Bayesian methods for small area estimation and spatial analysis

Juanjo Abellan

Small Area Health Statistics Unit (SAHSU)

Department of Epidemiology & Public Health

Imperial College London

ESRC Research Methods Festival, Oxford, 17-20 July 2006

Outline

- Part I: Bayesian small area estimation in Epidemiology:
  - Disease mapping
  - A Bayesian approach
- Part II: Bayesian Spatial analysis of SES indices
  - The English indices of deprivation (2004)
  - Spatial analysis of income, employment and education domains.

Part I:
Bayesian small area estimation in Epidemiology

Disease mapping

- Aim: to analyse the geographical variation of disease risk
- Geographical data: region of interest divided into a certain number N of areas.
- Data on the disease of interest:
  - Number of observed cases O_i, $i=1,\ldots,N$.
  - Number of expected cases E_i, $i=1,\ldots,N$.
- Disease risk measure: Standardised Incidence Ratio $SIR_i=O_i/E_i$, $i=1,\ldots,N$.

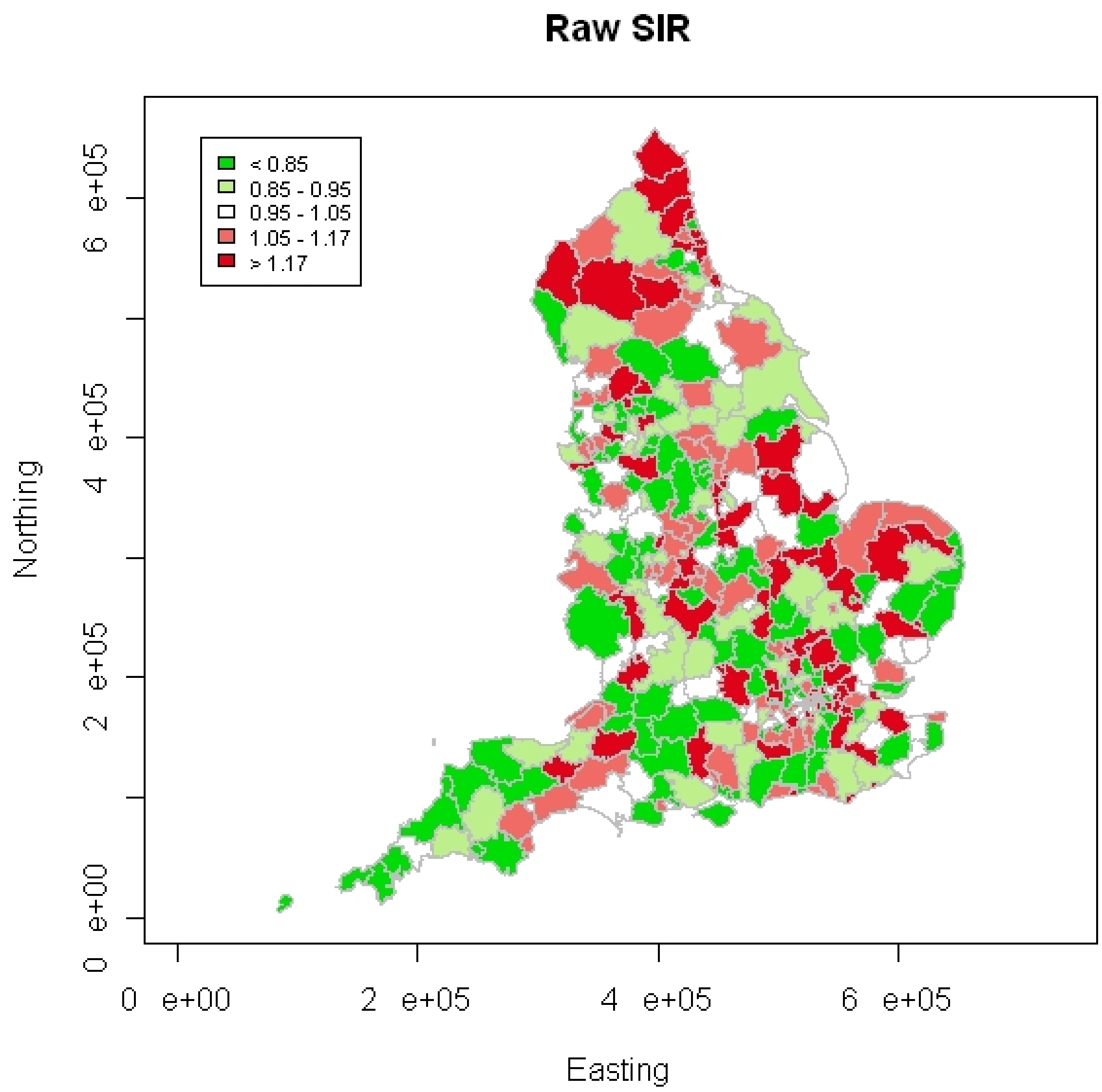
Disease mapping: an example

Disease: Hip fracture incidence in males aged 85+

Period: 1998-2000

Geographical level: District

Risk indicator: $SIR_i = O_i / E_i$



Statistical formulation

If we assume:

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

Then, the Maximum Likelihood Estimator is:

$$\hat{\lambda}_i = \frac{O_i}{E_i} = \text{SIR}_i ; \text{V}\left[\hat{\lambda}_i\right] = \frac{\text{SIR}_i}{E_i}$$

$$\text{CI95\%}(\lambda) = \hat{\lambda} \pm 1{,}96\sqrt{\frac{\hat{\lambda}}{E}} = \left[\hat{\lambda}_l , \hat{\lambda}_u\right]$$

Uncertainty: $1 \in \left[\hat{\lambda}_l , \hat{\lambda}_u\right]?$

A few problems...

- When working with small areas or rare diseases:
  - SIR=O/E may become unstable: extreme risks associated to low populated areas
  - SE(SIR) proportional to 1/E: significant risks associated to highly populated areas
  - Areas next to each other might show completely opposite risks

For example...

Total Pop	O	E	SIR	CI95% Lower	CI95% Upper
184 201	164	151.50	**1.08**	**0.92**	**1.26**
523	0	2.27	**0.00**	**0.00**	**11.12**

A Bayesian approach

Besag *et al* (1991) proposed the following Bayesian hierarchical model:

1st layer:
<u>Model</u>

$$\begin{cases} O_i \sim \text{Poisson}(\lambda_i E_i) \\ \log(\lambda_i) = \alpha + u_i + v_i \end{cases}$$

