Statistical Methodology for disease mapping:

Rate smoothing and issues of sensitivity and specificity

Sylvia Richardson

Department of Epidemiology and Public Health

In collaboration with Nicky Best and Andrew Thomson

Introduction

- Interest in conducting spatial analyses of health outcomes at the small area scale
  - Highlight sources of heterogeneity and spatial patterns
  - Suggest public health determinants or aetiological clues
- Small scale
  - less susceptible to ecological (aggregation) bias
  - more able to detect highly localised effects

BUT sparse data need more sophisticated statistical analyses techniques

Basic model for small area data

- Typically dealing with rare events in small areas A_i

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

Y_i is the observed count of disease in area

E_i is the expected count based on population size, adjusted for age, sex, other strata ….,

θ_i is a region specific relative risk : parameter of interest

assumes multiplicative model between area effect and age-sex in all strata

- Relative risk, θ_i, usually estimated by $SMR_i = Y_i / E_i$

Can be used to test an increase of risk in a single area:

$\theta_i > 1$

BUT:

- if interested in more than one area

→ problems of multiple testing and control of overall significance level (false detection rate)

- evidence of localised raised RR should be interpreted in the context of overall variability of disease rates in the region/country

Disease Mapping

- Common practice is to map SMR_i for each area $i = 1,\ldots,N$

BUT:

- $SE(SMR_i) \propto 1 / E_i \rightarrow SMR_i$ very imprecise for rare diseases and/or areas with small populations

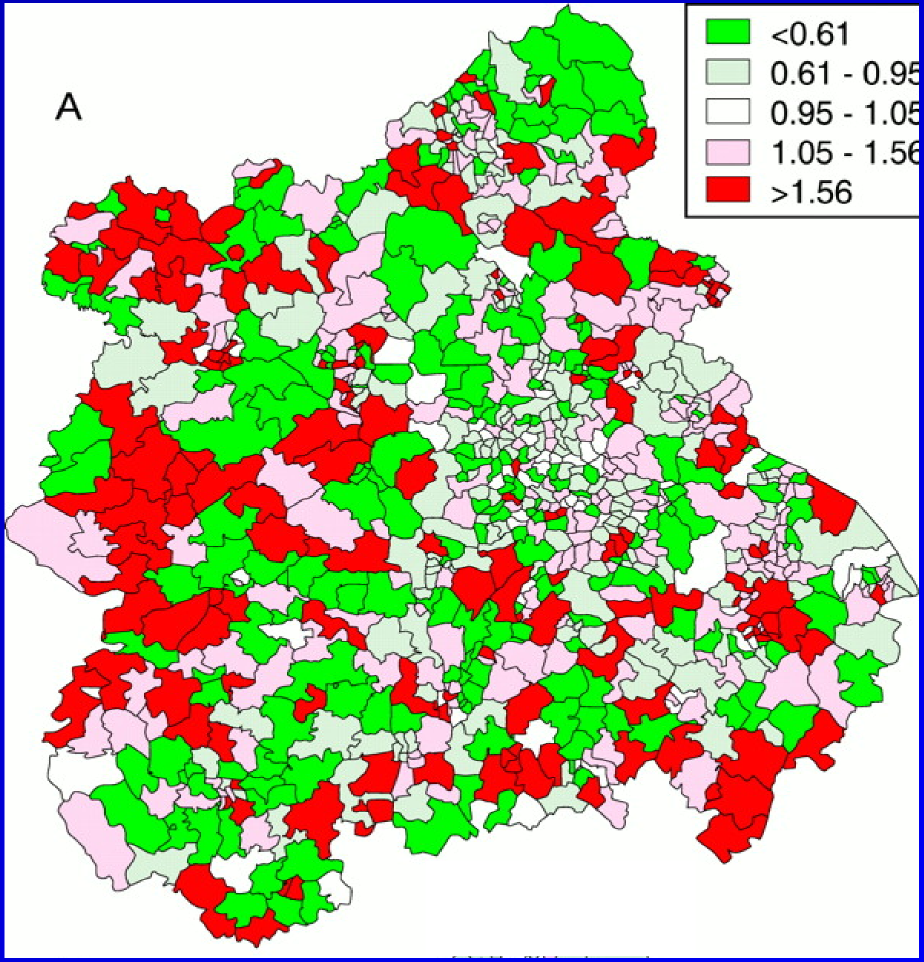
  → the precision can vary widely between areas

- SMR_i in each area is estimated independently

  → makes no use of risk estimates in other areas of the map, even though these are likely to be similar

  - highlights extreme risk estimates based on small numbers
  - ignores possible spatial correlation between disease risk in nearby areas due to possible dependence on spatially varying risk factors

Map of SMR of adult leukaemia in West Midlands Region, England 1974-86 (Olsen, Martuzzi and Elliott, *BMJ* 1996;313:863-866).



Is the variability real or simply reflecting unequal E_is ?

Have the highlighted areas truly a raised relative risk?

