Bayesian Hierarchical Models

Gavin Shaddick,

Millie Green, Matthew Thomas

University of Bath

6^{th} - 9^{th} December 2016

APPLICATIONS OF BAYESIAN HIERARCHICAL MODELS

OUTLINE

Spatial epidemiology

What is Spatial Epidemiology?

- Epidemiology is the study of the distribution of diseases in human populations.
- Disease risk depends on the classic epidemiological triad of person (genetics/behaviour), place and time.
- Spatial epidemiology focuses on the second of these.
- Place is a surrogate for exposures present at that location
  - environmental exposures in water/air/soil
  - lifestyle characteristics of those living in particular areas.

GROWING INTEREST IN SPATIAL EPIDEMIOLOGY

- Public interest in effects of environmental hazards/pollution.
- Epidemiological interest in differences in disease rates across different areas.
- Data availability: collection of health data at different geographical scales.
- Increase in computing power and methods
  - Geographical Informations Systems (GIS).
- Development of statistical/epidemiological methods for investigating disease 'clusters'.

The Need for Spatial Methods

- Many epidemiological studies are spatial
  - many are spatio-temporal!
- When do we need to 'worry'?
  - acknowledge the spatial component
  - are we explicitly interested in the spatial pattern of disease incidence?
    - disease mapping
    - cluster detection.
  - is the clustering a nuisance quantity that we wish to acknowledge, but are not explicitly interested in?
    - spatial regression.

Types of Spatial Data

- Point data
  - 'exact' residential locations exist for cases and controls.
- Count data
  - aggregation
  - typically over administrative units
  - ecological, in that they are collected across groups
  - in spatial studies the groups are geographical areas.

Disease mapping

Overview of Disease Mapping

- 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
  - estimation of background variability in underlying risk in order to place epidemiological studies in context.
- 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.

Example: Lip Cancer in Scotland

- Incidence rates of lip cancer in males in 56 counties of Scotland, registered in 1975–1980.
- Data
  - observed and 'expected' number of cases (based on the county age populations)
  - this allows the calculation of the standardised morbidity ratio
    - ratio of the observed to the expected cases.
  - a covariate measuring the proportion of the population engaged in agriculture, fishing, or forestry (AFF)
  - the projections of the longitude and latitude of the area centroid, and the 'position' of each county expressed as a list of adjacent counties.

EXAMPLE: LIP CANCER IN SCOTLAND

County No. i	Obs Cases Y_i	Exp Cases E_i	Prop AFF	SMR	Project N (km)	Project E (km)	Adjacent Counties
1	9	1.4	0.16	6.43	834.7	162.2	5,9,19
2	39	8.7	0.16	4.48	852.4	385.8	7,10
3	11	3.0	0.10	3.67	946.1	294.0	12
4	9	2.5	0.24	3.60	650.5	377.9	18,20,28
5	15	4.3	0.10	3.49	870.9	220.7	1,12,19
6	8	2.4	0.24	3.33	1015.2	340.2	Island
7	26	8.1	0.10	3.21	842.0	325.0	2,10,13,16,17
8	7	2.3	0.07	3.04	1168.9	442.2	Island
9	6	2.0	0.07	3.00	781.4	194.5	1,17,19,23,29
...							
47	2	5.6	0.01	0.36	640.8	277.0	24,31,46,48,49,53
48	3	9.3	0.01	0.32	654.7	282.0	24,44,47,49
49	28	88.7	0.00	0.32	666.7	267.8	38,41,44,47,48,52,53,54
50	6	19.6	0.01	0.31	736.5	342.2	21,29
51	1	3.4	0.01	0.29	678.9	274.9	34,38,42,54
52	1	3.6	0.00	0.28	683.7	257.8	34,40,49,54
53	1	5.7	0.01	0.18	646.6	265.6	41,46,47,49
54	1	7.0	0.01	0.14	682.3	267.9	34,38,49,51,52
55	0	4.2	0.16	0.00	640.1	321.5	18,24,30,33,45,56
56	0	1.8	0.10	0.00	589.9	322.2	18,20,24,27,55