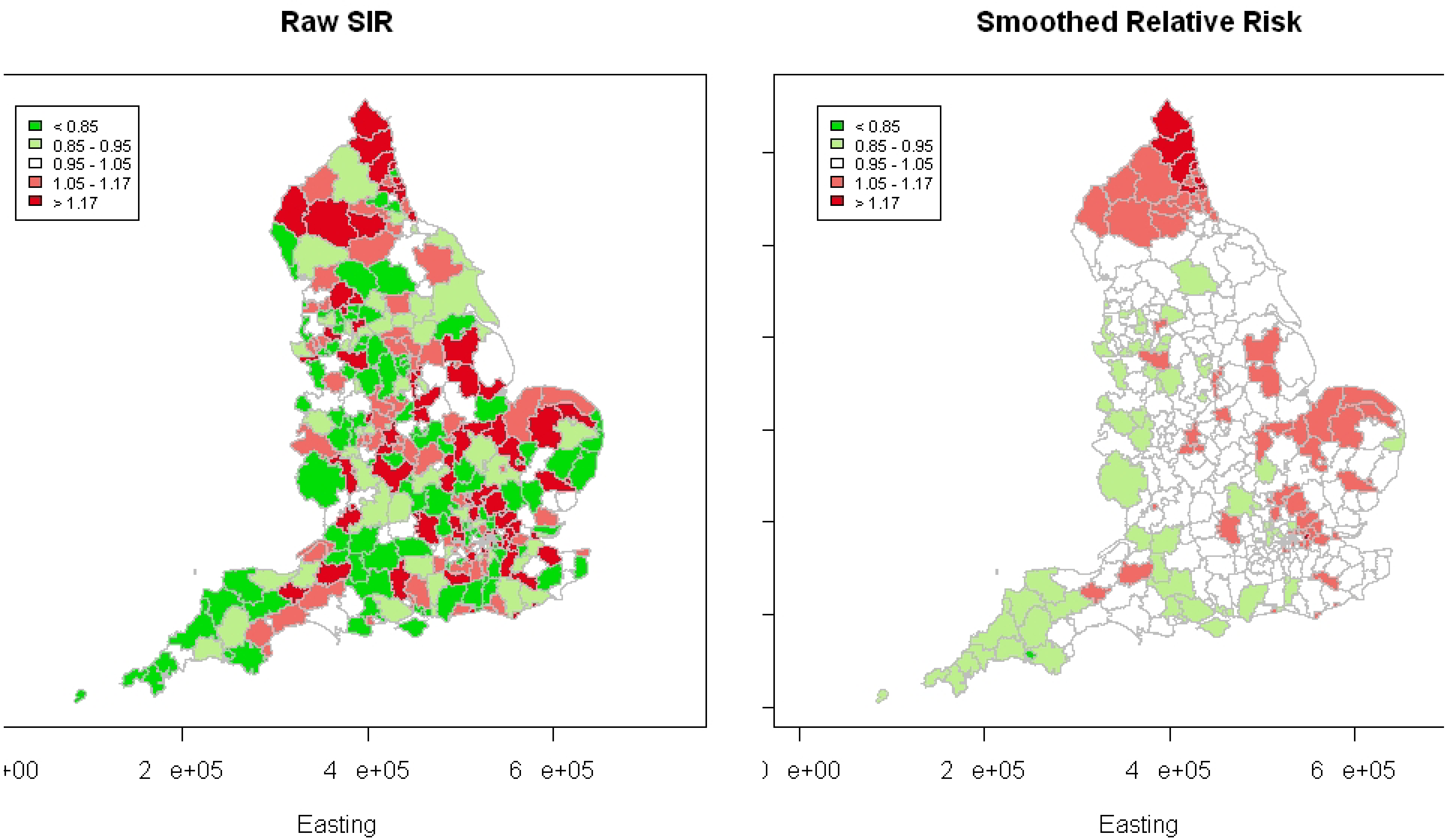
2nd layer:
<u>Priors</u>

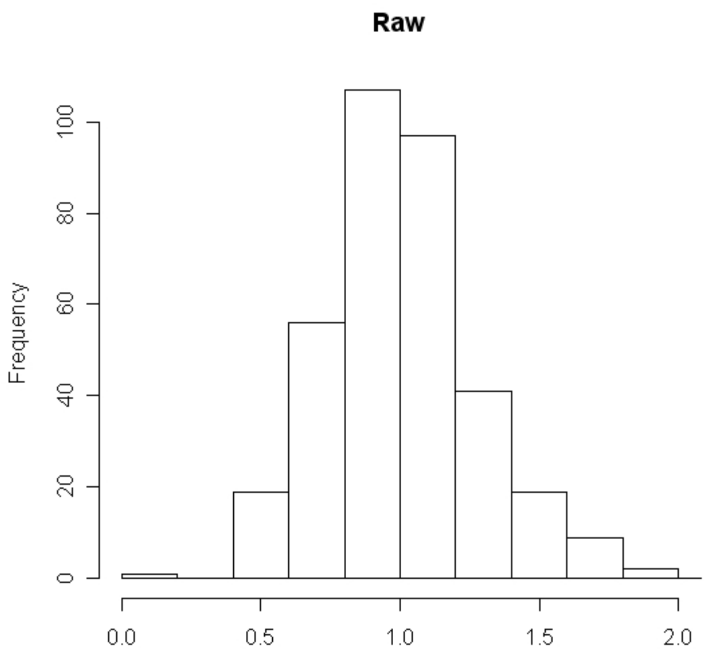
$$\begin{cases} \alpha \propto 1 \\ u_i \sim \text{Normal}(0, \sigma_{\text{u}}^2) \\ v_i \mid \mathbf{v}_{-i} \sim \text{Normal}\left(\sum_{j \sim i} v_j / n_i, \sigma_v^2 / n_i\right) \end{cases}$$