Bayesian Hierarchical Models

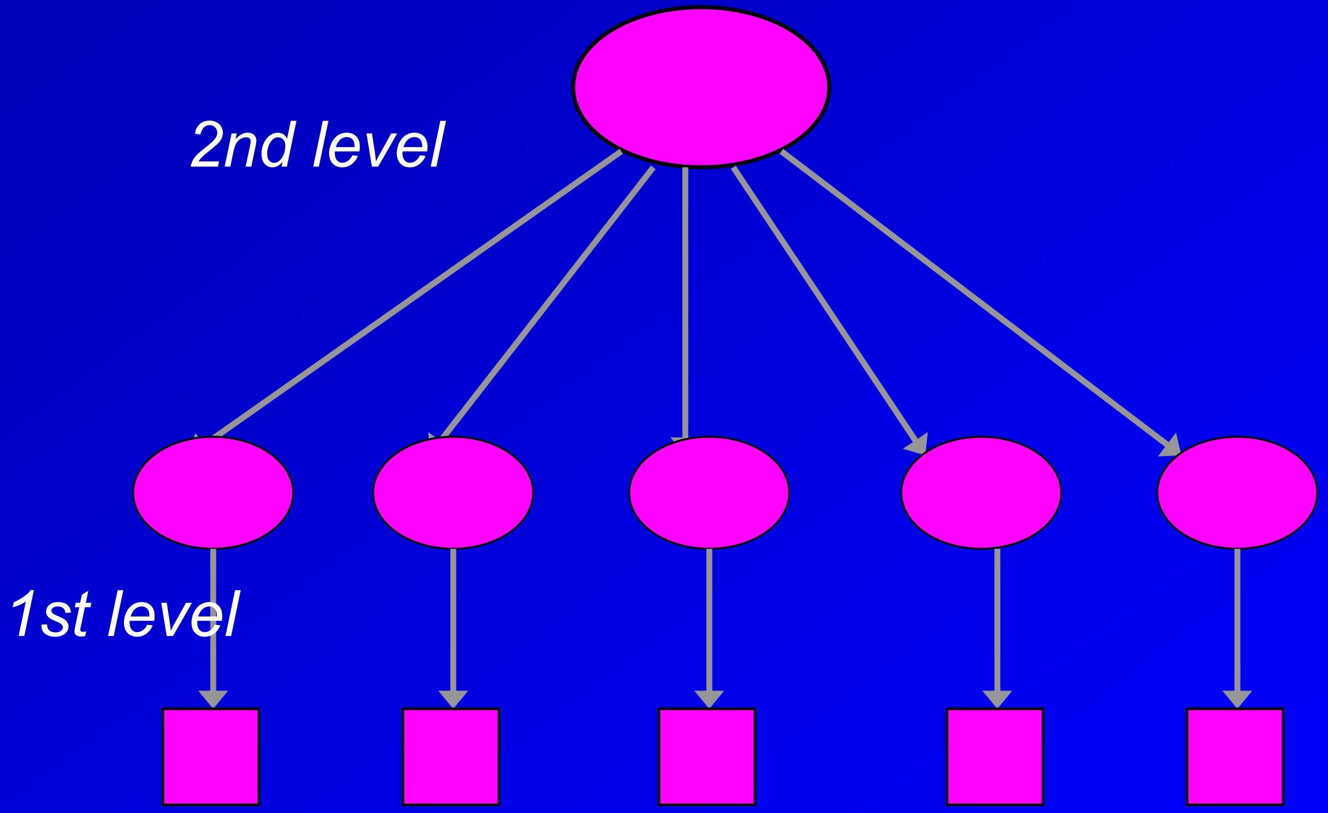
- These problems may be addressed using Bayesian ‘smoothing’ or ‘shrinkage’ estimators
- Assumes that the RRs $\{\theta_i\}$ come from a common distribution,

E.g.
$$\begin{cases} Y_i \sim \text{Poisson}(\theta_i E_i), \\ \log(\theta_i) \sim \text{Normal}(\mu, \sigma^2) \end{cases} \qquad \mathbf{i = 1,\ldots,N}$$

- Leads to estimate of the ‘true’ relative risk in area i that is a weighted average of the observed area-level risk ratio (SMR_i) and parameters reflecting the regional or national distribution of the relative risks, with weights depending on the population at risk in area i

⇨ relative risks are ‘shrunk’ and stabilised (smoothed)

