Lecture 11
Applications of Multi-level Models to Disease Mapping

Outline

- Multi-level models for spatially correlated data
  - Socio-economic and dietary factors of pellagra deaths in southern US
- Multi-level models for geographic correlation studies
  - The Scottish Lip Cancer Data

Data characteristics

- Data for disease mapping consists of disease counts and exposure levels in small adjacent geographical area
- The analysis of disease rates or counts for small areas often involves a trade-off between statistical stability of the estimates and geographic precision

An example of multi-level data in spatial epidemiology

- We consider approximately 800 counties clustered within 9 states in southern US
- For each county, data consists of observed and expected number of pellagra deaths
- For each county, we also have several county-specific socio-economic characteristics and dietary factors
  - % acres in cotton
  - % farms under 20 acres
  - dairy cows per capita
  - Access to mental hospital
  - % afro-american
  - % single women

Definition of Standardized Mortality Ratio

- Y_i is the observed number of deaths in area i
- E_i is the expected number of deaths in area i
- The "raw" Standardized Mortality Ratio is so defined:

$$SMR_i = (Y_i/E_i) \times 1000$$

Definition of the expected number of deaths

- The expected number of deaths in area i can be calculated as follows:

$$E_i = \sum_j p_j n_{ij}$$

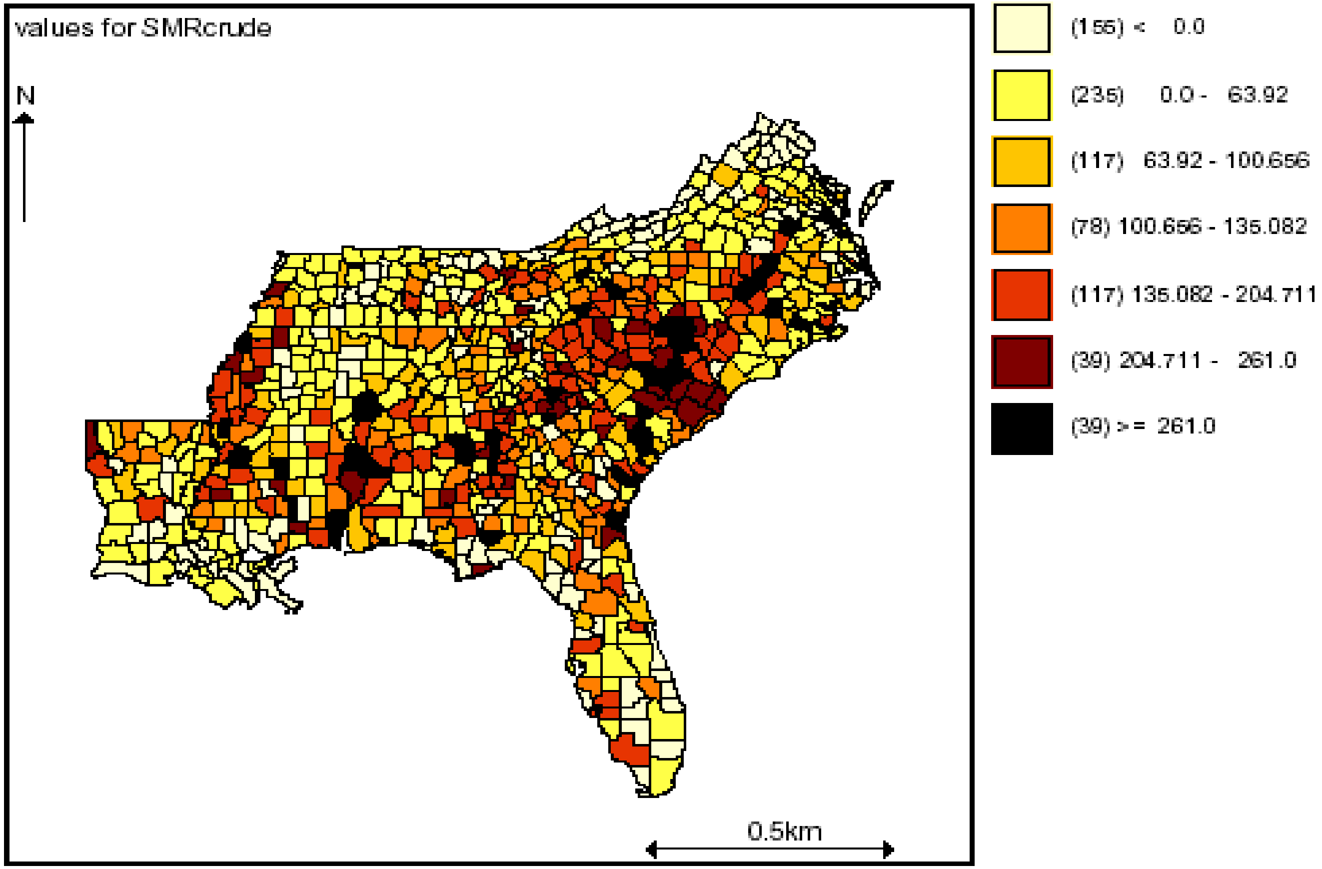
  where

- j is the population stratum generally defined by age×gender×race
- p_j is observed frequency of death in the reference population
- n_{ij} is the number of people at risk in area i in stratum j

Definition of Pellagra

- Disease caused by a deficient diet or failure of the body to absorb B complex vitamins or an amino acid.
- Common in certain parts of the world (in people consuming large quantities of corn), the disease is characterized by scaly skin sores, diarrhea, mucosal changes, and mental symptoms (especially a schizophrenia-likedementia). It may develop after gastrointestinal diseases or alcoholism.

Crude Standardized Mortality Ratio (Observed/Expected) of Pellagra Deaths in Southern USA in 1930 (*Courtesy of Dr Harry Marks)*



Scientific Questions

