Spatial Statistics for Epidemiologists

Part 2: Methods for Areal Data

Leonhard Held and Birgit Schrödle

27.-29. October 2010

The Besag *et al.* model

$\rightarrow$ Incorporate spatial structure in Bayesian hierarchical model:

$$\boxed{\text{Assume } y_i \sim \text{Poisson}(e_i\theta_i) \text{ with } \log(\theta_i) = \mu + u_i + v_i}$$

μ: Overall level

v_i: Unstructured component

u_i: Spatially structured component

- Gaussian Markov random field model for u_i:
  $u_i | u_{j\neq i} \sim N(\bar{u}_i, \sigma_u^2/m_i)$
  $\bar{u}_i$: Mean of the neighbours
  m_i: Number of neighbours
- Independent random effects $v_i \sim N(0, \sigma_v^2)$
- Unknown hyperparameters σ_u^2 and $\sigma_v^2 \rightarrow$ hyperpriors

Computational issues

- Statistical inference typically performed using Markov chain Monte Carlo (MCMC), this can be quite challenging and is non-trivial!
- Based on MCMC samples from the posterior we compute:
  - Point estimates of relative risk (posterior mean)
  - "Bayesian p-values": Posterior probabilities that the relative risk θ_i in area i is above the overall risk $\sum y_i / \sum e_i$ (or any other reference value)
- A more efficient approach is based on integrated nested Laplace approximations (INLA)

Hyperpriors

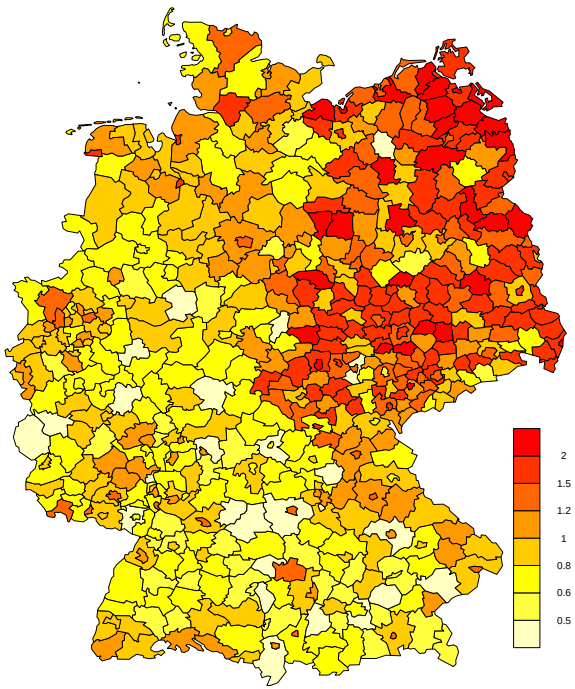
- Prior for variances σ_u^2 and σ_v^2 conveniently chosen from inverse gamma(a, b) distribution
- However no "non-informative" choice possible
- "Weekly informative" choices have been suggested such as
  (a) inverse gamma$(1, 0.005)$
  (b) inverse gamma$(0.5, 0.005)$
  (c) inverse gamma$(0.25, 0.0005)$ for σ_u^2 and inverse gamma$(0.25, 0.00025)$ for σ_v^2
- Prior (c) accounts for the fact that the u_is are independent of each other, but the v_is are conditioned on neighbouring areas. Therefore, (c) was used in all applications
- Prior sensitivity should be studied

SMRs and Relative risk estimates

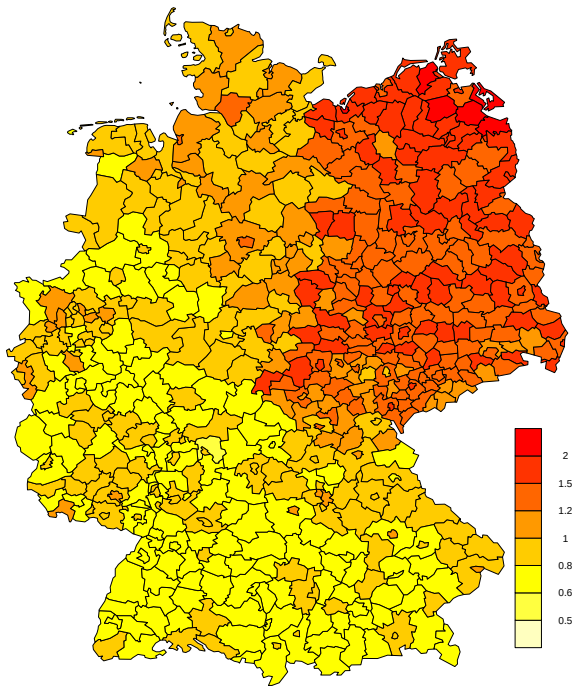
- Cancer type: Cervix
- Using INLA:

```
> bymcervix <- inla(formula, family = "poisson", data = cervixdata,
+     E = E, control.inla = list(h = 0.01), control.predictor = list(compute = TRUE,
+         cdf = c(0)))
```