Schematic representation of a hierarchical model



2nd level

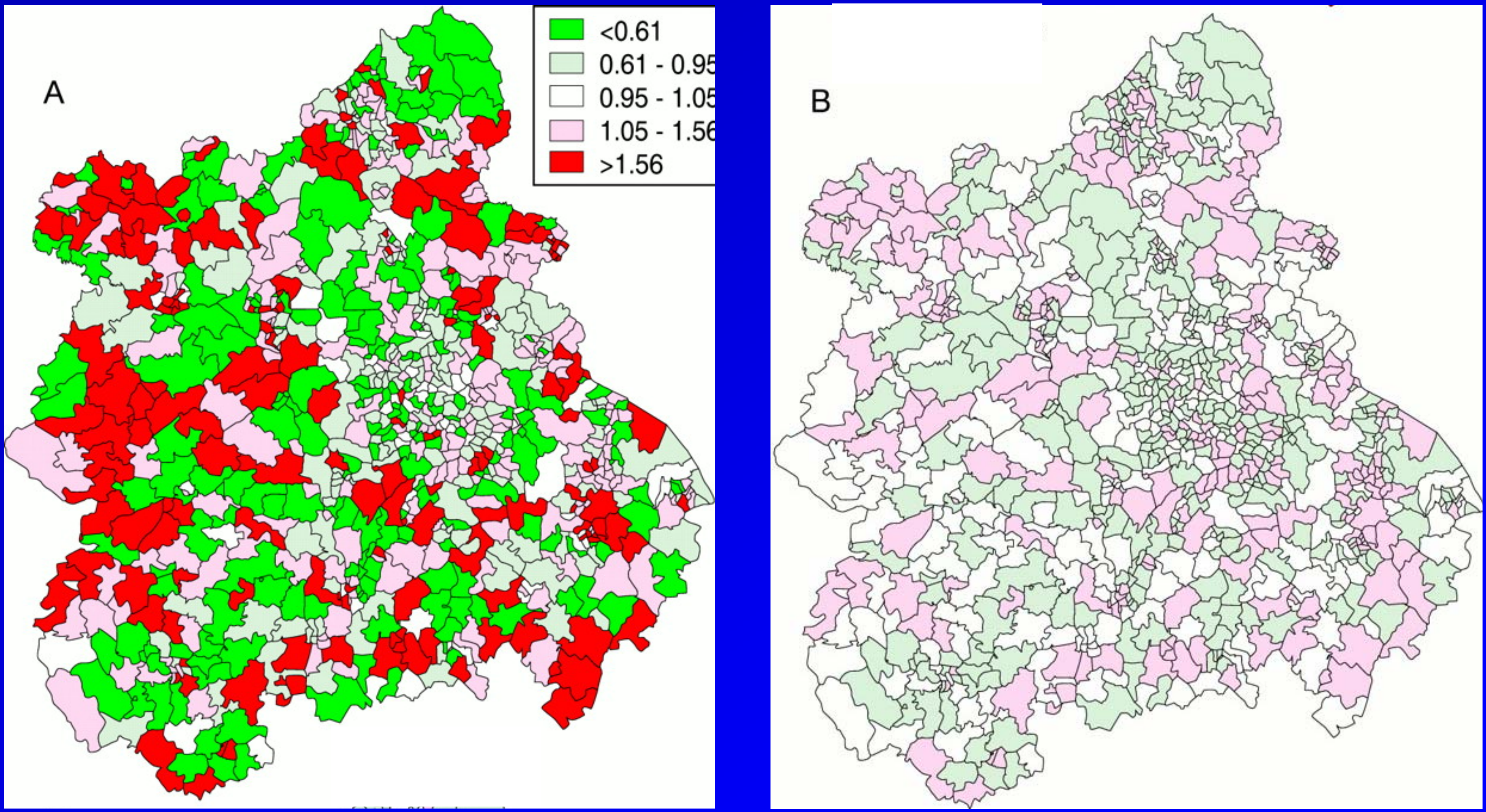
1st level

Overall disease risk

'True' disease risk θ_i

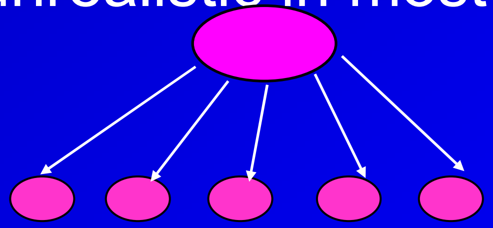
Observed disease risk (SMR_i)

Map of occurrences of adult leukaemia in West Midlands Region, England 1974-86: (A) unsmoothed SMR, (B) smoothed by Bayesian methods. (Olsen, Martuzzi and Elliott, *BMJ* 1996;313:863-866).



Building the hierarchical model

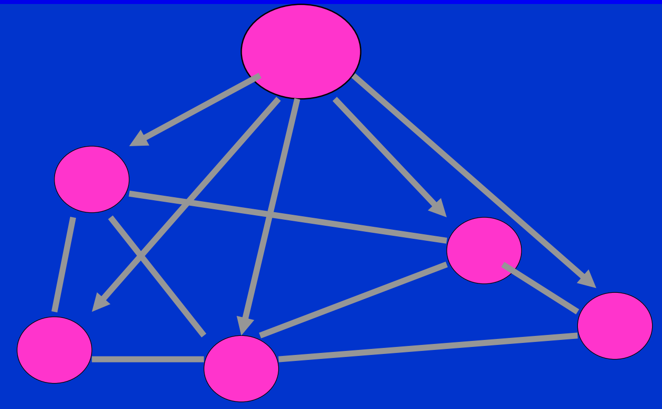
- Assuming that the relative risks $\{\theta_i\}$ are independently drawn from a common distribution is unrealistic in most epidemiological setting



- The θ_i are typically spatially correlated because they reflect in part spatially varying risk factors

⇒ Incorporation of spatial dependence in the distribution of θ_i

2nd level



Conditional Autoregressive (CAR) model

$\log(\theta_i) \sim \text{Normal}(\Sigma_k \theta_k / n_i, \sigma^2/n_i)$

for k = neighbour of i (n_i = #k)

Software

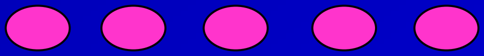
- Estimation may be carried out using

  Empirical Bayes (uses ‘plug-in’ estimate for parameters) or

  Hierarchical Bayes (fully accounts for uncertainty in all unknown parameters)

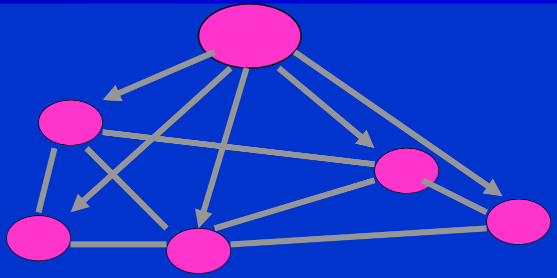
- Estimation of Bayesian hierarchical models requires computationally intensive simulation methods
  – Software (WinBUGS, GeoBUGS) developed at Imperial (N. Best)