EXAMPLE: LIP CANCER IN SCOTLAND

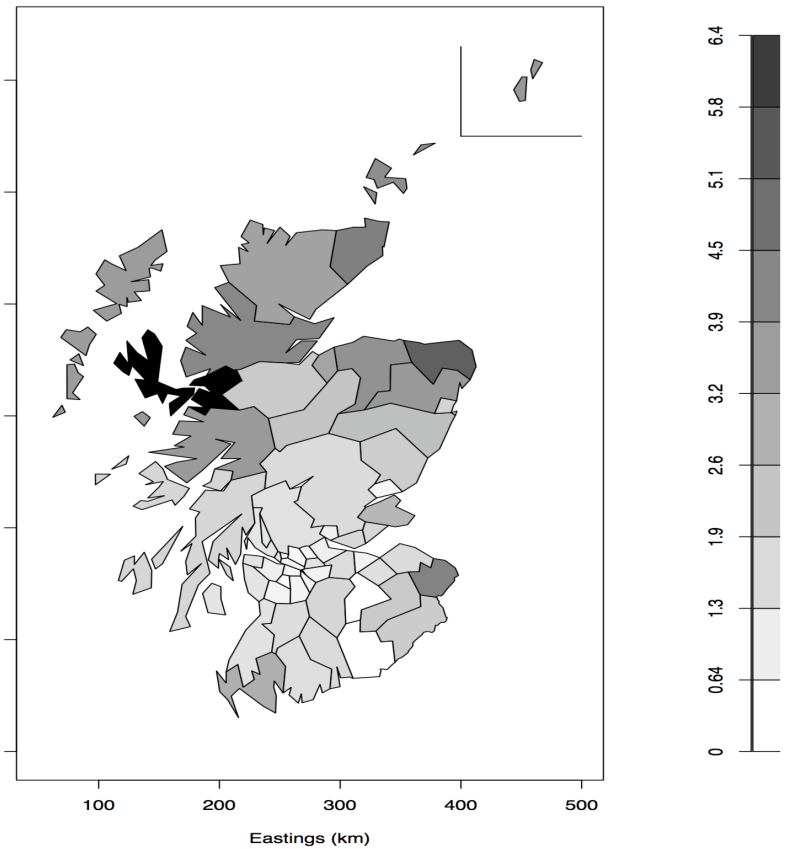


Figure: SMRs for male lip cancer in 56 counties of Scotland.

Example: Lung and Brain Cancer in North-West England

- Two tumors
  - one non-rare (lung)
  - and one rare (brain).
- Study period is 1981–1991.
- Analysis performed at ward level (144 wards)
  - incidence data by postcode.
- Brain cancer
  - the median number of cases per ward over the 11 year period is 6
  - range of 0 to 17.
- Lung cancer
  - the median number is 20
  - range 0–60.
- 'Expected counts' were based on ward-level populations from the 1991 census, by 5-year age bands and sex.

