Lecture 2
Spatial analysis of small area disease risk

MSc in Epidemiology @ Imperial College London

25 February 2019

1. Introduction

2. Spatial structure
   - Intrinsic ICAR model
   - BYM model

3. Ecological regression
   - Poisson regression with random effects
   - Childhood leukaemia and benzene

4. Joint analysis of several diseases
   - Shared component models
   - Oral cavity and oesophagus cancers

Disease mapping - background

- To summarise spatial and spatio-temporal variation in disease risk
- Rare disease and/or small areas: Poisson framework

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

where E_i = expected nb of cases, λ_i RR in area i

Non smoothed estimates of the RR

$$\text{SMR}_i = \frac{O_i}{E_i}, \text{ and } \hat{\text{Var}}(\text{SMR}_i) = \frac{O_i}{E_i^2}$$

- very imprecise: areas with small E_i have high associated variance
- estimated independently: does not account for spatial correlation

Disease mapping - background

- To summarise spatial and spatio-temporal variation in disease risk
- Rare disease and/or small areas: Poisson framework

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

where E_i = expected nb of cases, λ_i RR in area i

Non smoothed estimates of the RR

$$\text{SMR}_i = \frac{O_i}{E_i}, \text{ and } \hat{\text{Var}}(\text{SMR}_i) = \frac{O_i}{E_i^2}$$

- very imprecise: areas with small E_i have high associated variance
- estimated independently: does not account for spatial correlation

Disease mapping - background

- To summarise spatial and spatio-temporal variation in disease risk
- Rare disease and/or small areas: Poisson framework

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

where E_i = expected nb of cases, λ_i RR in area i

Non smoothed estimates of the RR

$$\text{SMR}_i = \frac{O_i}{E_i}, \text{ and } \hat{\text{Var}}(\text{SMR}_i) = \frac{O_i}{E_i^2}$$

- very imprecise: areas with small E_i have high associated variance
- estimated independently: does not account for spatial correlation

Disease mapping - background

- To summarise spatial and spatio-temporal variation in disease risk
- Rare disease and/or small areas: Poisson framework

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

where E_i = expected nb of cases, λ_i RR in area i

Non smoothed estimates of the RR

$$\text{SMR}_i = \frac{O_i}{E_i}, \text{ and } \hat{\text{V}}\text{ar}(\text{SMR}_i) = \frac{O_i}{E_i^2}$$

- very imprecise: areas with small E_i have high associated variance
- estimated independently: does not account for spatial correlation

⇒ Hierarchical modelling to smooth the rates

Smoothed estimates of the RR (non spatial) I

Poisson-logNormal model

$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log \lambda_i &= \alpha + V_i \\ V_i &\sim \text{Normal}(0, \sigma_v^2) \end{aligned}$$

Priors on the between-area variance σ_v^2 and the mean log relative risk α

where

- O_i, E_i: observed and expected nb of cases in area i (known data)

Smoothed estimates of the RR (non spatial) I

Poisson-logNormal model

$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log \lambda_i &= \alpha + V_i \\ V_i &\sim \text{Normal}(0, \sigma_v^2) \end{aligned}$$

Priors on the between-area variance σ_v^2 and the mean log relative risk α

where

- O_i, E_i: observed and expected nb of cases in area i (known data)
- V_i: area-specific random effects to take into account overdispersion
  - excess variation in the observed counts due to random noise
  - latent variable which captures the effects of unknown or unmeasured area level covariates

  ⇒ global smoothing

Smoothed estimates of the RR (non spatial) I

Poisson-logNormal model

$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log \lambda_i &= \alpha + V_i \\ V_i &\sim \text{Normal}(0, \sigma_v^2) \end{aligned}$$

Priors on the between-area variance σ_v^2 and the mean log relative risk α

where

- O_i, E_i: observed and expected nb of cases in area i (known data)
- V_i: area-specific random effects to take into account overdispersion
  - excess variation in the observed counts due to random noise
  - latent variable which captures the effects of unknown or unmeasured area level covariates

  ⇒ global smoothing
- $\lambda_i = \exp(\alpha + V_i)$: unknown RR in area i compared with expected risk based on age and sex of population

Smoothed estimates of the RR (non spatial) II

- Poisson-logNormal model based on the assumption that the observations in the data set are identically distributed and independent $\Rightarrow$ Independence makes much of the mathematics tractable

Smoothed estimates of the RR (non spatial) II

- Poisson-logNormal model based on the assumption that the observations in the data set are identically distributed and independent
  $\Rightarrow$ Independence makes much of the mathematics tractable

- However, data that occur close together in space are likely to be correlated
  $\Rightarrow$ Dependence between cases is a more realistic assumption

Smoothed estimates of the RR (non spatial) II

- Poisson-logNormal model based on the assumption that the observations in the data set are identically distributed and independent
  $\Rightarrow$ Independence makes much of the mathematics tractable

- However, data that occur close together in space are likely to be correlated
  $\Rightarrow$ Dependence between cases is a more realistic assumption

- Ignoring spatial dependence can lead to biased and inefficient inference

Smoothed estimates of the RR (non spatial) II

- Poisson-logNormal model based on the assumption that the observations in the data set are identically distributed and independent
  ⇒ Independence makes much of the mathematics tractable

- However, data that occur close together in space are likely to be correlated
  ⇒ Dependence between cases is a more realistic assumption

- Ignoring spatial dependence can lead to biased and inefficient inference

- Interest in estimating the relationship between location and outcome

  ⇒ Smooth in space: prior distribution for the random effects should allow for spatial correlation and help uncover spatial patterns

1 Introduction

2 Spatial structure
- Intrinsic ICAR model
- BYM model

3 Ecological regression
- Poisson regression with random effects
- Childhood leukaemia and benzene

4 Joint analysis of several diseases
- Shared component models
- Oral cavity and oesophagus cancers

A conditional spatial model

- Specify a distribution which takes into account information on neighbouring areas (adjacency matrix)
- Rule for determining the neighbours of each area: most common based on common boundary
- Estimate spatial random effect for each area as if we knew the values of the spatial random effects in neighbouring areas
- Use of conditional autoregressive distributions (we are conditioning on knowing the neighbours)

Intrinsic ICAR model (Besag et al. 1991)

$$U \sim \text{ICAR}(W, \sigma_u^2) \quad \Leftrightarrow \quad U_i | U_{j \ j \neq i} \sim \text{Normal}(u_i, \sigma_i^2)$$