Including spatial dependence in disease risk



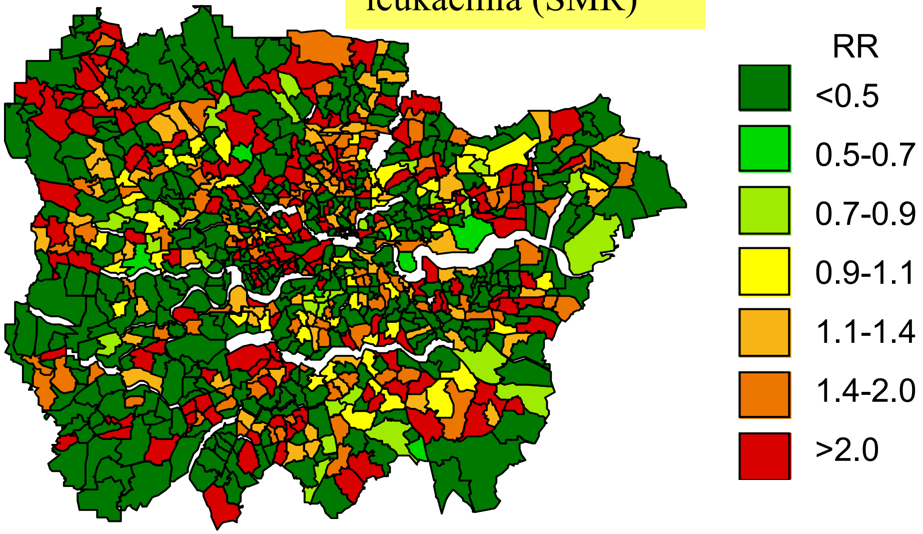
$Y_i \sim \text{Poisson}(\theta_i \, E_i)$, $i=1,\ldots,N$

$\theta_i = Y_i \,/\, E_i =$ SMR in area i

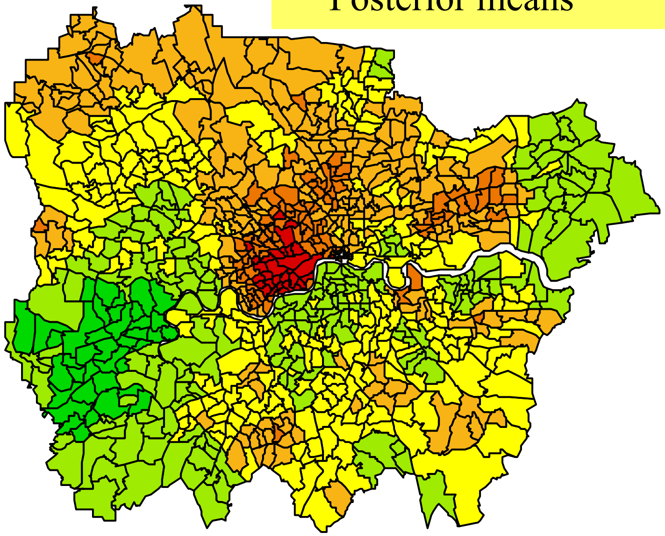


$Y_i \sim \text{Poisson}(\theta_i \, E_i)$, $i=1,\ldots,N$

$\theta_{\,i} \sim$ CAR model



Childhood leukaemia (SMR)



Childhood leukaemia Posterior means

Current methodological issues (1)

- Model choice for allowing spatial dependence in the second level
  - Different models have different shrinkage properties
- Model checking and diagnostics, predictive fit
  - Comparison of the performance of different spatial models for uncovering true pattern of heterogeneity
  - Use of an Bayesian model comparison criterion based on posterior deviance
- Sensitivity and specificity of smoothed estimates

Model choice

- For allowing spatial dependence in the second level – choices include:
  - Multivariate normal with spatial covariance matrix

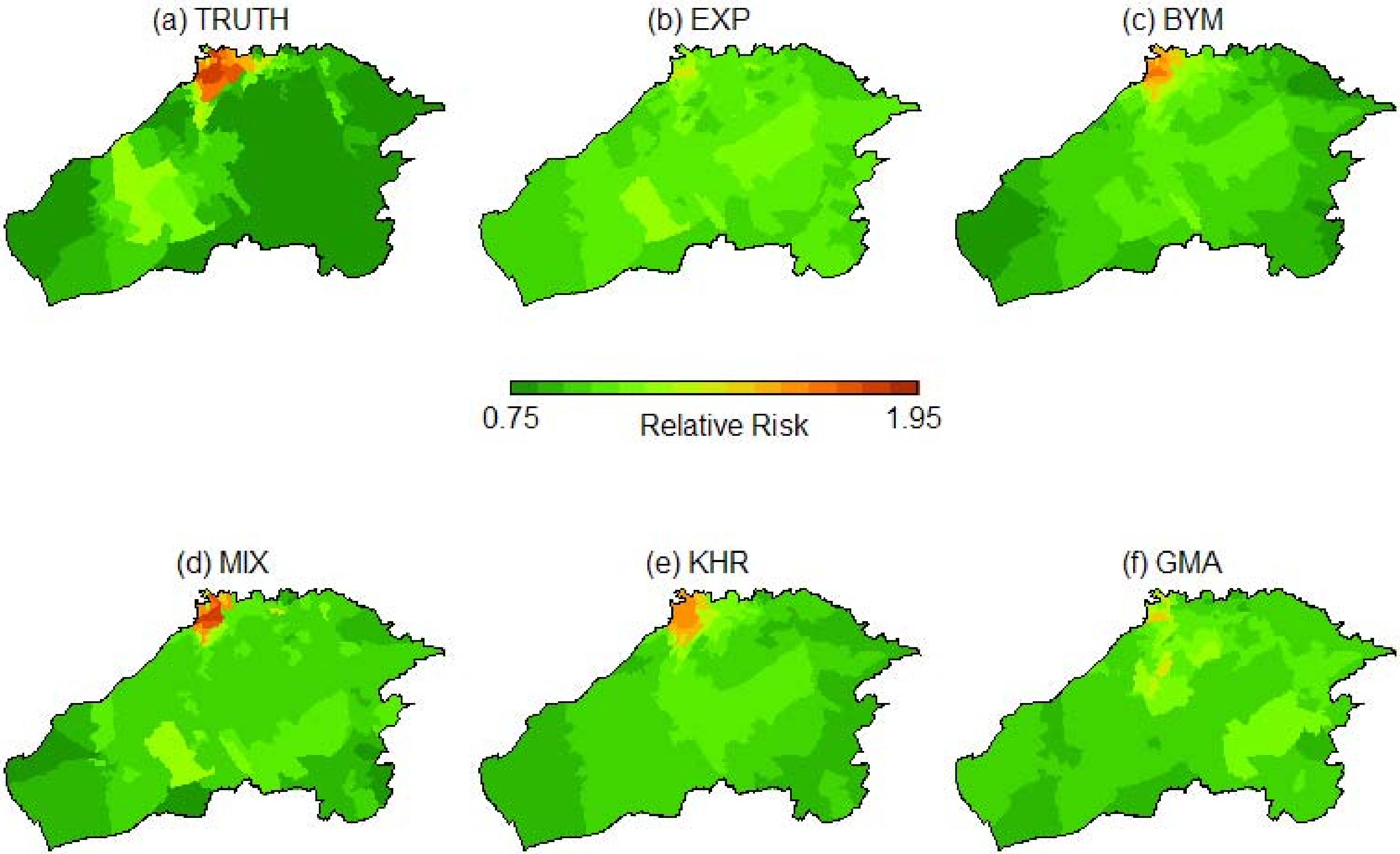
    e.g. with exponential decrease (EXP)
  - Markov Random Field models (Besag, York and Mollié, 1991)