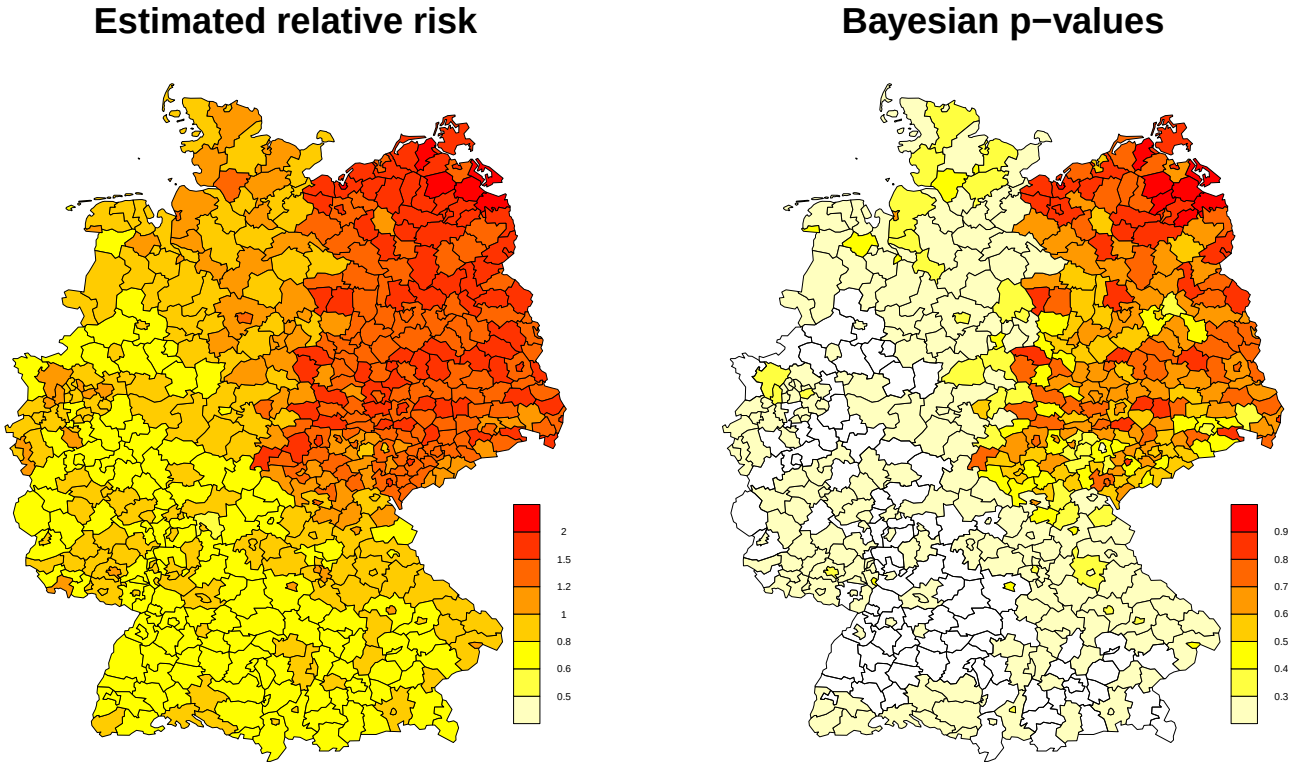
SMRs



Estimated relative risk



Relative risk estimates and Bayesian p-values



Discussion

The BYM model for estimating disease risk is a further improvement because

1. the spatial structure is incorporated
2. structured and unstructured heterogeneity can be distinguished

Sensitivity to the choice of hyperpriors should be studied

Interpretation of smoothed mortality maps is delicate/difficult:

1. Can only be suggestive due to observational study type
2. Can rarely be seen as an end in itself

Ecological regression

$\rightarrow$ Inclusion of area-level covariates $\mathbf{x}_i$, e.g. average information on exposure in area i:

$$\boxed{\text{Assume } y_i \sim \text{Poisson}(e_i\theta_i) \text{ with } \log(\theta_i) = \mu + u_i + v_i + \mathbf{x}_i'\boldsymbol{\beta}}$$