- Which social, economical, behavioral, or dietary factors best explain spatial distribution of pellagra in southern US?
- Which of the above factors is more important for explaining the history of pellagra incidence in the US?
- To which extent, state-laws have affected pellagra?

Statistical Challenges

- For small areas SMR are very instable and maps of SMR can be misleading
  - Spatial smoothing
- SMR are spatially correlated
  - Spatially correlated random effects
- Covariates available at different level of spatial aggregation (county, State)
  - Multi-level regression structure

Spatial Smoothing

- Spatial smoothing can reduce the random noise in maps of observable data (or disease rates)
- Trade-off between geographic resolution and the variability of the mapped estimates
- Spatial smoothing as method for reducing random noise and highlight meaningful geographic patterns in the underlying risk

Shrinkage Estimation

- Shrinkage methods can be used to take into account instable SMR for the small areas
- Idea is that:
  - *smoothed estimate for each area "borrow strength" (precision) from data in other areas, by an amount depending on the precision of the raw estimate of each area*

Shrinkage Estimation

- Estimated rate in area *A* is adjusted by combining knowledge about:
  - Observed rate in that area;
  - Average rate in surrounding areas
- The two rates are combined by taking a form of weighted average, with weights depending on the population size in area *A*

Shrinkage Estimation

- When population in area *A* is large
  - Statistical error associated with observed rate is small
  - High credibility (weight) is given to observed estimate
  - Smoothed rate is close to observed rate
- When population in area *A* is small
  - Statistical error associated with observed rate is large
  - Little credibility (low weight) is given to observed estimate
  - Smoothed rate is “shrunk” towards rate mean in surrounding areas

A Multi-level Model for Spatial Smoothing of SMR

$$
\begin{aligned}
Y_i \mid \mu_i &\sim \text{Poisson}(\mu_i) \\
\log \mu_i &= \log E_i + b_i \\
b_i \mid b_{j\neq i} &\sim N\left(\frac{\sum_{j\neq i} w_{ij} b_j}{\sum_{j\neq i} w_{ij}}, \sigma^2 \frac{1}{\sum_{j\neq} w_{ij}}\right)
\end{aligned}
$$

where:

- b_i are area-specific random effects with a spatially correlated random effect distribution
- w_{ij} are weights defining which regions j are neighbors to region i (by convention $w_{ii} = 0$, for all i)
- σ^2 is the variance controlling how similar the b_i is to its neighbors

Raw and Smoothed Standardized Mortality Rates

- Y_i are observed disease counts in area i
- E_i are expected disease counts in area i
- The raw and smoothed standardized mortality ratio (SMR_i and $\widehat{SMR_i}$) are so defined:

$$SMR_i = \frac{Y_i}{E_i}$$

$$\widehat{SMR_i} = \frac{\widehat{\mu}_i}{E_i}$$

- **In areas with abundant data:**

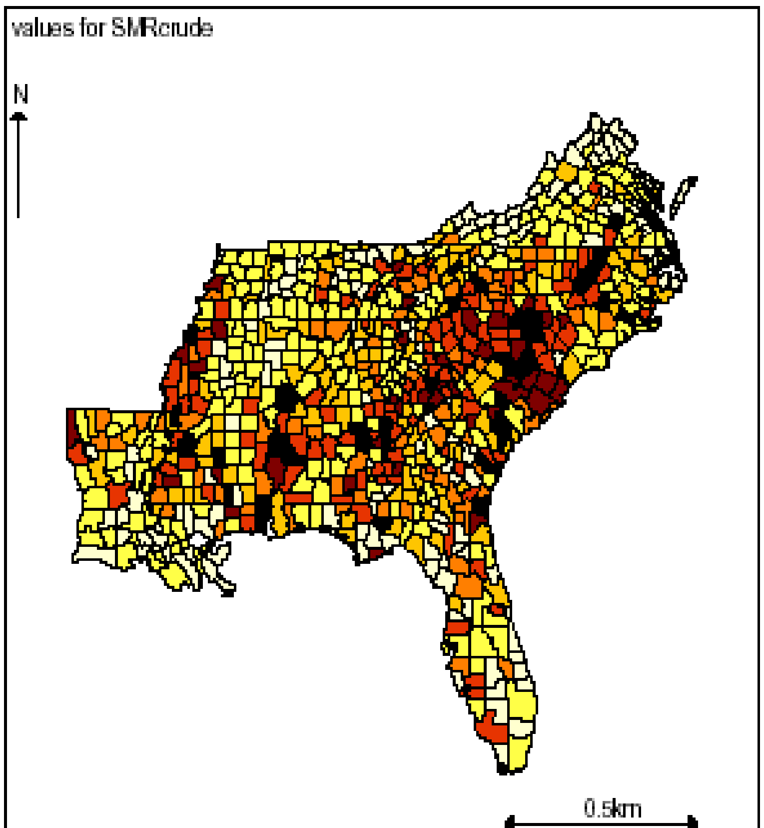
$$\widehat{SMR_i} \approx SMR_i$$

- **In areas with sparse data:**

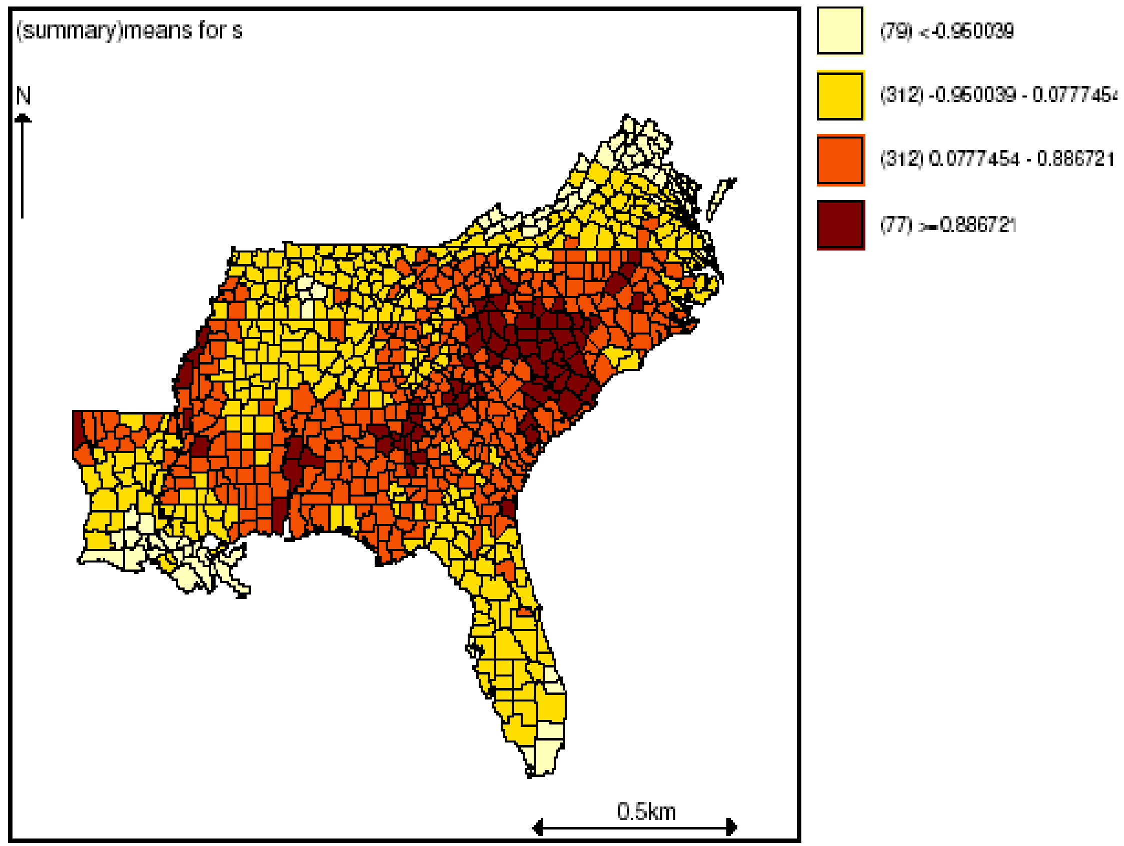
$\widehat{SMR_i} \approx$ weighted average of the SMR in the adjacent counties

SMR of pellagra deaths for 800 southern US counties in 1930

Crude SMR



Smoothed SMR



Multi-level Models for Geographical Correlation Studies

- Geographical correlation studies seek to describe the relationship between the geographical variation in disease and the variation in exposure

A Multilevel model for disease counts