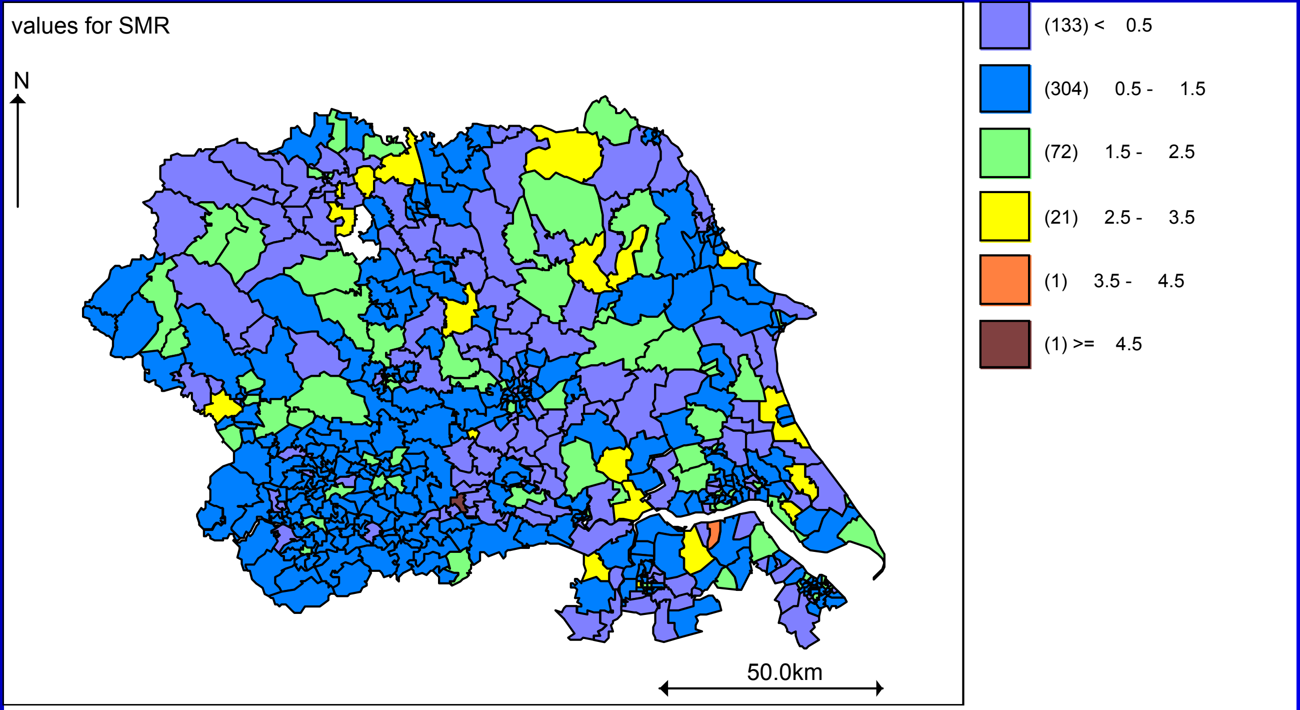
    CAR: assume dependence between adjacent areas,

    BYM = CAR + unstructured heterogeneity (allows more flexibility)
  - Spatial partition models (Knorr Held and Rasser, 2000) (KHR)
  - Spatial mixture models (Green and Richardson, 2002) (MIX)
  - Moving average models (Best et al, 2000)

    e.g with gamma distributed impulses (GMA)