- Idea: reduce ("explain") (spatial) variation through inclusion of relevant covariates

$\rightarrow$ Residual variation of adjusted relative risk $\exp(u_i + v_i)$ hopefully reduced

- Statistical inference by MCMC or INLA

Ecological bias and ecological fallacy

- Ecological studies can only use average information on exposure in an area.
- Can only generate hypotheses about potential risk factors, due to the problem of ecological bias, or ecological fallacy: the mistaken assumption that a statistical association observed on group level is equal to the individual level.
- However, ecological studies are typically much cheaper and can have advantages.
- It is quite common to replace potential exposures variables by surrogates such as deprivation or urbanization scores.

Example: German cancer mortality data, 1986-1990

- Cancer type: Oesophagus
- Covariates:

  x_1: Log lung cancer relative risk (from CK model) to adjust for smoking

  x_2: East-West indicator to account for different cancer registries, different screening methods etc.
- Following slides show relative risk estimates without adjustment and with
  1. adjustment for smoking
  2. adjustment for smoking and East-West indicator

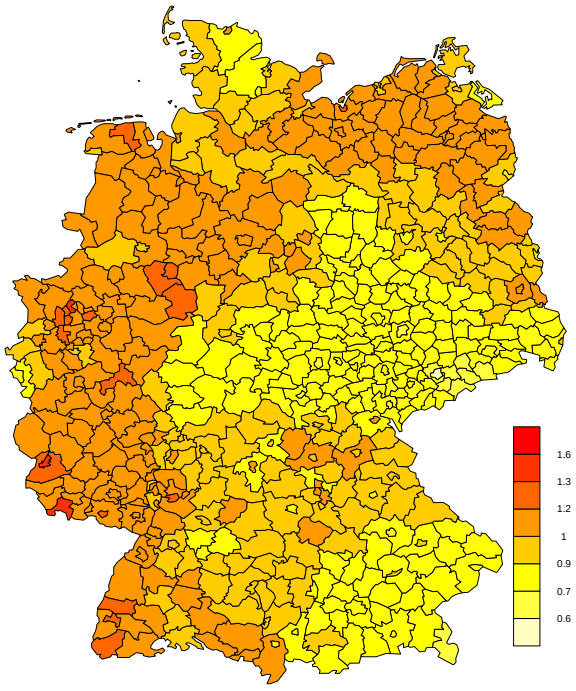
  and fixed covariate effects

Relative risk - adjustment for smoking

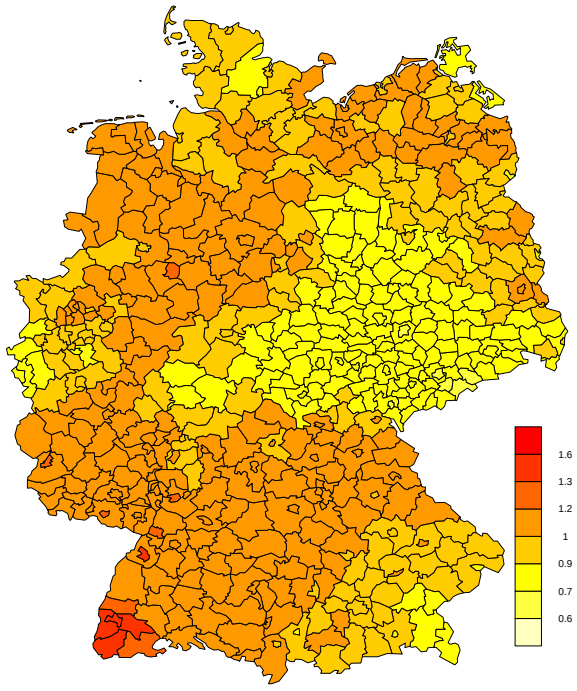
```
> bymsmoke$summary.fixed[, 1:5]
```

	mean	sd	0.025quant	0.5quant	0.975quant
(Intercept)	-0.04099483	0.01229404	-0.06527379	-0.04094766	-0.01702774
logcklung	0.54825769	0.08762168	0.37627863	0.54818053	0.72045330