Let ∂_i = set of areas adjacent to i

$$\text{E}(U_i|U_j) = u_i \ = \ \frac{\sum_{j \in \partial_i} U_j}{n_i} = \text{mean value of } U_j \text{ amongst neighbours}$$

$$\text{Var}(U_i|U_j) = \sigma_i^2 \ = \ \frac{\sigma_u^2}{n_i} \text{ where } n_i = \text{number of neighbours}$$

Intrinsic ICAR model (Besag et al. 1991)

$$U \sim \mathrm{ICAR}(W, \sigma_u^2) \quad \Leftrightarrow \quad U_i | U_{j\ j \neq i} \sim \mathrm{Normal}(u_i, \sigma_i^2)$$

Let $\partial_i =$ set of areas adjacent to i

$$\mathrm{E}(U_i | U_j) = u_i \quad = \quad \frac{\sum_{j \in \partial_i} U_j}{n_i} = \text{mean value of } U_j \text{ amongst neighbours}$$

$$\mathrm{Var}(U_i | U_j) = \sigma_i^2 \quad = \quad \frac{\sigma_u^2}{n_i} \text{ where } n_i = \text{number of neighbours}$$

- U_i is smoothed towards mean risk **in a set of neighbouring areas**

⇒ Local smoothing

Intrinsic ICAR model (Besag et al. 1991)

$$U \sim \text{ICAR}(W, \sigma_u^2) \quad \Leftrightarrow \quad U_i | U_{j\ j \neq i} \sim \text{Normal}(u_i, \sigma_i^2)$$

Let ∂_i = set of areas adjacent to i

$$\text{E}(U_i | U_j) = u_i \quad = \quad \frac{\sum_{j \in \partial_i} U_j}{n_i} = \text{mean value of } U_j \text{ amongst neighbours}$$

$$\text{Var}(U_i | U_j) = \sigma_i^2 \quad = \quad \frac{\sigma_u^2}{n_i} \ \text{ where } \ n_i = \text{number of neighbours}$$

- U_i is smoothed towards mean risk **in a set of neighbouring areas**

⇒ Local smoothing

- Conditional variance inversely proportional to the number of neighbours (so more neighbours, less variability)

Intrinsic ICAR model (Besag et al. 1991)

$$U \sim \text{ICAR}(W, \sigma_u^2) \quad \Leftrightarrow \quad U_i | U_{j \; j \neq i} \sim \text{Normal}(u_i, \sigma_i^2)$$

Let $\partial_i =$ set of areas adjacent to i

$$\text{E}(U_i | U_j) = u_i \quad = \quad \frac{\sum_{j \in \partial_i} U_j}{n_i} = \text{mean value of } U_j \text{ amongst neighbours}$$

$$\text{Var}(U_i | U_j) = \sigma_i^2 \quad = \quad \frac{\sigma_u^2}{n_i} \text{ where } \; n_i = \text{number of neighbours}$$

- U_i is smoothed towards mean risk **in a set of neighbouring areas**

$\Rightarrow$ Local smoothing

- Conditional variance inversely proportional to the number of neighbours (so more neighbours, less variability)
- More generally:

$$u_i = \frac{\sum_j w_{ij} U_j}{\sum_j w_{ij}}, \; \sigma_i^2 = \frac{\sigma_u^2}{\sum_j w_{ij}}$$

and if $w_{ij} = 1$ for $j \in \partial_i$, 0 otherwise, → as above

Intrinsic ICAR model II

- ICAR model defined above is improper: the overall mean of the U_i is not defined. So additional constraints need to be imposed:
  - **sum-to-zero constraint**: $\sum_i U_i = 0$
  - improper prior on α (dflat)

- The parameter σ_u^2 represents the **conditional** variance of the random effects (and not the marginal one) and its magnitude determines the amount of spatial variation

- No closed-form expression available for the **marginal** between-area variance of the spatial effects → estimate marginal spatial variance empirically

$$s^2_{\text{u.marginal}} = \sum_i (U_i - \overline{U})^2/(N-1)$$

The BYM model

- Besag, York and Mollie (BYM) recommend combining the ICAR prior and the standard normal prior to allow for both
  - spatially unstructured latent covariates V
  - spatially correlated latent covariates U

Convolution or BYM model

$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log \lambda_i &= \alpha + V_i + U_i \\ V_i &\sim \text{Normal}(0, \sigma_v^2) \\ \mathbf{U} &\sim \text{ICAR}(W, \sigma_u^2) \end{aligned}$$

Priors (vague, non informative): $\sigma_v^2, \sigma_u^2, \alpha$

Posterior quantification of spatial and unstructured variability in the BYM model

- σ_v^2 (unstructured heterogeneity variance) and σ_u^2 (spatial variance) NOT directly comparable
  - σ_v^2 reflects **marginal** variability of the unstructured REs between areas
  - σ_u^2/n_i reflects **conditional** variance of the spatial effect in area i conditional on values of neighbouring spatial effects

Posterior quantification of spatial and unstructured variability in the BYM model

- σ_v^2 (unstructured heterogeneity variance) and σ_u^2 (spatial variance) NOT directly comparable
  - σ_v^2 reflects **marginal** variability of the unstructured REs between areas
  - σ_u^2/n_i reflects **conditional** variance of the spatial effect in area i conditional on values of neighbouring spatial effects
- Estimate **marginal** between-area variance of the spatial effects *empirically:*

$$s^2_{\text{u.marginal}} = \sum_i (U_i - \overline{U})^2/(N-1)$$

- Relative contribution of spatial vs. unstructured heterogeneity:

$$\text{frac}_{\text{spatial}} = \sigma^2_{\text{u.marginal}}/(\sigma^2_{\text{u.marginal}} + \sigma^2_v)$$

  - $\text{frac}_{\text{spatial}} \to 1 \Rightarrow$ spatial heterogeneity dominates
  - $\text{frac}_{\text{spatial}} \to 0 \Rightarrow$ unstructured heterogeneity dominates

Use the BYM model with care!