- Y_{is} are observed disease counts in county i within state s
- E_{is} are expected disease counts in county i within state s
- **Stage I: County-level, within state model**

$$\begin{aligned} Y_{is} \mid \mu_{is} &\sim \text{Poisson}(\mu_{is}) \\ \log \mu_{is} &= \log E_{is} + \beta_{1s}(\text{cot}_{is} - \overline{\text{cot}}) + \beta_{2s}(\text{milk}_{is} - \overline{\text{milk}}) + b_i \\ b_i &\sim \text{spatially correlated random effects} \end{aligned}$$

- **Stage II: Between-states model**

$$\begin{aligned} \beta_{1s} &= \gamma_{11} + \gamma_{12}\text{state-taxes}_s + N(0, \sigma_1^2) \\ \beta_{2s} &= \gamma_{21} + \gamma_{22}\text{state-taxes}_s + N(0, \sigma_2^2) \end{aligned}$$

where:

- β_{1s} and β_{2s} are county-specific log-relative rates
- γ_{11} is the overall log-relative rate of pellagra mortality for the counties with average

Example: Scottish Lip Cancer Data
(*Clayton and Kaldor 1987 Biometrics*)

- Observed and expected cases of lip cancer in 56 local government district in Scotland over the period 1975-1980
- Percentage of the population employed in agriculture, fishing, and forestry as a measure of exposure to sunlight, a potential risk factor for lip cancer