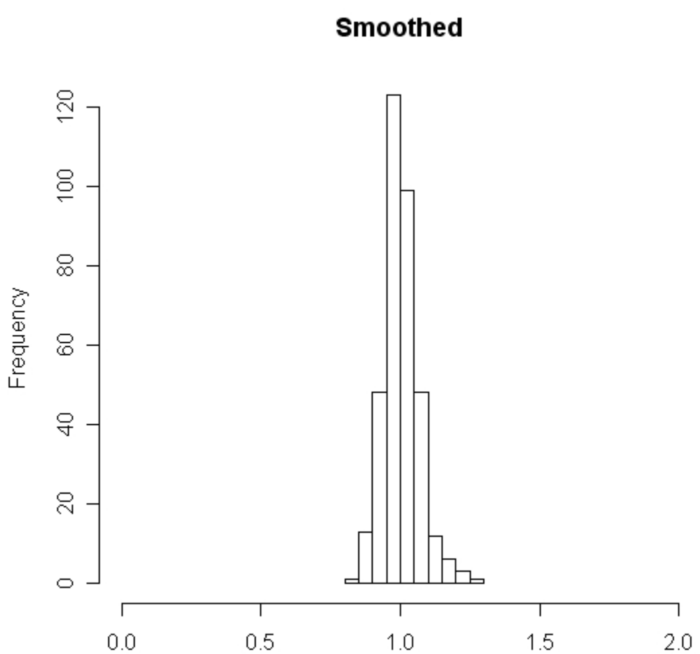
3rd layer:
<u>Hyperpriors</u>

$$\left\{ \sigma_u^2, \sigma_v^2 \sim \text{IGamma}(0.5, 0.0005) \right.$$

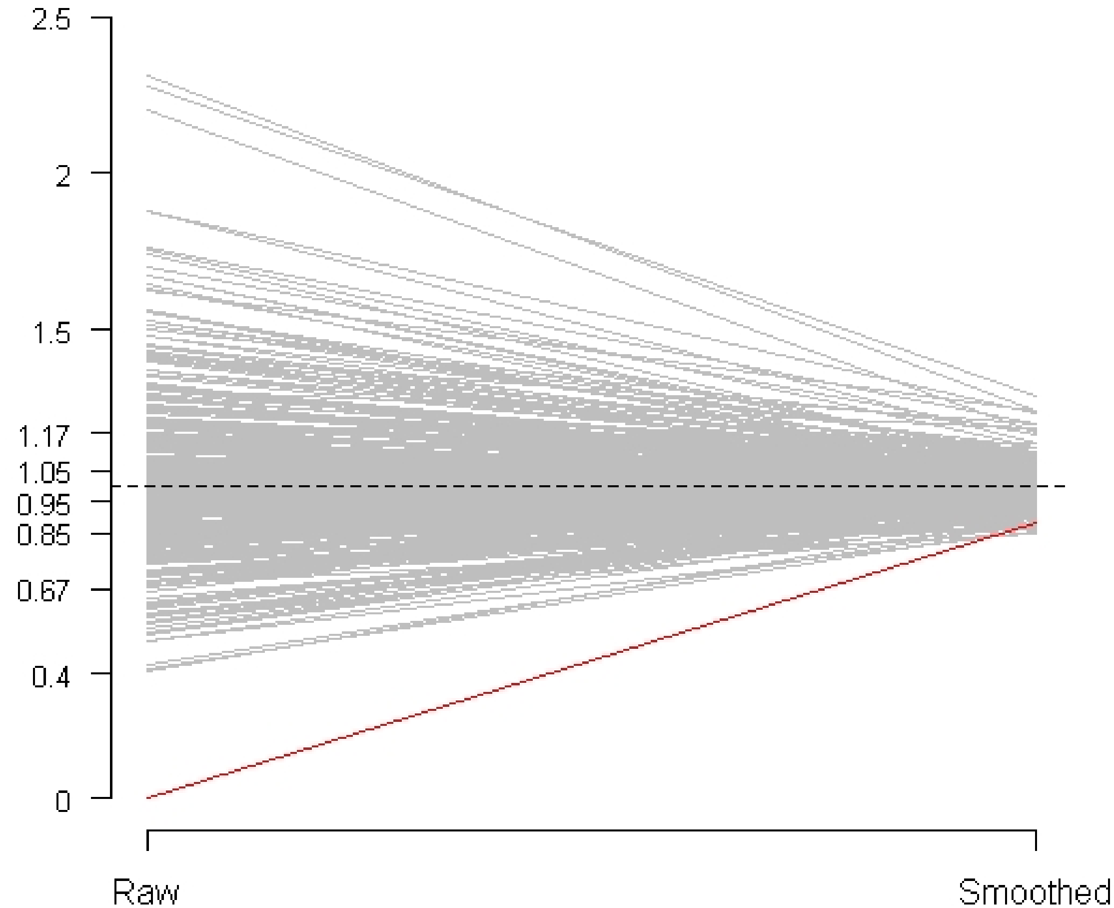
Raw vs Smoothed Risk




Raw
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0


Smoothed
Frequency
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0

Shrinkage

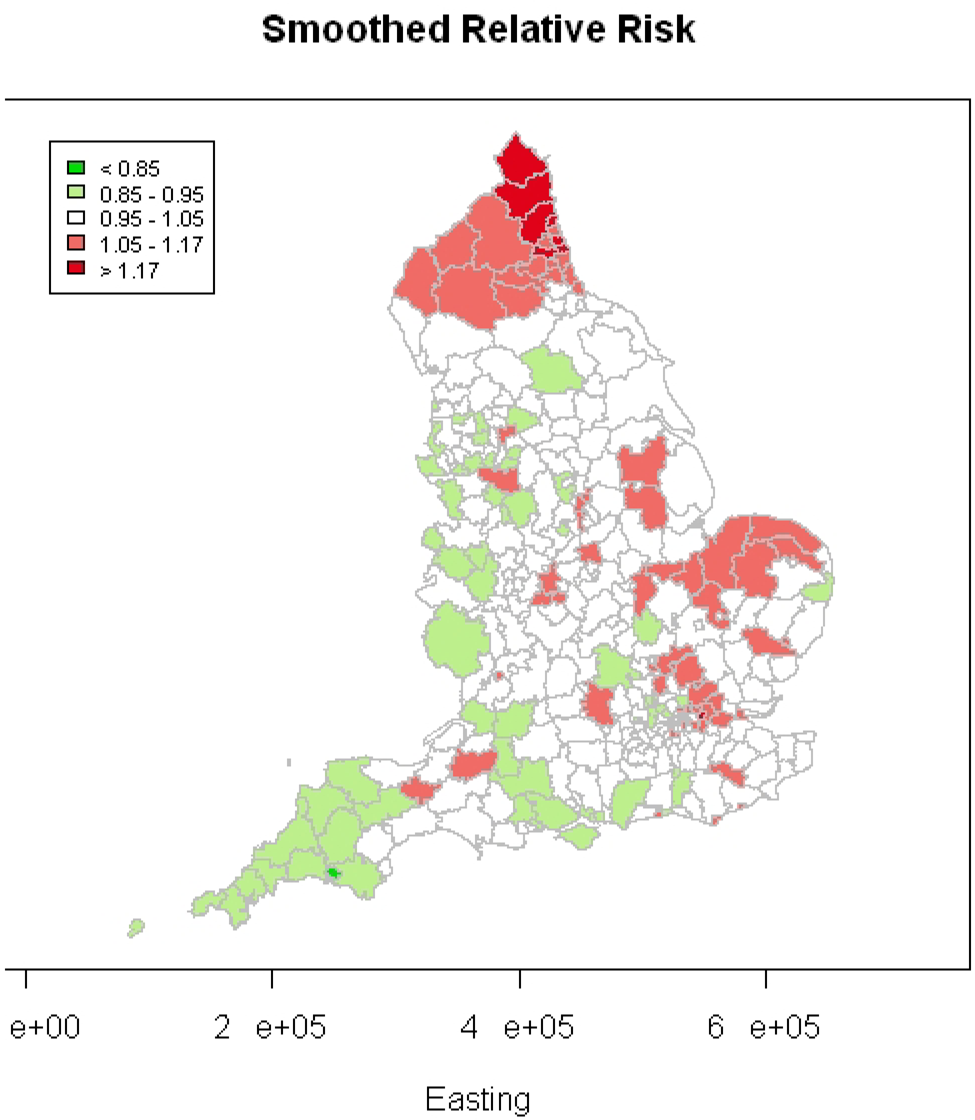

2.5
2
1.5
1.17
1.05
0.95
0.85
0.67
0.4
0
Raw
Smoothed

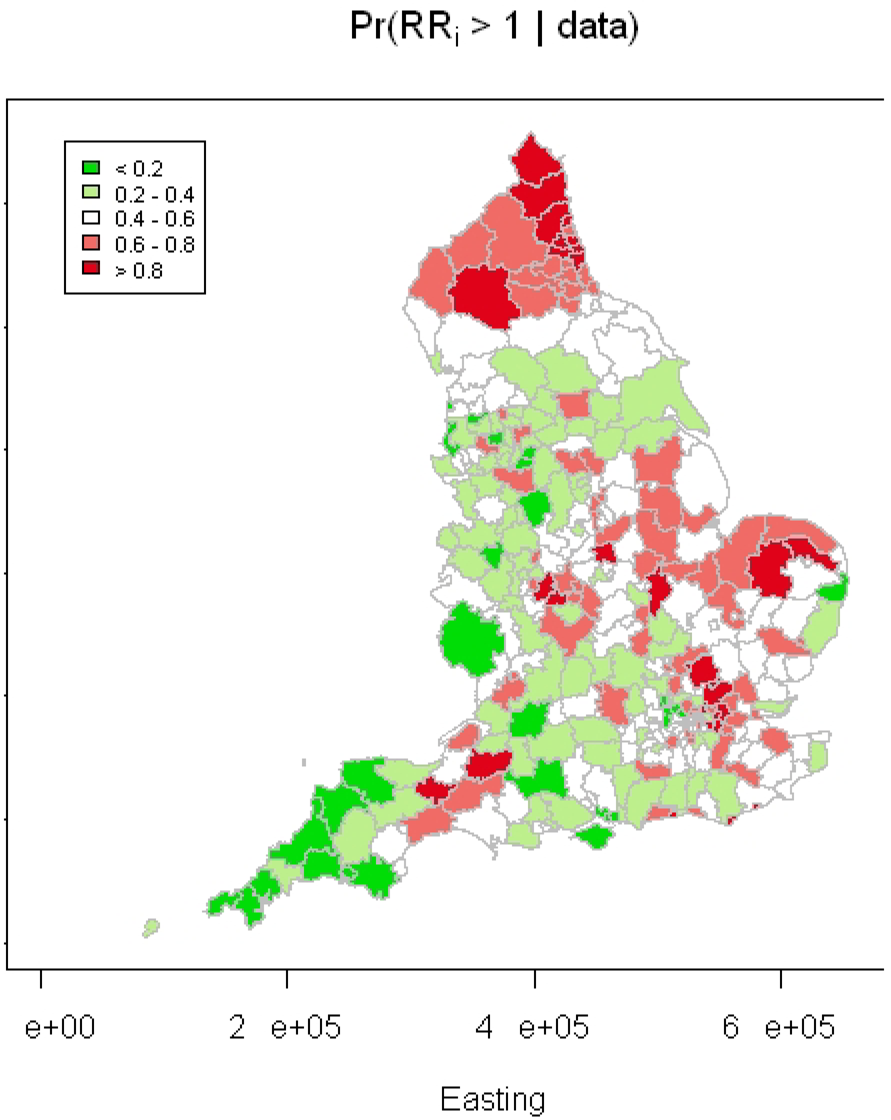
Example...continued

Total Pop	O	E	SIR	CI95% Lower	CI95% Upper
184 201	164	151.50	**1.08**	**0.92**	**1.26**
523	0	2.27	**0.00**	**0.00**	**11.12**

Smoothed RR	Q 2.5%	Q 97.5%
1.05	**0.93**	**1.19**
0.89	**0.57**	**1.29**

Uncertainty?





Discussion

- The Bayesian approach is a very natural framework for random effects.
- The inclusion of random effects allows to accomodate overdispersion.
- The inclusion of spatially structured random effects allows for borrowing information across areas.
- This model has been widely used in the last decade and has proved to be very robust.

Part II:
Bayesian spatial analysis of SES indicators

Joint work with

David Briggs and Daniela Fecht

Background

- ➢ Traditionally, in the UK, SES has been measured by relatively simple indices (Carstairs)
- ➢ More complex measures have been developed recently, designed to capture the multidimensional nature of deprivation.
- ➢ An initial series of country-level indices of multiple deprivation (IMDs) was developed in 2000.
- ➢ These were revised and extended to provide a new set of IMDs in England, Wales and Scotland in 2004.
- ➢ These indices not only use a wider range of variables, arranged in a series of separate domains, but also draw upon a wide variety of different (non-census) data sources.
- ➢ As such they provide the opportunity to recognise and represent different dimensions of deprivation, and perhaps better to characterise some of the complexity inherent in socio-economic associations with health.