- ICAR model defined above is improper: the overall mean of the U_i is not defined. So additional constraints need to be imposed:
  - **sum-to-zero constraint**: $\sum_i U_i = 0$ (implemented in OpenBUGS)
  - improper prior on α (dflat)
- Inference may be sensitive to choice of hyperprior for the random effects variance or precision parameters (σ_u^2 and σ_v^2)
- Interpretation of the variance parameters (conditional, marginal)
- Total variation of the log residual RR $= \sigma_v^2 + s^2_{\text{u.marginal}}$

Neighbourhood structure

→ Adjacency matrix

- An area cannot be specified as its own neighbour
- Adjacency matrix must be symmetric
- Islands (i.e single areas with no neighbours) are ignored and OpenBUGS fixes the value of U_i to zero if area i is an island
- The map must be fully connected

  There cannot be one or more groups of areas that are completely separated (i.e. have no neighbours in common) with the rest of the map. For a disconnected map, the overall mean and the individual spatial random effects are not well identified

Lung cancer incidence in males, 1985-2009, England and Wales

Here we replace the unstructured Normal random effects prior for the log relative risks by a convolution (CAR + unstructured Normal) prior:

$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log \lambda_i &= \alpha + V_i + U_i \\ V_i &\sim \text{Normal}(0, \sigma_v^2) \\ \mathbf{U} &\sim \text{ICAR}(W, \sigma_u^2) \end{aligned}$$

- **Data:**
- **Priors:**
- **Parameters of interest:**

Lung cancer incidence in males, 1985-2009, England and Wales

Here we replace the unstructured Normal random effects prior for the log relative risks by a convolution (CAR + unstructured Normal) prior:

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\lambda_i E_i) \\
\log \lambda_i &= \alpha + V_i + U_i \\
V_i &\sim \text{Normal}(0, \sigma_v^2) \\
\mathbf{U} &\sim \text{ICAR}(W, \sigma_u^2)
\end{aligned}
$$

- **Data:** O and E, observed and expected cases, W adjacency matrix
- **Priors:**
- **Parameters of interest:**

Lung cancer incidence in males, 1985-2009, England and Wales

Here we replace the unstructured Normal random effects prior for the log relative risks by a convolution (CAR + unstructured Normal) prior:

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\lambda_i E_i) \\
\log \lambda_i &= \alpha + V_i + U_i \\
V_i &\sim \text{Normal}(0, \sigma_v^2) \\
\mathbf{U} &\sim \text{ICAR}(W, \sigma_u^2)
\end{aligned}
$$

- **Data:** O and E, observed and expected cases, W adjacency matrix
- **Priors:** σ_v^2, σ_u^2, α
- **Parameters of interest:**

Lung cancer incidence in males, 1985-2009, England and Wales

Here we replace the unstructured Normal random effects prior for the log relative risks by a convolution (CAR + unstructured Normal) prior:

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\lambda_i E_i) \\
\log \lambda_i &= \alpha + V_i + U_i \\
V_i &\sim \text{Normal}(0, \sigma_v^2) \\
\mathbf{U} &\sim \text{ICAR}(W, \sigma_u^2)
\end{aligned}
$$

- **Data:** O and E, observed and expected cases, W adjacency matrix
- **Priors:** σ_v^2, σ_u^2, α
- **Parameters of interest:**
  - residual RR resRR $= \exp(V_i + U_i)$
  - marginal variance of the non-spatial effects V
  - empirical marginal variance of the spatial effects U
  - fraction of total variation in the log RR due to spatial effects

BUGS code

```
 for(i in 1:N) {
   O[i] ~ dpois(mu[i])
   log(mu[i]) <- log(E[i]) + alpha + V[i] + U[i]
   V[i] ~ dnorm(0, prec.v)             # unstructured RE
   RR[i] <- exp(alpha + V[i] + U[i])   # area specific RR
   resRR[i] <- exp(V[i] + U[i])        # area residual RR, after adjustment for area effect
  }

  U[1:N] ~ car.normal(adj[], weights[], num[], prec.u) # spatially corr effects (ICAR)

  for(k in 1:sumNumNeigh) { weights[k] <- 1 }

  alpha ~ dflat()                          # vague prior
  mean.RR <- exp(alpha)                    # overall mean risk

  prec.v ~ dgamma(0.5, 0.0005)
  sigma2.v <- 1/prec.v                     # variance of unstructured effects

  prec.u ~ dgamma(0.5, 0.0005)
  sigma2.u <- 1/prec.u                     # conditional variance of spatial effects
  sigma2.u.marginal <- sd(U[]) * sd(U[])   # empirical marginal var of spatial effects

 # fraction of total variation in log relative risks due to spatial effects
 frac.spatial <- sigma2.u.marginal / (sigma2.u.marginal + sigma2.v)

QR90 <- ranked(resRR[],8360)/ranked(resRR[],440)  # 90 percent quantile ratio
```

Arguments of the car.normal distribution in BUGS

car.normal(adj[], weights[], num[], prec.u)

- adj[]: sparse (vector) representation of adjacency matrix, e.g.

```
2, 5, 7, 8,
1, 4, 12,
5, 11, 17, 18, 23,
...
```

- num[]: vector of length N giving nb of neighbours for each area, e.g.

```
num=c(4, 3, 5, ...)
```

- weights[]: vector (same length as adj[]) giving (unnormalised) weights for each pair of neighbours

  $w_{ij} = 1$ if i and j are adjacent, 0 otherwise

- sumNumNeigh: length of adj[]

```
for(k in 1:sumNumNeigh) { weights[k] <- 1 }
```

- **R functions to construct these based on graph**

Data and initial values for BYM model

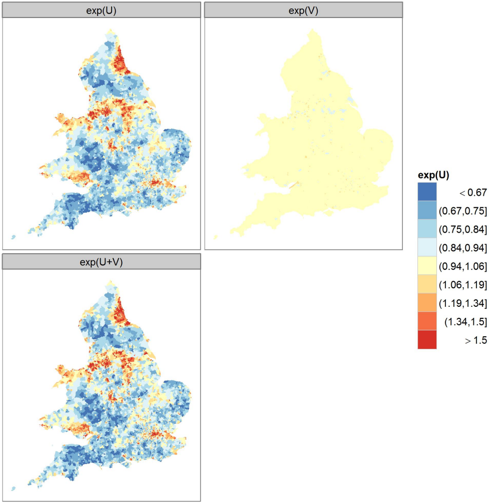
- Data: health and adjacency matrix

```
list(N=8800,
O=c(7,35,11,...),
E=c(17.49,39.78,12.20,...),
num = c(3, 3, 10, 4, 6, 7, 6, 6, 6, 6,
5, 5, 6, 6, 6, 5, 5, 4, 5, 5,
....
),
adj = c(
346, 3, 2,  #3 neighbours for area 1
346, 3, 1,  #3 neighbours for area 2
623, 566, 564, 349, 346, 218, 117, 4, 2, 1,  #10 neighbours for area 3
568, 566, 564, 3,  #4 neighbours for area 4
...
),
sumNumNeigh = 3576)
```

- Initial values

```
  list(alpha=0.01, prec.v=10,prec.u=10,
  V=c(0.01,0.01,0.01,...),
  U=c(0.01,0.01,0.01,..., -0.01,-0.01)#sum(U)=0
  )
```

Lung cancer incidence in males, 1985-2009, EW - I



Posterior means of the residual RR: spatial e^{U_i} and non-spatial e^{V_i} contributions, and total $e^{U_i+V_i}$. The spatial random effects dominate here.