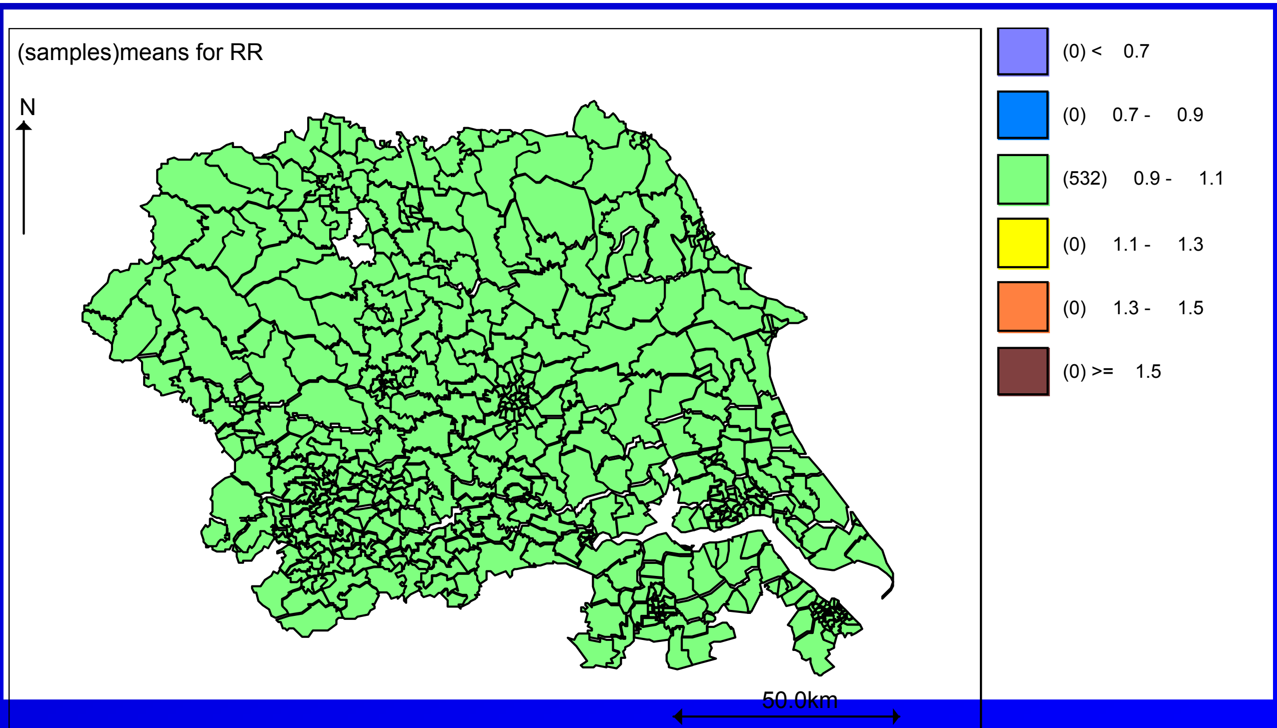
Simulation study comparing the smoothing of different spatial priors





Raw SMR

Prostate Cancer Yorkshire



Smoothed estimates:

Are they too smoothed?

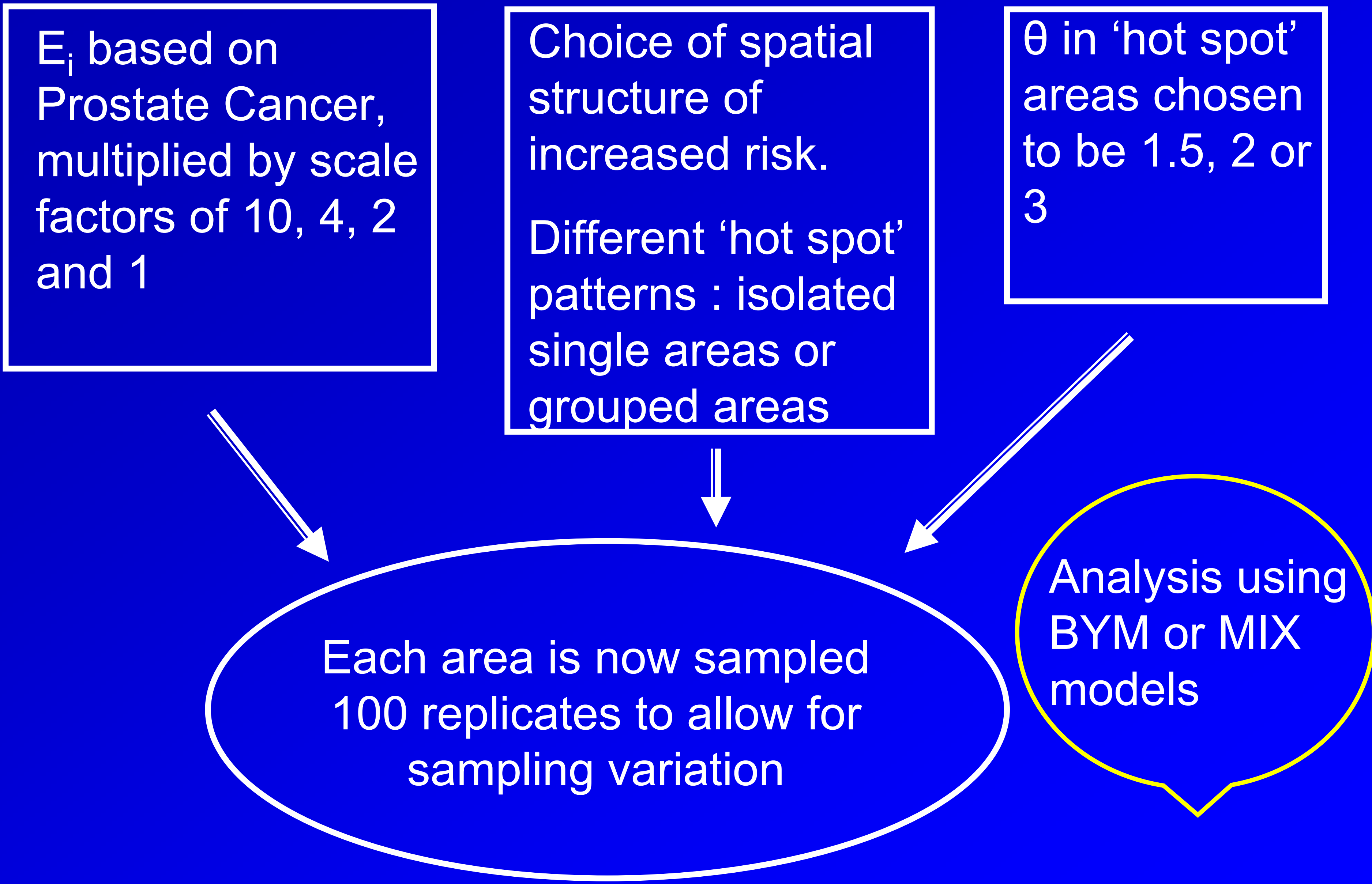
Current methodological issues (2)

- 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)
- Extensive simulation study to give guidelines for interpretation of posterior relative risk estimates derived by Bayesian smoothing methods

→ Highlights the advantage of using the whole posterior distribution of the RRs

and computing: Probability ($\theta_i > 1$)

How the Simulation is Carried out



E_i based on Prostate Cancer, multiplied by scale factors of 10, 4, 2 and 1

Choice of spatial structure of increased risk.