Data Set

- `county`: county identifier 1:59
- `o`: observed number of lip cancer cases
- `e`: expected number of lip cancer cases
- `x`: percentage of the population working in agriculture, fishing or forestry

Note: the expected number of lip cancer cases for a county is based on the age-specific lip cancer rates for the whole Scotland and the age-distribution of the counties

Crude standardized Mortality rates for each district, Note that there is a tendency for areas to cluster, with a noticeable grouping of areas with SMR> 200 to the North of the country

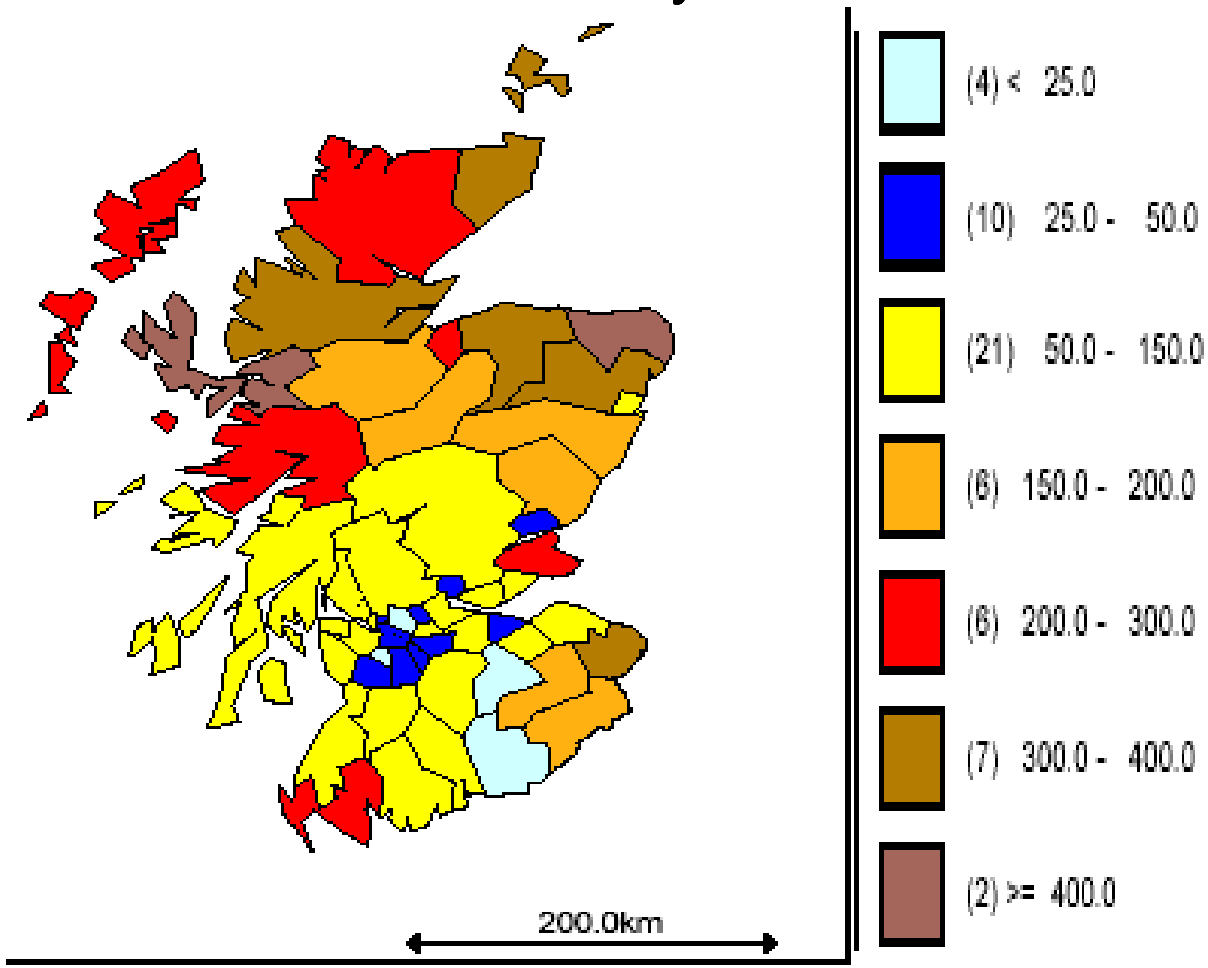


Table 6.2: Observed and expected numbers of lip cancer cases and various SMR estimates (in percentages) for Scottish counties