Lung cancer incidence in males, 1985-2009, EW - II

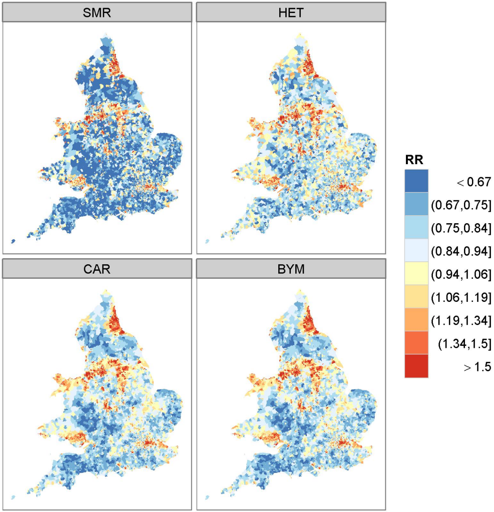
Comparison of the estimates of the residual RR using 4 methods

SMR non smoothed RR

HET non spatially smoothed residual RR $\exp(V)$

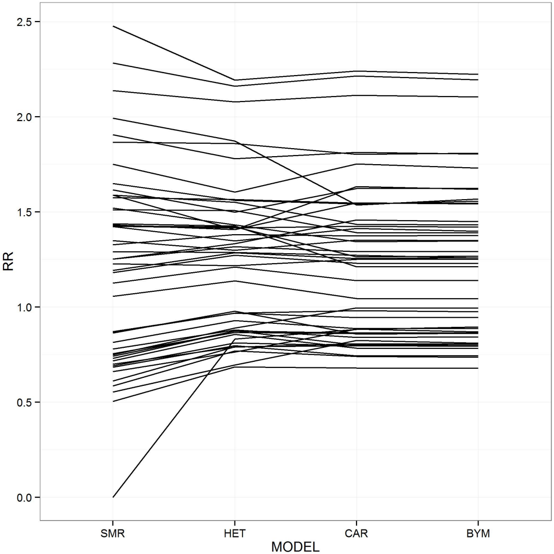
CAR spatially smoothed residual RR $\exp(U)$

BYM spatially and non spatially smoothed residual RR $\exp(V + U)$



Lung cancer incidence in males, 1985-2009, EW - III

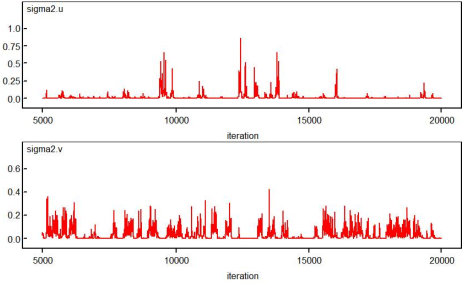
Comparison of estimates of the residual RR in selected areas



Other parameters of interest - Male lung cancer incidence

Model	DIC	σ_v^2	σ_u^2	σ_u^2 (marginal)	Spatial fraction
HET	65,867	0.094	-	-	-
ICAR	64,567	-	0.197	0.091	-
BYM = HET + ICAR	64,559	0.006	0.169	0.085	93%

Convergence



Interpretation of Disease Maps

- Smoothed relative risks are more stable (precise) than observed SMRs

  ⇒ geographical patterns of risk are easier to detect using smoothed maps

- Visual impact of maps can be very dependent on the choice of colours and cut-points used to shade each region

  - Should choose sensible cut-off values (eg symmetric and equally spaced on **logRR** scale)

  - Care must be taken not to over-interpret any patterns identified

Interpreting posterior risk estimates in disease mapping applications

- For sparse data, what is the sensitivity versus specificity of smoothed risk estimates?
  - Ability to detect true patterns (sensitivity)
  - Ability to discard false patterns (specificity)
- Smoothed rates usually result in higher specificity
  - Possible 'false positive' values shrunk towards mean
  - But, danger of over-smoothing (false negatives)
- Extensive simulation study to provide guidelines for interpretation of posterior relative risk estimates derived by Bayesian smoothing methods (Richardson et al (2004))

Posterior Probability

- Mapping the posterior mean RR does (or residual RR) not make full use of the output of the Bayesian analysis that provides, for each area, samples from the whole posterior distribution of the relative risk.
- Mapping the probability that a RR is greater than a specified threshold of interest has been proposed by several authors (e.g. Clayton and Bernardinelli (1992)).
- Very effective method to identify areas characterised by elevated risk

How to classify areas as having elevated risk

- We define the decision rule $D(c, RR_0)$, which depends
  - on a cutoff probability c
  - a reference threshold RR_0

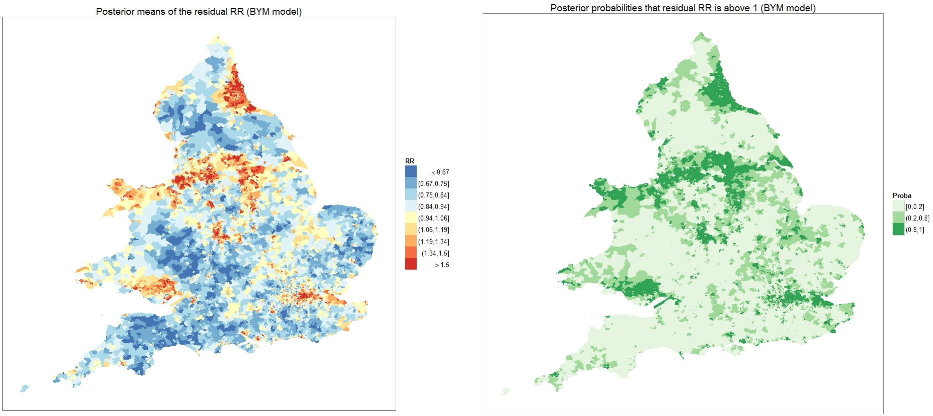
How to classify areas as having elevated risk

