Introduction to disease mapping - Part 2

Bayesian modelling for spatial and spatio-temporal data

MSc in Epidemiology

Week 6

Aims of disease mapping

Disease maps used variously for:

- Descriptive purposes: to summarise spatial and spatio-temporal variation in disease risk $\rightarrow$ a visual summary of geographical risk;
- To generate etiological hypotheses: informal examination of disease maps with exposure maps (followed by formal examination via spatial regression);
- For surveillance, to highlight areas at apparently high risk;
- To aid policy formation and resource allocation.

What to map?

- **Mortality**
  - most readily available source of data for all diseases
  - should be complete and relatively accurate
- **Morbidity**
  - **Incidence**, that is the number of new cases within a specific time frame
    - incidence data usually routinely available for cancer registries
  - **Prevalence**, that is the total number of existing cases over specific time frame
    - available from registries and hospital admissions

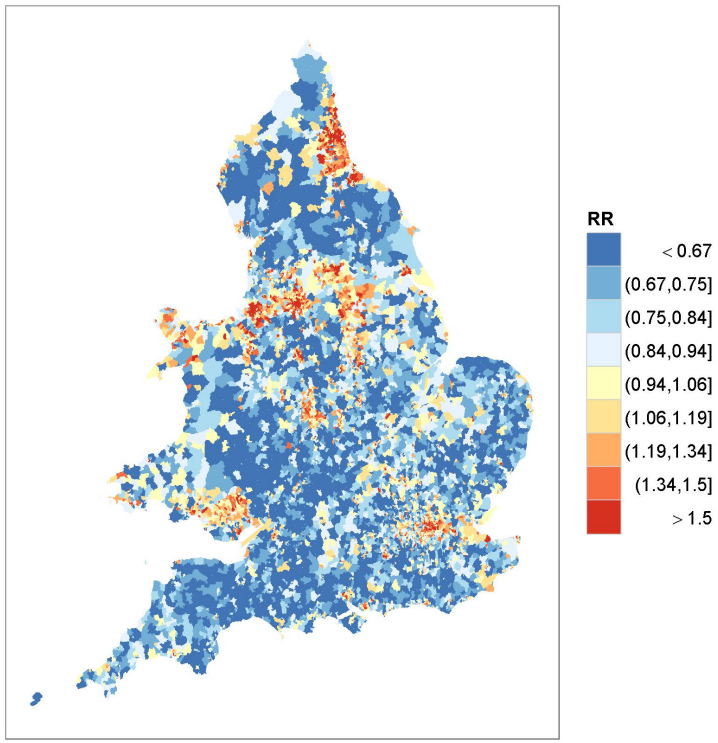
Geographical scales

Disease mapping may be carried out at a variety of geographical scales:

- **International**:
  - Comparisons between countries, e.g World Health Organization (WHO) or International Agency for Research on Cancer (IARC) reports
- **National**:
  - Comparisons between e.g. regions, districts
  - Most published disease atlases fall into this category
- **Small area studies**
  - Sub-national scale, e.g. wards, municipalities

Small area case study: lung cancer incidence in males - England and Wales

Map of the SMRs for the period 1985-2009 at ward level



- Is the variability real or simply reflecting unequal E_is?
- Have the highlighted areas truly a raised relative risk?

	Min	Q1	Median	Q3	Max
O	0	26	47	84	456
E	3.25	32.14	53.60	82.47	390.49
SMR	0	0.70	0.89	1.13	2.63

Problems with mapping SMRs

- It is common practice to map SMRs, however:
  - SMR_i very imprecise for rare diseases and/or areas with small populations,
  - For the model $O_i \sim \text{Poisson}(\lambda_i E_i)$, the MLE is $\text{SMR}_i = \frac{O_i}{E_i}$, and its estimated variance is $\hat{\text{Var}}(\hat{\lambda}) = \frac{O_i}{E_i^2}$,
  - areas with small E_i have high associated variance.
- SMR in each area is estimated independently:
  - $\rightarrow$ makes no use of risk estimates in other areas of the map, even though these are likely to be similar.

$\Rightarrow$ Highlights extreme risk estimates based on small numbers, thus risks tend to be unstable.

$\Rightarrow$ Ignores possible spatial correlation between disease risk in nearby areas due to possible dependence on spatially varying risk factors.

Smoothing

- One method for addressing these problems and reducing the noise in rates or risks associated with geographical areas is the 'spatial smoothing'.

- The basic idea is to borrow information from neighbouring areas to produce a better (i.e. more stable and less noisy) estimate of the risk associated with each area and thus separate out the signal from the noise.