Example: Lung and Brain Cancer in North-West England

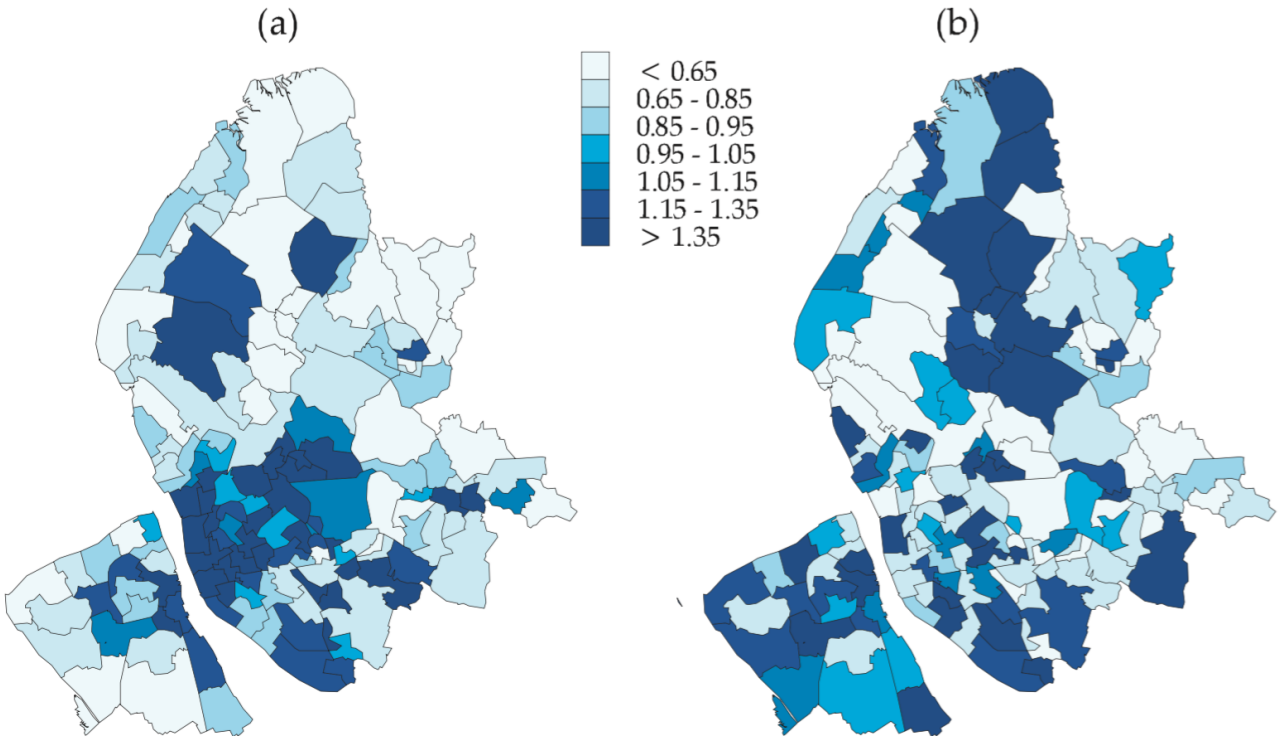


Figure: SIRs for (a) lung cancer, and (b) brain cancer in the North-West of England.

NON-STATISTICAL ISSUES

- There is a trade-off when a geographical scale is chosen
  - larger geographical areas provide more stable rates
  - 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
  - high relative risks will be diluted under aggregation.
- The size of the areas chosen also determines the sort of questions that can be posed
  - larger areas often mean greater contrasts in relative risks and exposures
  - localized effects can only be detected with data at a smaller level of aggregation.

Statistical Issues

- 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

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

statistically, the best estimate of the rate $\widehat{\theta_i}$ will be

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

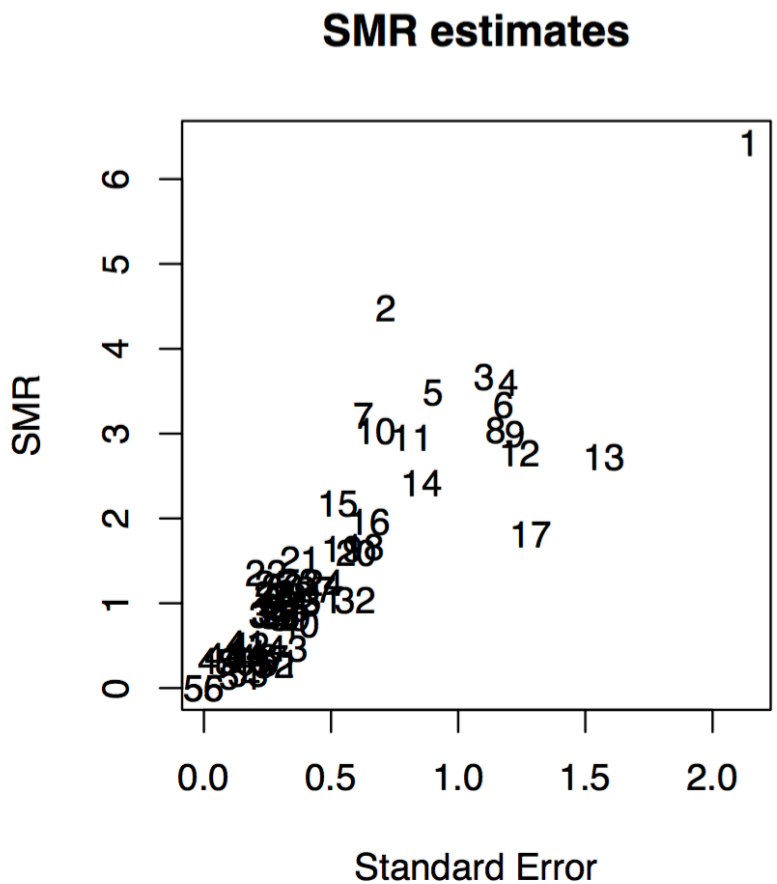
with variance

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

so that areas with small E_i have high associated variance.