- We define the decision rule $D(c, RR_0)$, which depends
  - on a cutoff probability c
  - a reference threshold RR_0
- Area i is classified as having an elevated risk according to $D(c, RR_0) : \text{Prob}(RR_i > RR_0) > c$
- Recommended values $c = 0.8$ and $RR_0 = 1$ (Richardson et al. 2004)
- Posterior probability of interest: $\text{Prob}(resRR_i > 1) = \text{Prob}(e^{V_i + U_i} > 1)$
- In OpenBUGS, add the statement:

  ```
  proba.resRR[i]<-step(resRR[i]-1)
  ```

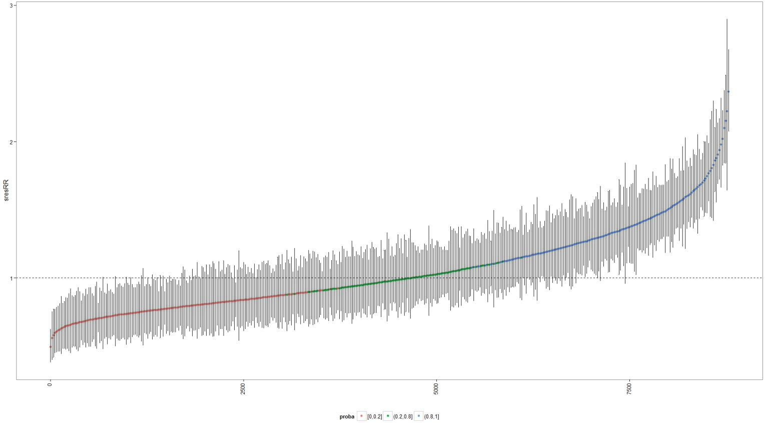
  (step(x) $= 1$ if $x >= 0$; 0 otherwise)

Lung cancer incidence in males, 1985-2009, EW - I



Map of the smoothed residual RRs and posterior probabilities that the resRR is above the average risk

Lung cancer incidence in males, 1985-2009, EW - II



CI95% of residual RR - red, green, blue identify the posterior probability that residRR>1

- Even if 1 in CI, probability can be above 0.8 or below 0.2

1. Introduction

2. Spatial structure
   - Intrinsic ICAR model
   - BYM model

3. Ecological regression
   - Poisson regression with random effects
   - Childhood leukaemia and benzene

4. Joint analysis of several diseases
   - Shared component models
   - Oral cavity and oesophagus cancers

Ecological regression: Extension to account for covariates

Straightforward extension of the BYM model:

Ecological regression with BYM structure

$$\begin{aligned} \mathrm{O}_i &\sim \text{Poisson}(E_i \lambda_i); \quad i = 1, ..., N \\ \log \lambda_i &= \alpha + V_i + U_i \\ \text{residual RR}_i &= \exp(V_i + U_i) \\ V_i &\sim \text{Normal}(0, \sigma_v^2) \\ \mathbf{U} &\sim \text{ICAR}(\mathbf{W}, \sigma_u^2) \ , \sum_i U_i = 0 \end{aligned}$$

where

- **V**: random effects without spatial structure, i.i.d.
- **U**: random effects with spatial structure, conditional distribution

Ecological regression: Extension to account for covariates

Straightforward extension of the BYM model:

Ecological regression with BYM structure

$$
\begin{aligned}
\mathrm{O}_i &\sim \text{Poisson}(E_i\lambda_i); \quad i = 1, ..., N \\
\log \lambda_i &= \alpha + \beta X_i + V_i + U_i \\
\text{residual RR}_i &= \exp(V_i + U_i) \\
V_i &\sim \text{Normal}(0, \sigma_v^2) \\
\mathbf{U} &\sim \text{ICAR}(\mathbf{W}, \sigma_u^2)\ , \sum_i U_i = 0
\end{aligned}
$$

where

- X area-level covariate of interest
- β: parameter associated with the covariate (assigned a prior)
- **V**: random effects without spatial structure, i.i.d.
- **U**: random effects with spatial structure, conditional distribution

Interpretation of the parameters

- $\exp(\beta)$ is the relative risk of disease associated with a unit increase in exposure X

- $U_i + V_i$ is the random effect in area i

- $\exp(U_i + V_i)$ is the residual or adjusted relative risk of disease in area i after accounting for the effects of measured covariates and the overall mean risk

- The variance of the random effects reflects the amount of overdispersion and residual spatial correlation in the data $(= \sigma_v^2 + s_{u.marginal}^2)$

Poisson regression with random effects - BUGS code Continuous covariate X

For each area $i = 1, ..., N$	`for(i in 1:N) {`
$O_i \sim \text{Poisson}(\mu_i)$	`O[i] ~ dpois(mu[i])`
$\log \mu_i = \log E_i + \alpha + \beta X_i + V_i + U_i$	`log(mu[i])<-log(E[i])+alpha+beta*X[i]+V[i]+U[i]`
residual $\text{RR}_i = \exp(V_i + U_i)$	`resRR[i] <- exp(V[i] + U[i])`
$V_i \sim \text{Normal}(0, \sigma_v^2)$	`V[i] ~ dnorm(0, prec.v)`
Posterior prob resRR>1	`proba.resRR[i]<-step(resRR[i]-1)`
	`}`
$\mathbf{U} \sim \text{ICAR}(\mathbf{W}, \sigma_u^2)$, $\sum_i U_i = 0$	`U[1:N]~car.normal(adj[],weights[],num[],prec.u)`
Weights	`for(k in 1:sumNumNeigh) {weights[k] <- 1}`
Priors	`#Priors`
β vague prior	`beta ∼ dnorm(0,0.000001)`
α vague prior	`alpha ~ dflat()`
σ_v^2 vague prior	`prec.v ~ dgamma(0.5,0.0005)`
σ_u^2 vague prior	`prec.u ~ dgamma(0.5,0.0005)`
	`#Parameters of interest`
e^β	`RR.X<-exp(beta)`

Poisson regression with RE - Data for R2OpenBUGS Continuous covariate X

- **Health and exposure data**

```
data = list(N=8800,
 O=c(7,35,11,...),
 E=c(17.49,39.78,12.20,...),
 X=c(1.23, 0.34,...))
```