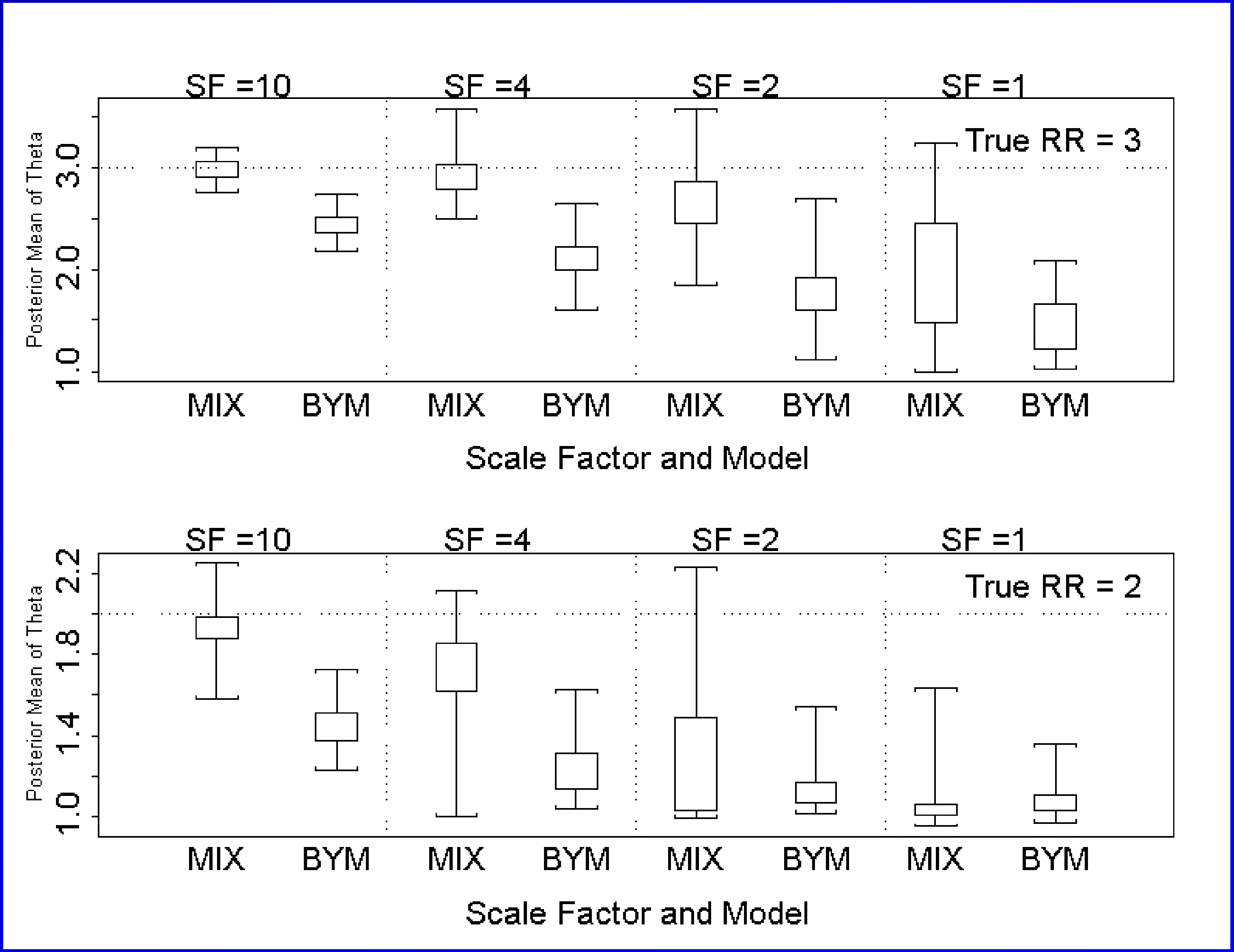
Different ‘hot spot’ patterns : isolated single areas or grouped areas

θ in ‘hot spot’ areas chosen to be 1.5, 2 or 3

Each area is now sampled 100 replicates to allow for sampling variation

Analysis using BYM or MIX models

Smoothing of the RRs of hot spots (4 contiguous areas with average expected counts ≈ 5) for different spatial models



Comparison

- All models lead to considerable smoothing unless the expected count is high
- MIX performs less shrinkage than BYM models (Gaussian or median based)
- Mapping the mean posterior value of θ_i does not make full use of the posterior distribution θ_i that is obtained for each area