EXAMPLE: SCOTTISH LIP CANCER

- The variance of the estimate is $\text{var}(\text{SMR}_i) = \text{SMR}_i/E_i$.
- This 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.


SMR estimates
SMR
Standard Error
0
1
2
3
4
5
6
0.0
0.5
1.0
1.5
2.0

Smoothing Models

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

Statistical methods for smoothing risks

Spatial Models

- In general we might expect 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.

- Consider the model

$$Y_i|\beta,\gamma,\mathbf{U_i},\mathbf{V_i} \sim \text{Poisson}(\mathbf{E_i}\mu_\mathbf{i}\mathbf{e}^{\mathbf{U_i}+\mathbf{V_i}})$$

$$\log \mu_i = g(\mathbf{S}_i,\gamma) + \mathbf{f}(\mathbf{x_i},\beta),$$

- $\mathbf{S}_i = (S_{i1}, S_{i2})$ denotes spatial location, the centroid of area i,
- $f(\mathbf{x_i},\beta)$ is a regression model,
- $g(\mathbf{S}_i,\gamma)$ is an expression that we may include to capture large-scale spatial trend, e.g. long-term spatial trend

$$f(\mathbf{S}_i) = \gamma_1 S_{i1} + \gamma_2 S_{i2},$$

- The random effects $V_i \sim_{iid} N(0,\sigma_v^2)$ represent non-spatial overdispersion,
- U_i are random effects with spatial structure.

- One approach is to specify the distribution of the random effect for a particular area, U_i, as if we knew the values of the spatial random effects, U_j, in 'neighbouring areas'
- We therefore need to specify a rule for determining the 'neighbours' of each area.
- Commonly areas i and j are taken to be neighbours if they share a *common boundary*.

- Various other neighbourhood/weighting schemes are possible.
- The neighbourhood structure could depend on the distance between area centroids.
- Determine the distance within which regions are considered neighbours.

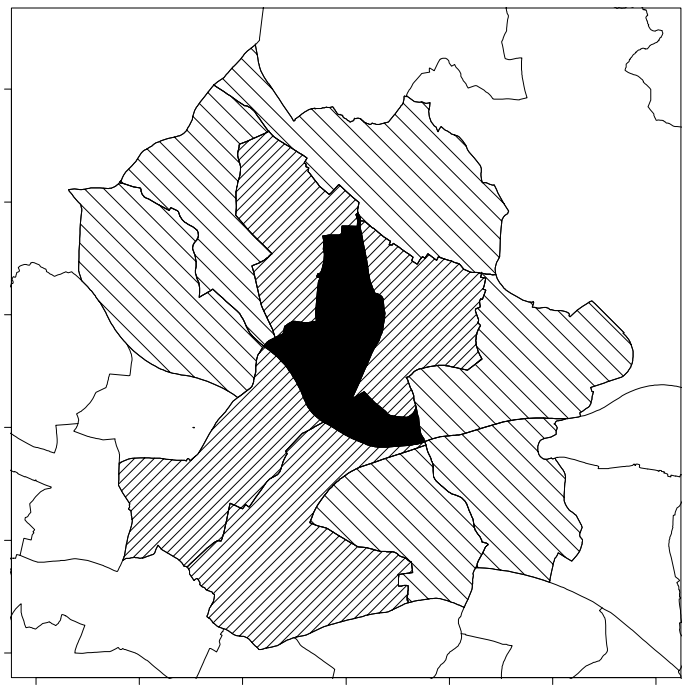


Figure: Close-up of a region of Birmingham.

The ICAR model

- A common model is to assign the spatial random effects an intrinsic conditional autorgressive (ICAR) prior.
- Under this specification it is assumed that

$$U_i | U_j, j \in \partial_i \sim N\left(\overline{U}_i, \frac{\omega_u^2}{m_i}\right),$$

  where ∂i is the set of neighbours of area i, m_i is the number of neighbours, and $\overline{U}_i$ is the mean of the spatial random effects of these neighbours.
- The parameter ω_u^2 is a conditional variance and its magnitude determines the amount of spatial variation.
- The variance parameters σ_v^2 and ω_u^2 are on different scales, σ_v is on the log odds scale while ω_u is on the log odds scale, *conditional* on $U_j, j \in \partial i$; hence they are not comparable.

- Notice that if ω_u^2 is 'small' then although the residual is strongly dependent on the neighboring value the overall contribution to the residual relative risk is small.
- This is a little counterintuitive but stems from spatial models having two aspects, strength of dependence and total amount of spatial dependence, and in the ICAR model there is only a single parameter which controls both aspects.