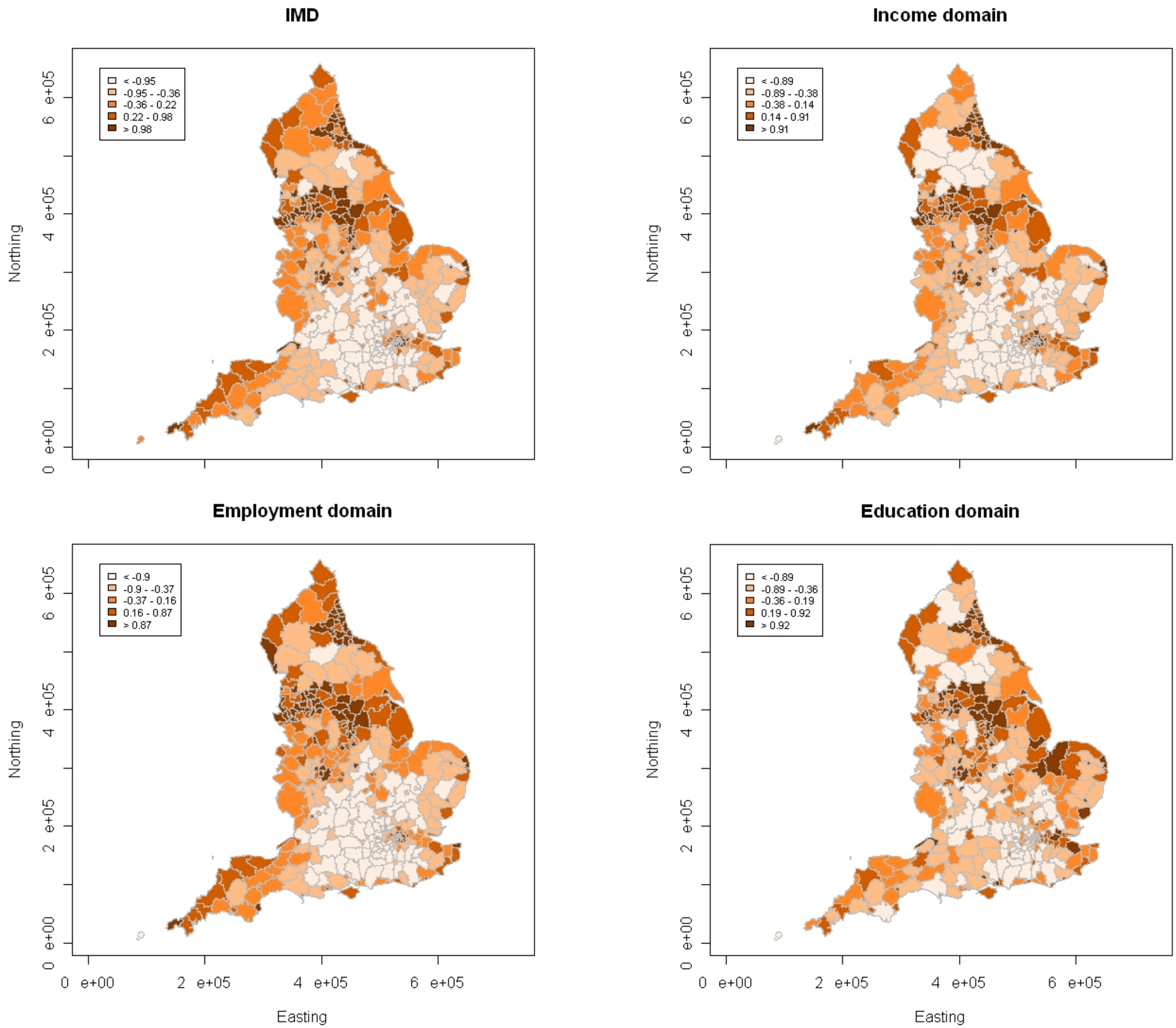
IMD2004 in England

It was derived from 37 variables arranged in seven domains:

1. Income
2. Employment
3. Education
4. Health deprivation and disability
5. Barriers to housing and services
6. Living environment
7. Crime

Primary factors

The English indices of deprivation (2004)
Office of the Deputy Prime Minister


IMD
< -0.95
-0.95 - -0.36
-0.36 - 0.22
0.22 - 0.98
> 0.98
Income domain
< -0.89
-0.89 - -0.38
-0.38 - 0.14
0.14 - 0.91
> 0.91
Employment domain
< -0.9
-0.9 - -0.37
-0.37 - 0.16
0.16 - 0.87
> 0.87
Education domain
< -0.89
-0.89 - -0.36
-0.36 - 0.19
0.19 - 0.92
> 0.92
Northing
0 e+00
2 e+05
4 e+05
6 e+05
Easting

What do they have in common?

Correlations	Income	Employment	Education
IMD	0.97	0.94	0.81
Income		0.91	0.78
Employment			0.79

To analyse their respective geographies (at district level); specifically, to extract their common patterns by splitting them into shared and specific components

A Bayesian spatial model

1st layer: Model

$$\text{domain}_{ij} \sim \text{Normal}(\mu_{ij}, \sigma_i^2) \quad i = 1,2,3;\ j = 1,\ldots,354$$