- We will use Bayesian 'smoothing' estimators in a hierarchical formulation:
  - Poisson-lognormal model: non-spatial (global) smoothing
  - Poisson-lognormal spatial model: spatial and non-spatial smoothing

Poisson-lognormal model (non-spatial smoothing)

Model:

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\lambda_i E_i) \\
\log(\lambda_i) &= \alpha + \theta_i \\
\theta_i &\sim \text{Normal}(0, \sigma_\theta^2) \\
\alpha &\sim \text{Normal}(0, 10^4)
\end{aligned}
$$

- O_i and E_i are observed and expected number of cases in area i
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ are *exchangeable parameters* and each θ_i follows the same prior distribution, but, in addition, we assign a prior distribution (hyperprior) to the unknown parameter of that prior distribution: $1/\sigma_\theta^2 \sim \text{Gamma}(1, 0.01)$

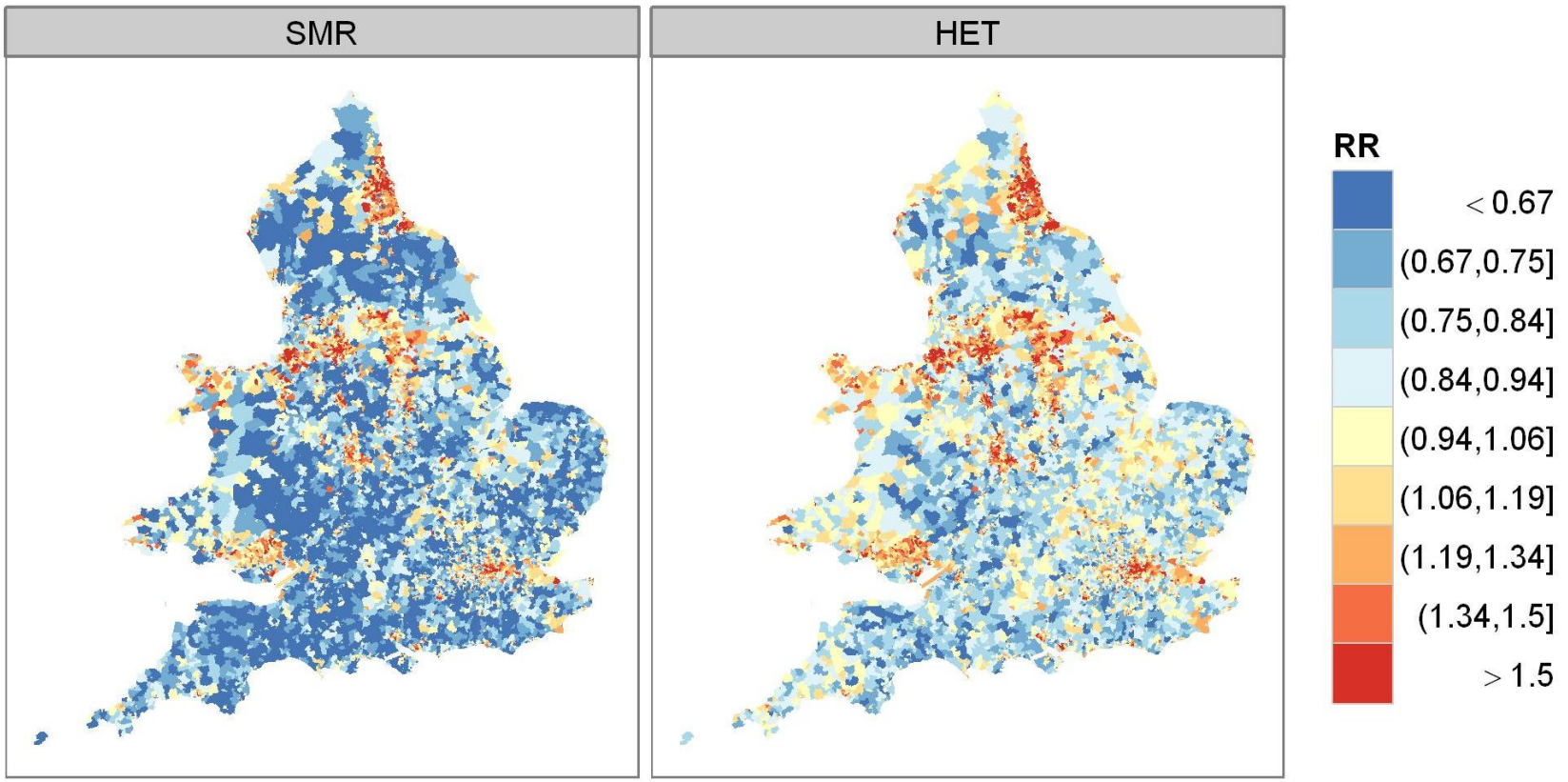
Recall: the reciprocal of the variance is the precision $1/\sigma_\theta^2 = \tau_\theta$