		Obs	Exp	Crude	Predicted SMRs	
County	#	o_j	e_j	SMR	Norm.	NPML
Skye, Lochalsh	1	9	1.4	652.2	470.7	342.6
Banf, Buchan	2	39	8.7	450.3	421.8	362.4
Caithness	3	11	3.0	361.8	309.4	327.1
Berwickshire	4	9	2.5	355.7	295.2	321.6
Ross, Cromarty	5	15	4.3	352.1	308.5	327.6
Orkney	6	8	2.4	333.3	272.0	311.1
Moray	7	26	8.1	320.6	299.9	322.2
Shetland	8	7	2.3	304.3	247.8	292.5
Lochaber	9	6	2.0	303.0	239.0	280.1
Gordon	10	20	6.6	301.7	279.1	319.9
W. Isles	11	13	4.4	295.5	262.5	315.5
Sutherland	12	5	1.8	279.3	219.2	254.3

Poisson model with random intercept

$$y_j \sim Poisson(m_j)$$

$$\log m_j = \log e_j + b_1 + V_j$$

$$V_j \sim N(0, t^2)$$

Unobserved heterogeneity between the counties

$$\log(SMR_j) = b_1 + V_j$$

$$SMR_j = m_j / e_j$$

We include an offset to that $\beta_1 + V_j$ can be interpreted as the county-specific log-SMR