$$\mu_{ij} = \alpha + \delta_i \,\text{shared}_j + \text{specific}_{ij}$$

2nd layer: Priors

$$\alpha \propto 1$$

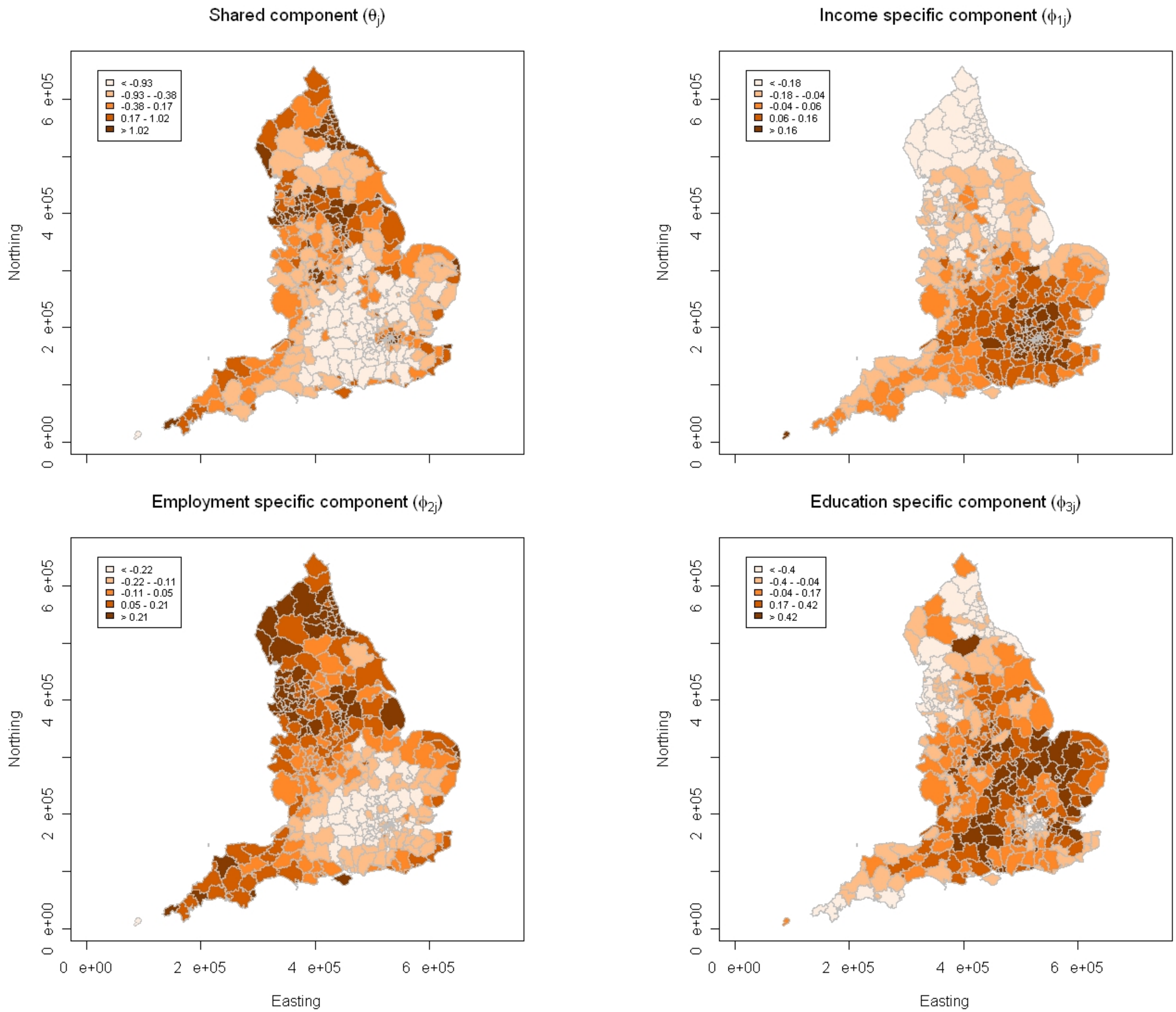
$$\text{shared}_j \sim \text{CARNormal}(\mathbf{W}, \sigma^2_{\text{shared}})$$

$$\text{specific}_{ij} \sim \text{CARNormal}(\mathbf{W}, \sigma^2_{\text{specific}_i})$$

$$\log \delta_i \sim \text{Normal}(0, 5)$$

3rd layer: Hyperpriors

$$\sigma^2\text{'s} \sim \text{IGamma}(0.5, 0.0005)$$



Shared component (θ_j)
< -0.93
-0.93 - -0.38
-0.38 - 0.17
0.17 - 1.02
> 1.02
Income specific component (ϕ_{1j})
< -0.18
-0.18 - -0.04
-0.04 - 0.06
0.06 - 0.16
> 0.16
Employment specific component (ϕ_{2j})
< -0.22
-0.22 - -0.11
-0.11 - 0.05
0.05 - 0.21
> 0.21
Education specific component (ϕ_{3j})
< -0.4
-0.4 - -0.04
-0.04 - 0.17
0.17 - 0.42
> 0.42
Northing
0 e+00
2 e+05
4 e+05
6 e+05
Easting

Variance explained and correlations

	Shared comp.	Specific Income	Specific Employ	Specific Educat	Residual Income	Residual Enploy	Residual Educat
Variance	2.588	0.111	0.092	0.582	0.003	0.001	0.055
% Var	75.42	3.22	2.67	16.96	0.10	0.02	1.61

Correlations	Specific Income	Specific Employment	Specific Education
Shared	-0.30	0.57	-0.38
Specific Income		-0.85	-0.06
Specific Employment			-0.36

Discussion

• We split the three scores into four components, one shared by all of them and one specific to each of them.

• That common geography explained 75.42% of the total variability in the scores, and turned out to be very similar to the IMD itself.

• The specific patterns of the three domains have less relevance in terms of the proportion of variance (all together explain 22.85% of the variance). They reflect very different patterns, of which income and employment are rather antagonistic.

• This decomposition can potentially be very useful to assess inequities in health in relation to SES, as we can consider the shared component or a specific one, or a combination of them.

• The results we obtained in this study merit further investigation and analyses.

Thanks for you attention!