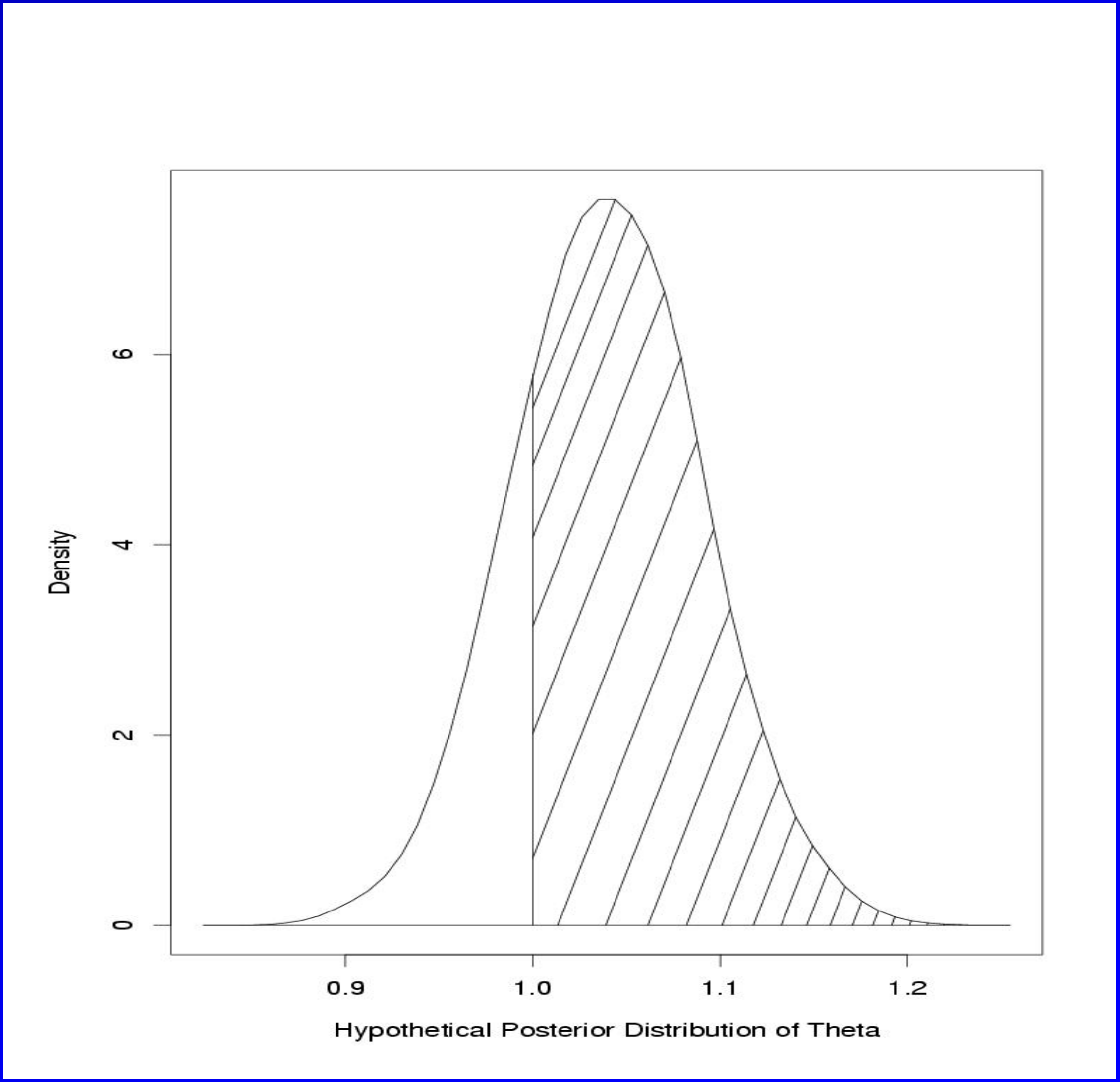
Investigate the use of the decision rule based on

Probability (θ_i > threshold)

e.g. Probability (θ_i > 1)

Decision rule: an example

- Compute Probability ($\theta_i > 1$)
- Classify an area as having an elevated risk if [Prob ($\theta_i > 1$)] > 0.8
- This rule has high specificity in most cases (% false detection < 10%)
- Sensitivity ?



Sensitivity of the decision rule: [Prob (θ_i > 1) > 0.8] to declare an area as having an elevated risk for the BYM model

		Scale factor = 1			Scale factor = 2		
	BYM	Θ=1.5	Θ=2	Θ=3	Θ=1.5	Θ=2	Θ=3
Single raised area	(E=1.10)	0.36	0.48	0.38	0.20	0.24	0.36
	(E=1.92)	0.32	0.48	0.40	0.16	0.32	0.66
	(E=5.37)	0.08	0.30	0.74	0.12	0.52	0.98
	(E=7.38)	0.12	0.22	0.74	0.10	0.64	0.98
Grouped	(E=5.42)	0.18	0.42	0.95	0.30	0.74	1

Scale factor = 4			Scale factor = 10		
Θ=1.5	Θ=2	Θ=3	Θ=1.5	Θ=2	Θ=3
0.20	0.50	0.82	0.28	0.54	1
0.24	0.66	0.98	0.30	0.96	1
0.22	0.76	1	0.66	1	1
0.34	0.88	1	0.88	1	1
0.53	0.97	1	0.90	1	1

RR of 1.5 are not detected unless E > 20

RR of 2 are detected, with E ≈ (10-20) with prob 0.75

RR of 3 are detected, with E ≈ 5

Richardson, Thompson, Best, Elliott, 2004

Conclusions

- Beneficial to implement a variety of flexible spatial models in order to gain practical insights into their properties

- Useful to investigate and compare their performance by simulation studies

  - Some improvement linked to the use of partition or mixture models

Conclusions (continued)

- Decision rules based on the posterior distribution of the relative risks shows:
  - Good specificity of Bayesian disease mapping models
  - Low sensitivity for detecting small excess risk
  - Trade off between size of areas and size of expected counts, anticipated magnitude and structure of the putative risks

→ Borrowing information between diseases
Introduction of area level covariates

Thank you