Poisson model with random intercept and a covariate

$$y_j \sim Poisson(m_j)$$

$$\log m_j = \log e_j + b_1 + b_2 x_j + V_j$$

offset

$$V_j \sim N(0, t^2)$$

$$\log(SMR_j) = b_1 + V_j$$

$$SMR_j = m_j / e_j$$

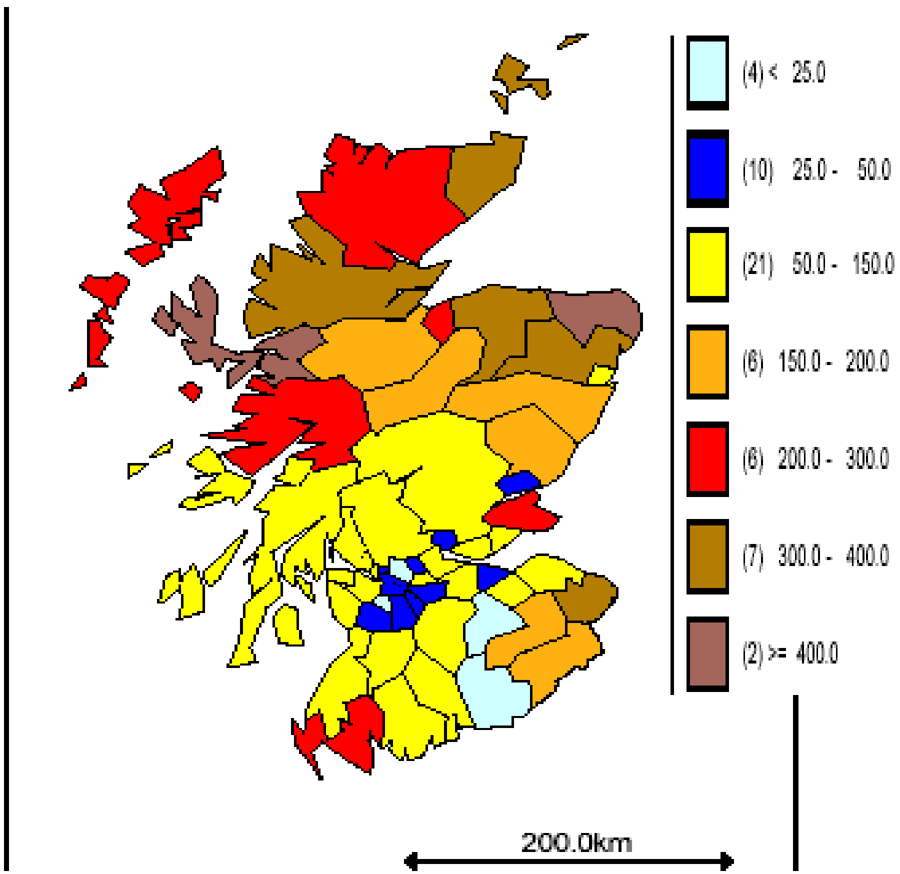
Parameter estimates

intercept	0.08	-0.49
	(SE = 0.12)	(SE = 0.16)
slope		0.07
		(SE = 0.01)
Variance	0.58	0.35
of the random effect	(SE = 0.15)	(SE = 0.1)

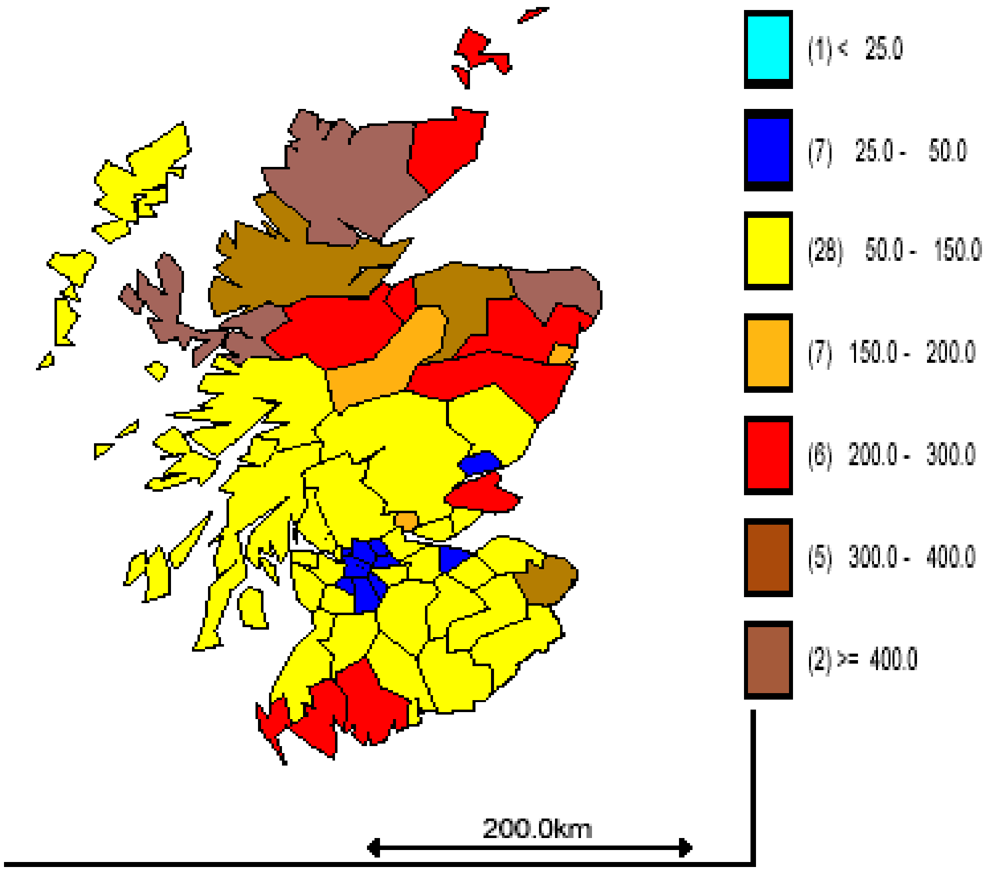
MLE obtained by using gllamm

Disease Mapping

Crude SMR



Smoothed SMR



The y=x line has been superimposed as well as the lines x=108 and y=108 representing the SMR in percent when the random effect is equal to zero. Shrinkage is apparent, since counties with particularly high crude SMRs lie below the y=x line (have predictions lower than the crude SMR) and counties with particularly low crude SMR lie above the y=x line