Recall the model

$$Y_i|\beta,\gamma,U_i,V_i \sim_{ind} \text{Poisson}(E_i\mu_i e^{U_i+V_i})$$

with

$$\log\mu_i = g(\mathbf{S}_i,\gamma) + f(\mathbf{x_i},\beta), \tag{1}$$

where

- $\mathbf{S}_i = (S_{i1}, S_{i2})$ denotes spatial location, the centroid of area i,
- $f(\mathbf{x_i},\beta)$ is a regression model,
- $g(\mathbf{S}_i,\gamma)$ is an expression that we may include to capture large-scale spatial trend – the form

$$f(\mathbf{S}_i) = \gamma_1 S_{i1} + \gamma_2 S_{i2},$$

  is a simple way of accommodating long-term spatial trend.
- The random effects $V_i \sim_{iid} N(0,\sigma_v^2)$ represent non-spatial overdispersion,
- U_i are random effects with spatial structure which we now consider 'jointly' rather than considering neighbours.

A Joint Model

- Assume that $\mathbf{U} = (U_1, ..., U_n)$ arise from a zero mean multivariate normal distribution with variances $\text{var}(U_i) = \sigma_u^2$ and correlations $\text{corr}(U_i, U_j) = \exp(-\phi d_{ij}) = \rho^{d_{ij}}$ where d_{ij} is the distance between the centroids of areas i and j, and $\rho > 0$ is a parameter that determines the extent of the correlation.
- This model is *isotropic* since it assumes that the correlation is the same in all spatial directions. We refer to this as the *joint* model, since we have specified the joint distribution for $\mathbf{U}$.
- More generally the correlations can be modeled as $\text{corr}(U_i, U_j) = \exp(-(\phi d_{ij})^{\kappa})$.

Bayesian Analysis

- Inference for models with a spatial component is often not straightforward using likelihood based approaches, 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.
- Markov chain Monte Carlo (MCMC)
  - `Win/OpenBUGS` is a package that allows very general Bayesian modeling
  - `GeoBUGS` module contains a number of useful spatial models, and mapping facilities
  - Packages in `R`, for example `CARBayes`.
- Approximate Bayesian inference, for fast computation on big datasets
  - `R-INLA`.