Parameter interpretation

- $\lambda_i = \exp(\alpha + \theta_i)$: unknown RR in area *i* compared with expected risk based on age and sex of population
- Parameter α: mean log relative risk, i.e. overall risk in the region of study
- Parameters θ_i: **area-specific random effects** to capture region-wide *heterogeneity*. They take into account the extra-Poisson variability, i.e. overdispersion, in the log-relative risks. The excess variation can be due to:
  - errors in numerators (observed counts) and denominators (expected counts)
  - unobserved or unknown risk factors (confounders)
- residual RR = $\exp(\theta_i)$
- σ^2_θ: is between-area variance and controls the magnitude of the θ_i

Mapping smoothed *vs* raw estimates of λ_i

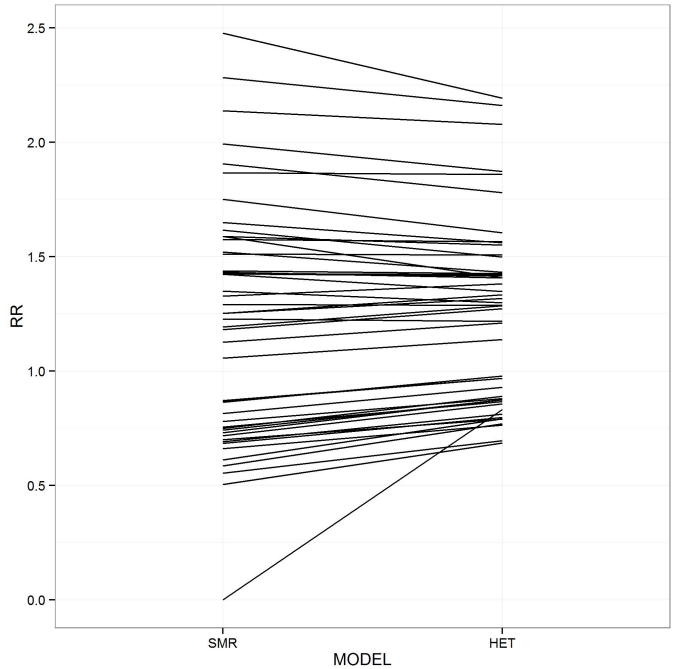
Lung cancer incidence in males, 1985-2009, England and Wales



SMRs and non-spatially smoothed RRs

Note: It is important to keep the same cut points when comparing maps

SMRs versus non-spatially smoothed RRs in selected areas for lung cancer incidence in males, 1985-2009, England and Wales



- Shrinkage towards the mean when using a hierarchical model
- Posterior mean and 95% credible intervals (CI) = 1.0 (95%CI: 1.09-1.1)

Binomial framework

- For more common diseases, Binomial model may be preferred

$$O_i \sim \text{Binomial}(p_i, n_i)$$

  where
  - n_i = population at risk
  - p_i = probability of disease
- $\text{logit}(p_i) = \alpha + \theta_i$
- Parameter of interest: Odds ratio $\text{OR}_i = \exp(\alpha + \theta_i)$

A few notes on ecological bias [1]

- In our module, we work often with aggregated data
- Aggregation has implications for the type of inference that are possible
- Ecological inference is the process where aggregated data are used to infer individual level relationships. Reasons:
  - individual data are not available for confidentiality reasons
  - individual data are not reliable or too expensive to be collected
- Ecological (or aggregation) bias in the difference between the estimates of relationships obtained using grouped data and those estimates obtained using individual data (e.g. the association observed at the area level do not hold for the individuals within areas).

A few notes on ecological bias [2]

- Ecological bias can manifest itself in a variety of ways and results in an information loss (e.g. it can be due to a model specification bias that arises because a nonlinear risk model changes its form under aggregation). Researchers should specify the conditions under which the estimates are reasonable.

- The converse, using individual level estimates uncritically to infer group level relationships (ignoring the possibility of group level or contextual effects) is called as atomistic or individualistic fallacy.

Reference

- Banerjee S., Carlin B.P, Gelfand A. (2014), Hierarchical Modeling and Analysis for Spatial Data. (2nd ed.) CRC Press - Section 6.4.1 (partially Section 6.4.3.2)

- Haining R., Guangquan L. (2020), Modelling Spatial and Spatial-Temporal Data. A Bayesian Approach, CRC Press - Section 7.4.2