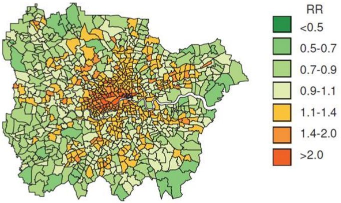
- **Initial values**

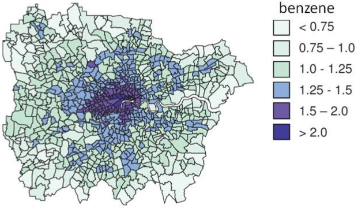
```
initial = list(alpha=0.01, prec.v=10, prec.u=10,
 beta=0.01,
 V=c(0.01,0.01,0.01,...) #same length as O and E
 U=c(0.01,0.01,..., -0.01, -0.01) #sum=0
```

Example 1: Childhood leukaemia and benzene

Best et al, 2001, JRSSA

- Can we explain some of the variation in risk of leukaemia by environmental exposure to benzene?
- Let X_i = average benzene emissions (tonnes per annum) in ward i (following cube-root transformation to reduce skew)





Results

- e^{β}: RR of leukaemia associated with unit increase in cube root benzene emissions in area of residence $= 2.23$ (1.64, 2.96)

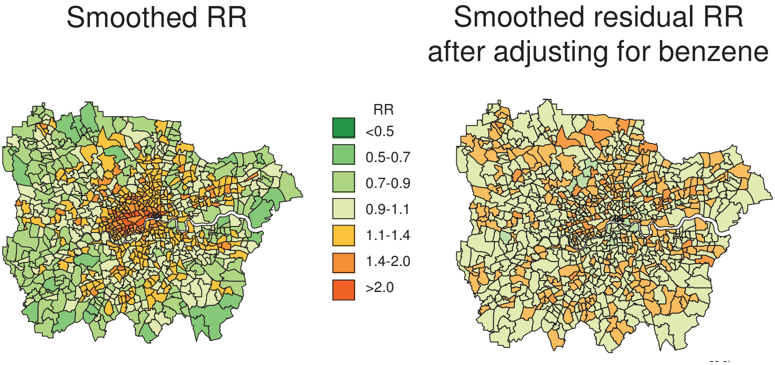
- e^{α}: overall mean RR of leukaemia in London $= 0.38$ (0.25, 0.54)

- $QR90$: % quantile ratio indicates that there is a 3.9-fold (95% 1.8 to 5.0-fold) CI variation in residual relative risk between the top and bottom 5% of areas *after adjusting for effects of benzene*

- $e^{(\alpha+\beta X_i+V_i+U_i)}$: RR of leukaemia in area i relative to London average

- $e^{(V_i+U_i)}$: $=$ residual RR of leukaemia in area i relative to London average after *adjusting for effects of benzene*

Maps of leukaemia RR



Smoothed RR

Smoothed residual RR after adjusting for benzene

Extension to several variables

$$
\begin{array}{rcl}
\mathrm{O}_i & \sim & \mathrm{Poisson}(E_i\lambda_i); \quad i = 1, ..., N \\
\log\lambda_i & = & \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + V_i + U_i \\
\text{residual } \mathrm{RR}_i & = & \exp(V_i + U_i) \\
V_i & \sim & \mathrm{Normal}(0, \sigma_v^2) \\
\mathbf{U} & \sim & \mathrm{CAR}(\mathbf{W}, \sigma_u^2) \; , \sum_i U_i = 0
\end{array}
$$

- $\exp(\beta_1)$ is the relative risk of disease associated with a unit increase in exposure X_1, after adjustment for X_2
- $\exp(\beta_2)$ is the relative risk of disease associated with a unit increase in exposure X_2, after adjustment for X_1
- $\exp(U_i + V_i)$ is the residual or adjusted relative risk of disease in area i after accounting for the effects of measured covariates and the overall mean risk

1. Introduction

2. Spatial structure
   - Intrinsic ICAR model
   - BYM model

3. Ecological regression
   - Poisson regression with random effects
   - Childhood leukaemia and benzene

4. Joint analysis of several diseases
   - Shared component models
   - Oral cavity and oesophagus cancers

Joint disease mapping: extension to several diseases

- Spatial modelling of disease risk almost exclusively for a single disease
- But: many diseases share common risk factors, e.g. smoking
- Current practice: Incidence rates from other diseases are used as surrogate exposure measures in a non-symmetric regression fashion

Joint disease mapping: extension to several diseases

- Spatial modelling of disease risk almost exclusively for a single disease
- But: many diseases share common risk factors, e.g. smoking
- Current practice: Incidence rates from other diseases are used as surrogate exposure measures in a non-symmetric regression fashion

Joint formulation seems more appropriate

→ Improves precision of risk estimates

→ Provides greater aetiological insight by identification of common and disease-specific geographical variations

Models for joint disease mapping

Data: observed and expected number of cases in area i for disease $d = 1, 2$, O_{di} and E_{di}

First-stage model

$$O_{1i} \sim \text{Poisson}(\mathrm{E}_{1i}\lambda_{1i}) \text{ and } O_{2i} \sim \text{Poisson}(\mathrm{E}_{2i}\lambda_{2i})$$

Second stage model for λ_{1i} and λ_{2i}

Models for joint disease mapping

Data: observed and expected number of cases in area i for disease $d = 1, 2$, O_{di} and E_{di}

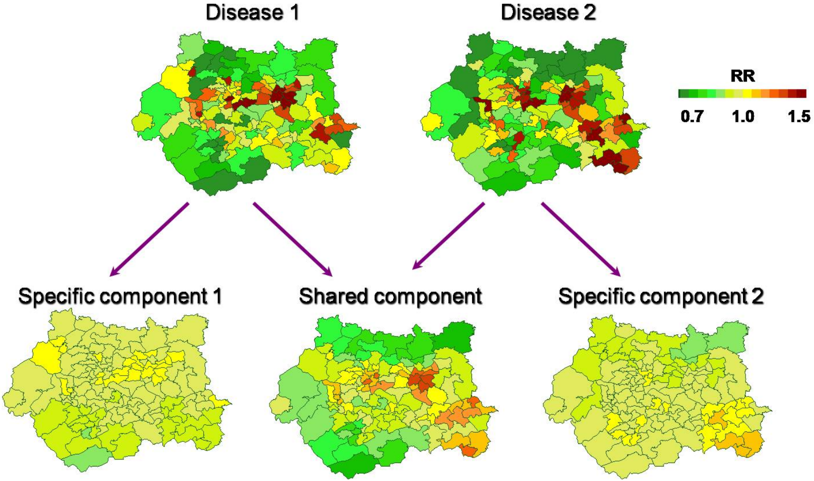
First-stage model

$$O_{1i} \sim \text{Poisson}(\mathrm{E}_{1i}\lambda_{1i}) \text{ and } O_{2i} \sim \text{Poisson}(\mathrm{E}_{2i}\lambda_{2i})$$