EXAMPLE: LIP CANCER IN SCOTLAND

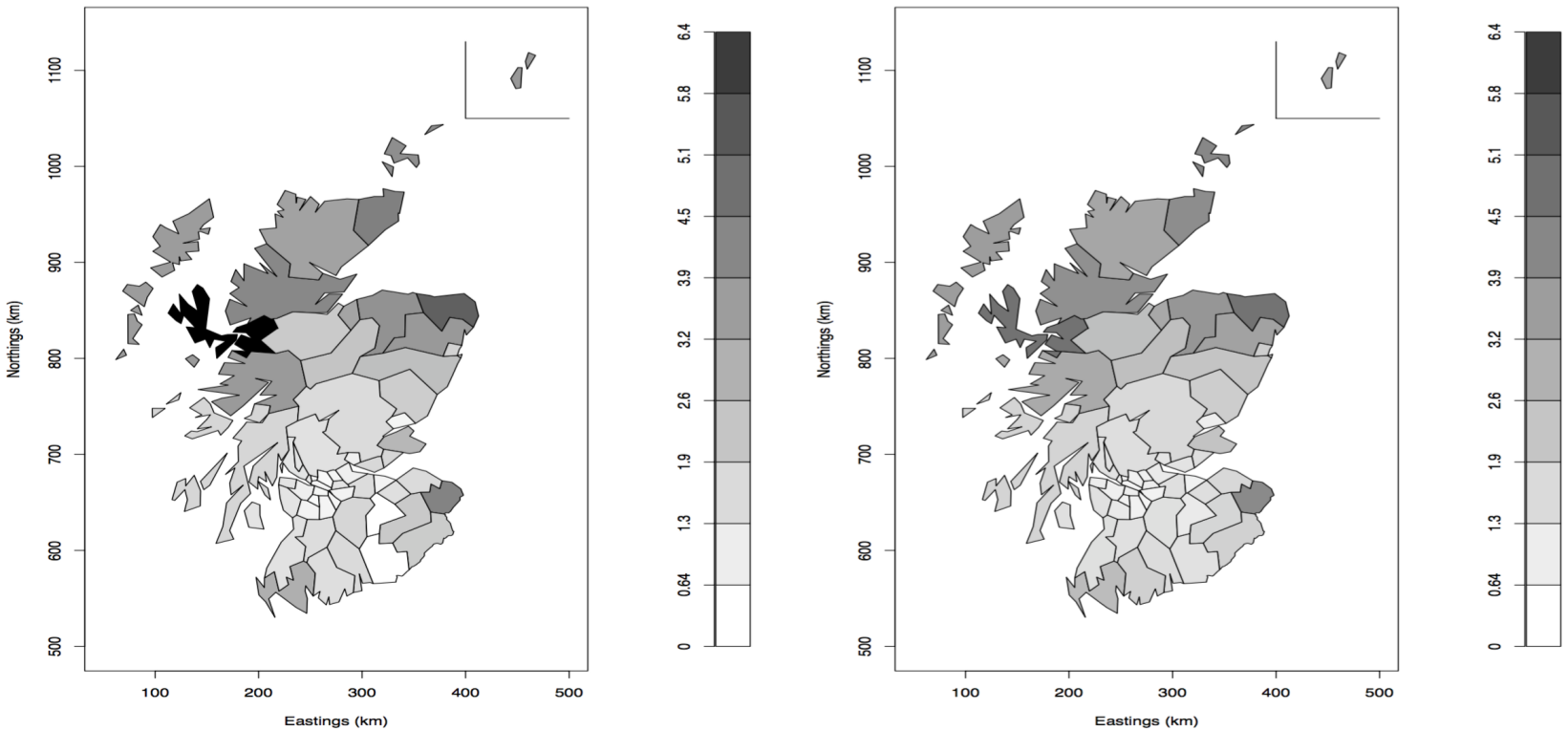


Figure: (a) Unsmoothed SMRs and (b) Smoothed SMRs for lung cancer in the North-West of England.