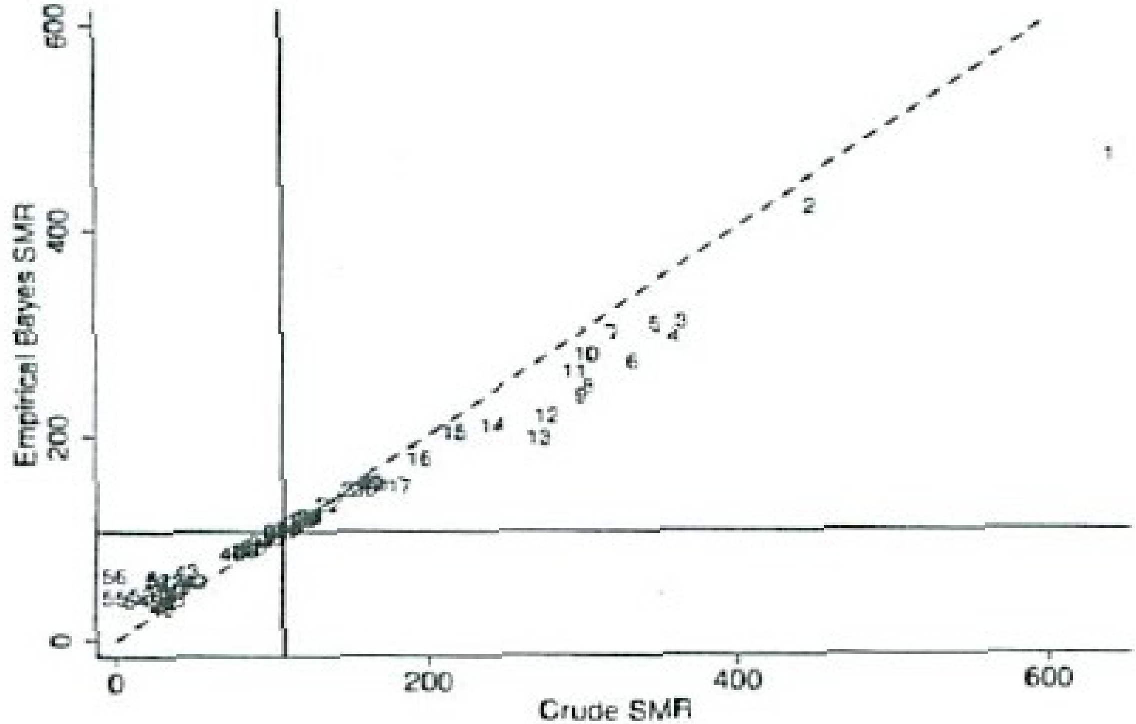


Figure 6.3: Empirical Bayes SMRs versus crude SMRs

Discussion

- In multi-level models is important to explore the sensitivity of the results to the assumptions inherent with the distribution of the random effects
- Specially for spatially correlated data the assumption of global smoothing, where the area-specific random effects are shrunk toward and overall mean might not be appropriate

Discussion

- Multilevel models are a natural approach to analyze data collected at different level of spatial aggregation
- Provide an easy framework to model sources of variability (within county, across counties, within regions etc..)
- Allow to incorporate covariates at the different levels to explain heterogeneity within clusters
- Allow flexibility in specifying the distribution of the random effects, which for example, can take into account spatially correlated latent variables

Key Words

- Spatial Smoothing
- Disease Mapping
- Geographical Correlation Study
- Hierarchical Poisson Regression Model
- Spatially correlated random effects
- Posterior distributions of relative risks