Second stage model for λ_{1i} and λ_{2i}

- Multivariate models
  - → Bivariate ICAR prior (extension of BYM model)
  - → Focus on correlation structure between latent disease-specific variables
- Shared component models: Separate risk surface of each disease into
  - a shared component and
  - a disease specific component

Shared component models I



Shared component models II

- Model for log relative risk:

$$\begin{aligned}\log(\lambda_{1i}) &= \alpha_1 + \phi_i.\delta + \psi_{1i}\\ \log(\lambda_{2i}) &= \alpha_2 + \phi_i/\delta + \psi_{2i}\end{aligned}$$

where
α_1, α_2: disease-specific intercepts
ϕ : shared component
ψ_1 and ψ_2 : disease-specific components
δ : scaling factor

- Independent BYM/ICAR models for ϕ, ψ_1 and ψ_2.

Shared component models II

- Model for log relative risk:

$$\begin{aligned}\log(\lambda_{1i}) &= \alpha_1 + \phi_i.\delta + \psi_{1i} \\ \log(\lambda_{2i}) &= \alpha_2 + \phi_i/\delta + \psi_{2i}\end{aligned}$$

  where
  α_1, α_2: disease-specific intercepts
  ϕ : shared component
  ψ_1 and ψ_2 : disease-specific components
  δ : scaling factor

- Independent BYM/ICAR models for ϕ, ψ_1 and ψ_2.
- $\delta \sim \text{LogNormal}(0, \tau^2) \rightarrow$ Formulation is invariant for switches of disease indices
- If $\delta > 1$, effect of the shared component on disease 1 is larger than on disease 2: the unobserved risk factors that are common to both diseases are associated with a higher risk of disease 1 than of disease 2

Shared component model

$$O_{1i} \sim \text{Poisson}(\text{E}_{1i}\lambda_{1i}) \text{ and } O_{2i} \sim \text{Poisson}(\text{E}_{2i}\lambda_{2i})$$

$$\begin{aligned} \log(\lambda_{1i}) &= \alpha_1 + \phi_i.\delta + \psi_{1i} \\ \log(\lambda_{2i}) &= \alpha_2 + \phi_i/\delta + \psi_{2i} \end{aligned}$$

$\phi_i = Ush_i + Vsh_i$

$Ush \sim \text{ICAR}(\sigma^2_{ush})$

$Vsh \sim \text{N}(0, \sigma^2_{vsh})$

$\psi_{1i} = Usp_{1i} + Vsp_{1i}$

$Usp_1 \sim \text{ICAR}(\sigma^2_{usp1})$

$Vsp_{1i} \sim \text{N}(0, \sigma^2_{vsp1})$

$\psi_{2i} = Usp_{2i} + Vsp_{2i}$

$Usp_2 \sim \text{ICAR}(\sigma^2_{usp2})$

$Vsp_{2i} \sim \text{N}(0, \sigma^2_{vsp2})$

Shared component model

$$O_{1i} \sim \text{Poisson}(\text{E}_{1i}\lambda_{1i}) \text{ and } O_{2i} \sim \text{Poisson}(\text{E}_{2i}\lambda_{2i})$$

$$\begin{aligned} \log(\lambda_{1i}) &= \alpha_1 + \phi_i.\delta + \psi_{1i} \\ \log(\lambda_{2i}) &= \alpha_2 + \phi_i/\delta + \psi_{2i} \end{aligned}$$

$$\phi_i = Ush_i + Vsh_i \qquad \psi_{1i} = Usp_{1i} + Vsp_{1i} \qquad \psi_{2i} = Usp_{2i} + Vsp_{2i}$$

$$Ush \sim \text{ICAR}(\sigma^2_{ush}) \qquad Usp_1 \sim \text{ICAR}(\sigma^2_{usp1}) \qquad Usp_2 \sim \text{ICAR}(\sigma^2_{usp2})$$

$$Vsh \sim \text{N}(0, \sigma^2_{vsh}) \qquad Vsp_{1i} \sim \text{N}(0, \sigma^2_{vsp1}) \qquad Vsp_{2i} \sim \text{N}(0, \sigma^2_{vsp2})$$

- Priors on all the variance parameters
- Parameters of interest
  - residual RR specific to disease d: $\exp(\psi_{di})$
  - shared component of the risk common to both diseases: $\exp(\phi_i)$
  - fraction of total variation in RR for each disease that is explained by the shared component
  - Ratio of the two risk gradients δ

BUGS code

```
model {
for (i in 1:Nareas) {
for (k in 1:Ndiseases) {
O[i,k] ~ dpois(mu[i,k])
log(mu[i,k]) <- log(E[i,k]) + alpha[k] + eta[i, k] }}

for(i in 1:Nareas) {
# Define log relative risk in terms of disease-specific (psi) and shared (phi) random effects
eta[i,1] <- phi[i] *delta + psi[1, i]
eta[i,2] <- phi[i] /delta + psi[2, i]

# Spatial priors (BYM) for the disease-specific random effects
for (k in 1:Ndiseases) {
for (i in 1:Nareas) {
psi[k, i] <- V.sp[k, i] + U.sp[k, i] #BYM
V.sp[k, i] ~ dnorm(0, tau.unstr[k]) }  #unstructured disease-specific random effects
U.sp[k,1:Nareas] ~ car.normal(adj[], weights[], num[], tau.spatial[k]) } #spatial disease-spec effects

# Spatial priors (BYM) for the shared random effects
for (i in 1:Nareas) {
phi[i] <- V.sh[i] + U.sh[i] #BYM
V.sh[i] ~ dnorm(0, omega.unstr)} #unstructured shared random effects
U.sh[1:Nareas] ~ car.normal(adj[], weights[], num[], omega.spatial) # spatial shared effects

for (k in 1:sumNumNeigh) {weights[k] <- 1}

#.....continued next page.......
```