Example: Lung and Brain Cancer in North-West England

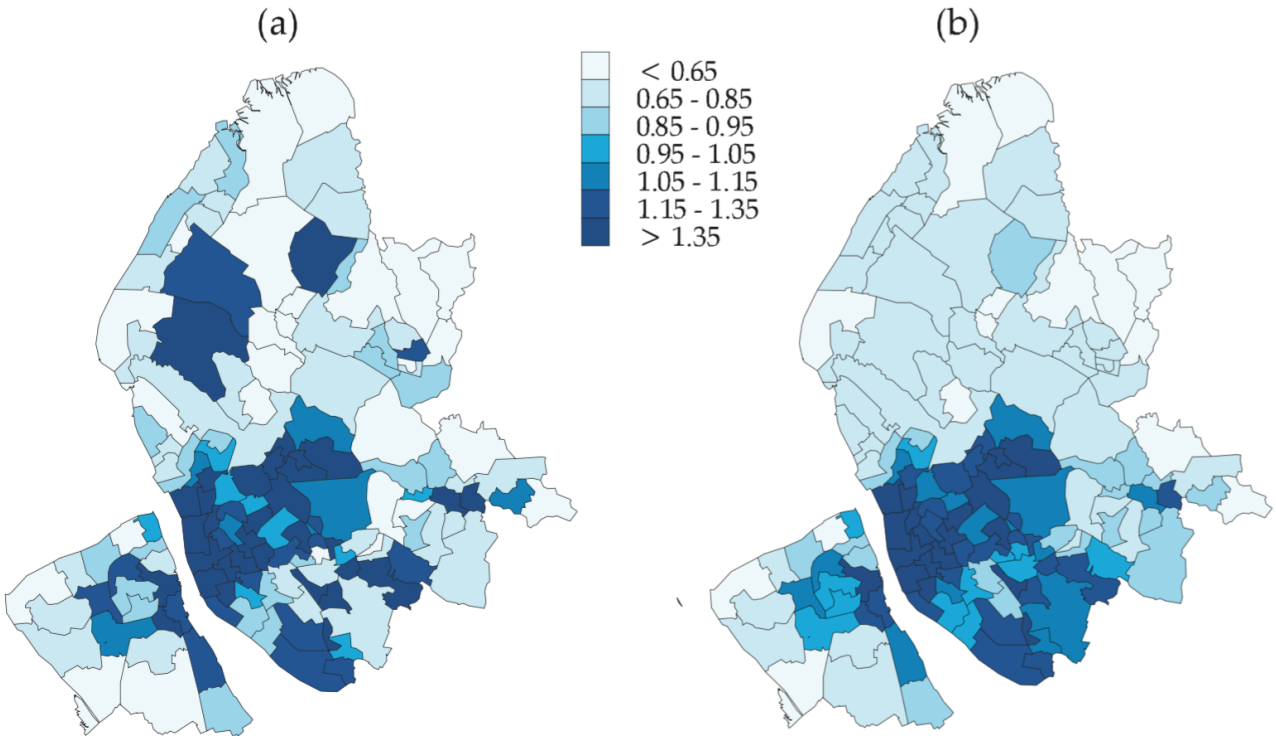


Figure: (a) Unsmoothed SIRs and (b) Smoothed SIRs for lung cancer in the North-West of England.

Example: Lung and Brain Cancer in North-West England

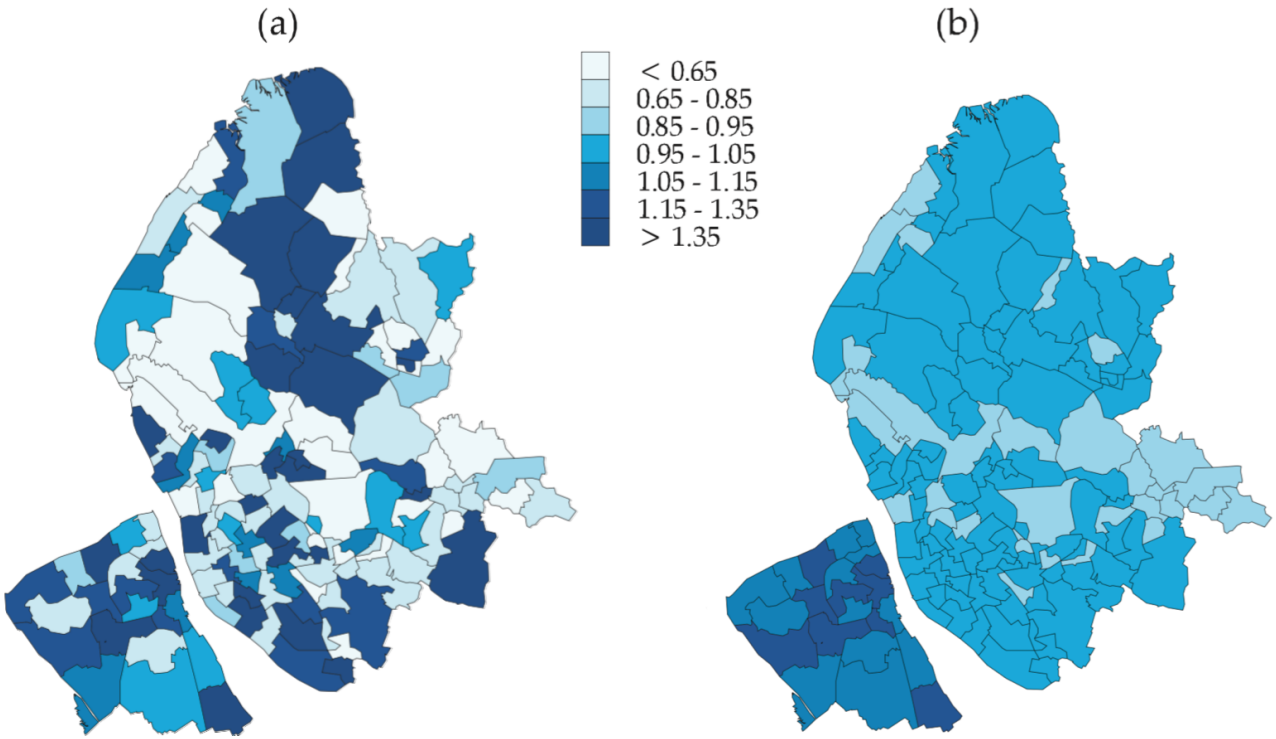


Figure: (a) Unsmoothed SIRs and (b) Smoothed SIRs for brain cancer in the North-West of England.

Summary of Smoothing in 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.
- Comparing SMRs with smoothed estimates is important
  - if there are big changes, are they appropriate?
  - were the expected numbers small?
  - did the regression model fit this area well?