Without adjustment



With adjustment

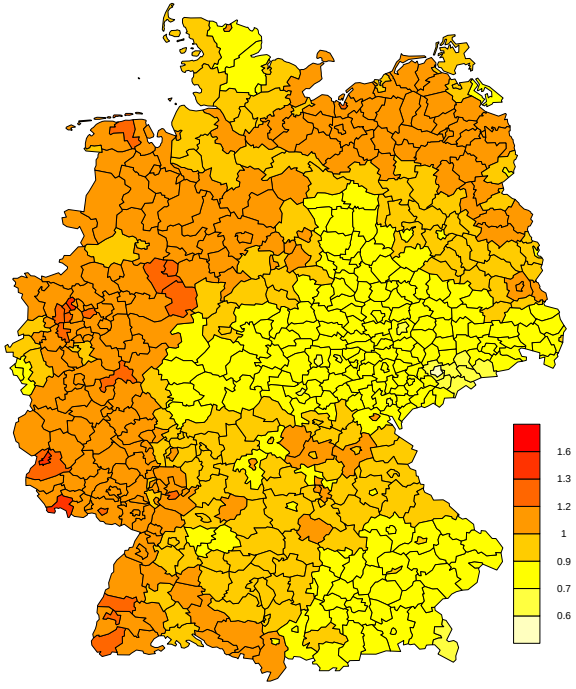


Adjustment for smoking and East-West indicator

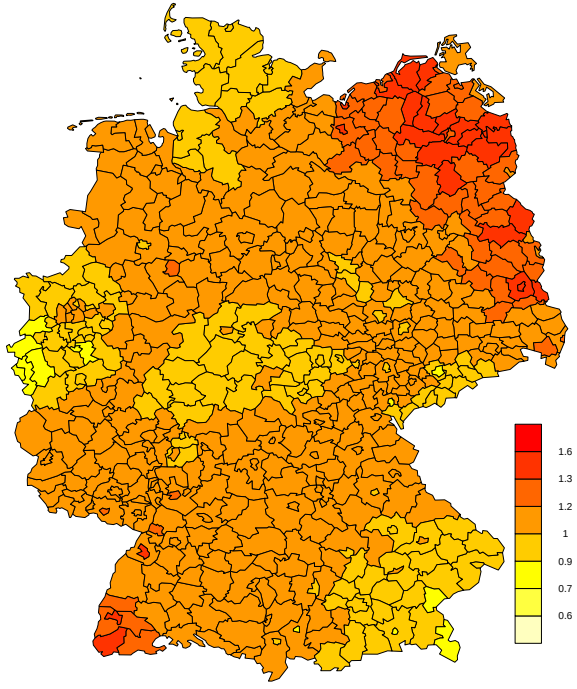
```
> bymsmokeeast$summary.fixed[, 1:5]
```

	mean	sd	0.025quant	0.5quant	0.975quant
(Intercept)	0.06518608	0.02391720	0.01846797	0.06505168	0.1126393
logcklung	0.57996878	0.08384448	0.41551127	0.57984754	0.7449167
eastwest	-0.29175874	0.05831973	-0.40796262	-0.29125689	-0.1784920

Without adjustment



With adjustment



Confounding by location

Ecological regression can also be performed by simple GLM without spatially structured random effects:

$$\boxed{\text{Assume } y_i \sim \text{Poisson}(e_i\theta_i) \text{ with } \log(\theta_i) = \mu + x_i'\beta}$$

If covariate is spatially structured then

- covariate effect from simple GLM will be diluted in BYM ecological regression

Example: Cervical cancer with East-West indicator as covariate

- Model without spatial clustering effect:

```
> nostructured <- glm(cervix[, 1] ~ eastwest, family = poisson,
+     offset = log(cervix[, 2]))
> nostructured$coefficients

(Intercept)    eastwest
 -0.1354765   0.5358304
```

- Model with spatial clustering effect:

```
> withstructured <- inla(cervixeast, family = "poisson", data = cervixdata,
+     E = E, control.inla = list(h = 0.01), control.compute = list(dic = TRUE))

> withstructured$summary.fixed[, 1:5]

                  mean         sd 0.025quant   0.5quant 0.975quant
(Intercept) -0.1585814 0.02281853 -0.2026954 -0.1588642 -0.1128625
eastwest     0.4870337 0.04988466  0.3852342  0.4882812  0.5818754
```

Discussion

Ecological regression can be helpful in order to

- reduce/explain overdispersion and spatial variation

However, interpretation is often difficult because

- covariate information is on area level rather than individual level
  $\rightarrow$ ecological fallacy
- covariate information is often very crude and only vaguely related to potential exposure
  $\rightarrow$ adjustments for measurement error have been suggested