BUGS code - priors

```
for (k in 1:Ndiseases) {
alpha[k] ~ dflat()
tau.unstr[k] ~ dgamma(0.5, 0.0005)
tau.spatial[k] ~ dgamma(0.5, 0.0005)
}
omega.unstr ~ dgamma(0.5, 0.0005)
omega.spatial ~ dgamma(0.5, 0.0005)
logdelta ~ dnorm(0, 5.9)  # scaling factor for relative strength of shared component
                          # for each disease
delta <- exp(logdelta)    # (prior assumes 95% probability that delta is between 0.44 and 2.2;
                          # lognormal assumption is invariant to which disease is labelled 1
                          # and which is labelled 2)
```

BUGS code - parameters of interest

```
for (i in 1:Nareas) {
specificRR1[i]<- exp(psi[1,i])      # residual RR specific to disease 1
specificRR2[i]<- exp(psi[2,i])      # residual RR specific to disease 2

sharedRR[i] <- exp(phi[i])         # shared component of risk common to both diseases
logsharedRR1[i] <- phi[i]*delta    # Note that this needs to be scaled by delta or 1/delta if the
logsharedRR2[i] <- phi[i]/delta    # absolute magnitude of shared RR for each disease is of interest
}

var.shared[1] <- sd(logsharedRR1[])*sd(logsharedRR1[])  # empirical var of shared effects
                                                        # (scaled for disease 1)
var.shared[2] <- sd(logsharedRR2[])*sd(logsharedRR2[])  # empirical var of shared effects
                                                        # (scaled for disease 2)
var.specific[1] <- sd(psi[1,])*sd(psi[1,])     # empirical var of disease 1 specific effects
var.specific[2] <- sd(psi[2,])*sd(psi[2,])     # empirical var of disease 2 specific effects

# fraction of total variation in relative risks for each disease that is explained
#by the shared component
frac.shared[1] <- var.shared[1] / (var.shared[1] + var.specific[1])
frac.shared[2] <- var.shared[2] / (var.shared[2] + var.specific[2])
```

Data for the joint modelling

When using WinBUGS/OpenBUGS:

- Data

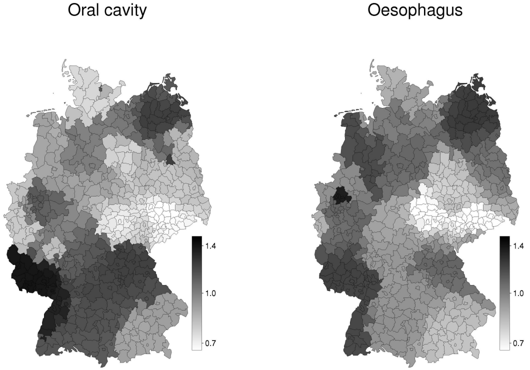
```
list(Nareas=8800, Ndiseases=2,
O=structure(.Data= c(7,35,11,...),.Dim=c(8800, 2)),
E=structure(.Data= c(17.49,39.78,12.20,...),.Dim=c(8800, 2),
...)
```

- Initial values for the unknown parameters for each chain

```
list(alpha=c(0.01, 0.01), tau.unstr=c(10,10),
     V.sh=c(0.01,0.01,0.01,...),
     U.sh=c(0.01,0.01,...,-0.01,-0.01) #sum U = 0
     ...)
```

Analysis of oral cavity and oesophagus cancers I

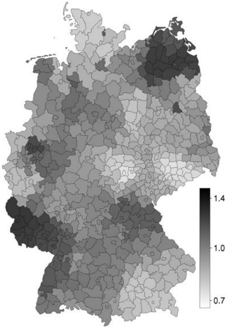
- Oral cavity and oesophagus cancer mortality, 544 districts in Germany, 1986-1990 (Knorr-Held and Best, 2001)
- Established risk factors: tobacco and alcohol
- Separate analysis: map of the smoothed RR


Oral cavity
Oesophagus

Spatial structure similar for both cancers: high values in the NE and SW

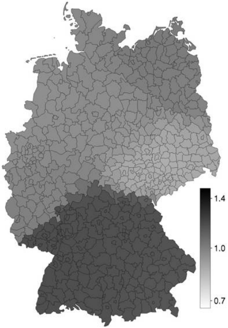
Analysis of oral cavity and oesophagus cancers II

Shared RR



2 large clusters in the NE and SW (regions where alcohol or tobacco consumptions are high,respectively)

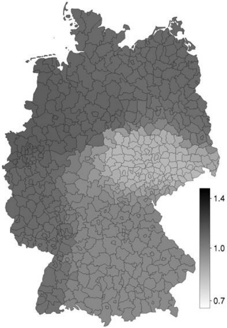
Oral cavity-specific RR



Clear spatial pattern: higher RRs in the S and lower RRs in the N

⇒ existence of additional risk factors that are relevant only to oral cavity but not to oesophageal cancer

Oesophageal-specific RR



Different spatial pattern with less variation and slightly higher RRs in the W and N.

Adjacency matrix with R

- Import the shapefile into R
- Convert the polygons to a list of neighbours using `poly2nb` function (spdep package)
- Include in the data a list of 3 components is created: adj, num and weights

```
#convert the polygons to a list of neighbours
shp_nb<-poly2nb(shp)
summary(shp_nb)

#convert to BUGS format
#a list of 3 components is created: adj, num and weights
nbWB <- nb2WB(shp_nb)
names(nbWB)
summary(nbWB)
```

Summary

- Smoothing of small area risks is important to help to separate spatial pattern from 'noise'

- Many methods available in the literature, but BYM most used method for disease mapping
  - Global smoothing achieved by iid distribution
  - Local smoothing achieved by spatial distribution (borrowing information from neighbouring regions)

- Natural extension to include covariates (ecological regression)

- Natural extension to model 2 or more diseases (joint disease mapping)

References and further reading

Besag, J.; York, J. and Mollie, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics* **43**:1-59

Best, N. and Hansell, A. (2009) Geographic Variations in Risk: Adjusting for Unmeasured Confounders Through Joint Modeling of Multiple Diseases. *Epidemiology* **20**(3):400-410.

Knorr-Held, L. and Best, N. (2001) A Shared Component Model for Detecting Joint and Selective Clustering of Two Diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)***164**(1):73-85.