priors will be beneficial.

Lognormal Priors

It is convenient to specify lognormal priors for positive parameters $\exp(\beta_j)$, since one may specify two quantiles of the distribution, and directly solve for the two parameters of the lognormal.

Denote by $\text{LN}(\mu, \sigma)$ the lognormal distribution for a generic parameter θ with $\mathsf{E}[\log \theta] = \mu$ and $\text{var}(\log \theta) = \sigma^2$, and let θ_1 and θ_2 be the q_1 and q_2 quantiles of this prior. Then it is straightforward to show that

$$
\mu = \log(\theta_1) \left(\frac{z_{q_2}}{z_{q_2} - z_{q_1}} \right) - \log(\theta_2) \left(\frac{z_{q_1}}{z_{q_2} - z_{q_1}} \right), \ \sigma = \frac{\log(\theta_1) - \log(\theta_2)}{z_{q_1} - z_{q_2}}.
$$

As an example, suppose that for the ecological relative risk e^{β_1} we believe there is a 50% chance that the relative risk is less than 1 and a 95% chance that it is less than 5; with $q_1 = 0.5, \theta_1 = 1.0$ and $q_2 = 0.95, \theta_2 = 5.0$, we obtain lognormal parameters $\mu = 0$ and $\sigma = \log 5/1.645 = 0.98$.

Prior for σ_v^2

Bottom line: the priors $\sigma_v^{-2} \sim \mathsf{Ga}(1, 0.0260)$ or $\sigma_v^{-2} \sim \mathsf{Ga}(0.5, 0.0005)$ will often be suitable in a mapping context, though sensitivity of the results to the specification should be carried out, particularly if the number of areas is not large.

Prior for σ_v^2 †

It is not straightforward to specify a prior for σ_v, which represents the standard deviation of the log residual relative risks, a difficult parameter to interpret.

We specify a gamma prior $\mathsf{Ga}(a, b)$ for the precision $\tau_v = 1/\sigma_v^2$. The choice of a gamma distribution is convenient since it produces a marginal distribution for the residual relative risks in closed form.

Specifically the two-stage model

$$V_i|\sigma_v \sim_{iid} N(0, \sigma_v^2), \quad \tau_v = \sigma_v^{-2} \quad \sim \quad \mathsf{Ga}(a, b)$$

produces a marginal distribution for V_i which is $t_d(0, \lambda^2)$, a Student's t distribution with $d = 2a$ degrees of freedom, location zero, and scale b/a; this is equivalent to the residual relative risks following a log t distribution.

We specify the range $\exp(\pm R)$ within which the residual relative risks lie with probability q, and use the relationship $\pm t_{q/2}^d \sqrt{b/a} = \pm R$, where t_q^d is the q-th quantile of a Student t random variable with d degrees of freedom, to give $a = d/2$, $b = R^2 d/2(t_{q/2}^d)^2$.

For example, if we assume *a priori* that the residual relative risks follow a log Student t distribution with 2 degrees of freedom, with 95% of these risks falling in the interval (0.5,2.0), we obtain the prior, $\tau_v \sim \mathsf{Ga}(1, 0.0260)$, an exponential distribution.

In terms of σ_v this results in (2.5%, 97.5%) quantiles of (0.084,1.01) with posterior median 0.19.

Non-Spatial Analysis of the Scottish Lip Cancer Data

We now report a fully Bayesian version of the normal model, (9), with log-linear cubic model.

The covariates are centered here in order to reduce dependence in the parameter estimates, which reduces the computational burden; this model was fitted using so-called Markov chain Monte Carlo via the WinBUGS software.

Flat priors were placed on $\beta_0, \beta_1, \beta_2, \beta_3$ and the previously-discussed gamma prior, $\text{Ga}(1, 0.0260)$, was assumed for σ_v^{-2}.

WinBUGS code

Below we give code for fitting the cubic log-linear Poisson-lognormal model.

We define the random variables $\exp(\pm 1.96 \times \sigma_v)$ as the endpoints of a 95% interval for the residual relative risks.

In Figure 5 we see that the estimates under the empirical Bayes gamma and fully Bayesian normal model, at positions 3 and 4 respectively, each with cubic mean model, are very similar, illustrating that the most important aspect is not the inferential method or the choice of gamma or lognormal random effects, but the judicious choice of the covariate model.

```
model {
  for (i in 1 : N) {
       Y[i]    ˜ dpois(mu[i])
       X1c[i] <- X[i]-mean(X[1:N])
       X2c[i] <- X1c[i]*X1c[i]
       X3c[i] <- X1c[i]*X1c[i]*X1c[i]
       log(mu[i]) <- log(E[i]) + beta0 +
                     beta1*X1c[i] + beta2*X2c[i] + beta3*X3c[i] + V[i]
       RR[i] <- exp(beta0 + beta1*X1c[i] + beta2*X2c[i]+ beta3*X3c[i] + V[i])
       V[i]    ˜ dnorm(0,tau.V)
  }
# The gamma prior corresponds to df=2, q=0.95, R=log 2.
  tau.V ˜ dgamma(1,0.0260)
  beta0 ˜ dflat()
  beta1 ˜ dflat()
  beta2 ˜ dflat()
  beta3 ˜ dflat()
# Functions of interest:
  sigma.V <- sqrt(1/tau.V)        # standard deviation of non-spatial
  RRRlo <- exp(-1.96*sigma.V)
  RRRhi <- exp(1.96*sigma.V) }
```
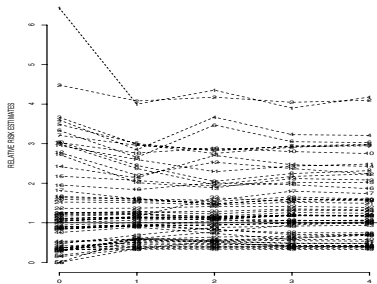
Figure: Features of Washington state, created using a GIS.

- In general we might expect residual relative risks in areas that are "close" to be more similar than in areas that are not "close".
- We would like to exploit this information in order to provide more reliable relative risk estimates in each area.
- This is analogous to the use of a covariate x, in that areas with similar x values are likely to have similar relative risks.
- Unfortunately the modelling of spatial dependence is much more difficult since spatial location is acting as a surrogate for unobserved covariates.
- We need to choose an appropriate spatial model, but do not directly observe the covariates whose effect we are trying to mimic.