Analysing geoadditive regression data: a mixed model approach

Thomas Kneib
Institut für Statistik, Ludwig-Maximilians-Universität München

Joint work with
Ludwig Fahrmeir & Stefan Lang

I B S

25.11.2005

LMU

Spatio-temporal regression data

- Regression in a general sense:
  - Generalised linear models,
  - Multivariate (categorical) generalised linear models,
  - Regression models for survival times (Cox-type models, AFT models).
- Common structure: Model a quantity of interest in terms of categorical and continuous covariates, e.g.

$$\mathbb{E}(y|u) = h(u'\gamma) \qquad \text{(GLM)}$$

  or

$$\lambda(t|u) = \lambda_0(t)\exp(u'\gamma) \qquad \text{(Cox model)}$$

- Spatio-temporal data: Temporal and spatial information as additional covariates.

- Spatio-temporal regression models should allow
  - to account for spatial and temporal correlations,
  - for time- and space-varying effects,
  - for non-linear effects of continuous covariates,
  - for flexible interactions,
  - to account for unobserved heterogeneity.

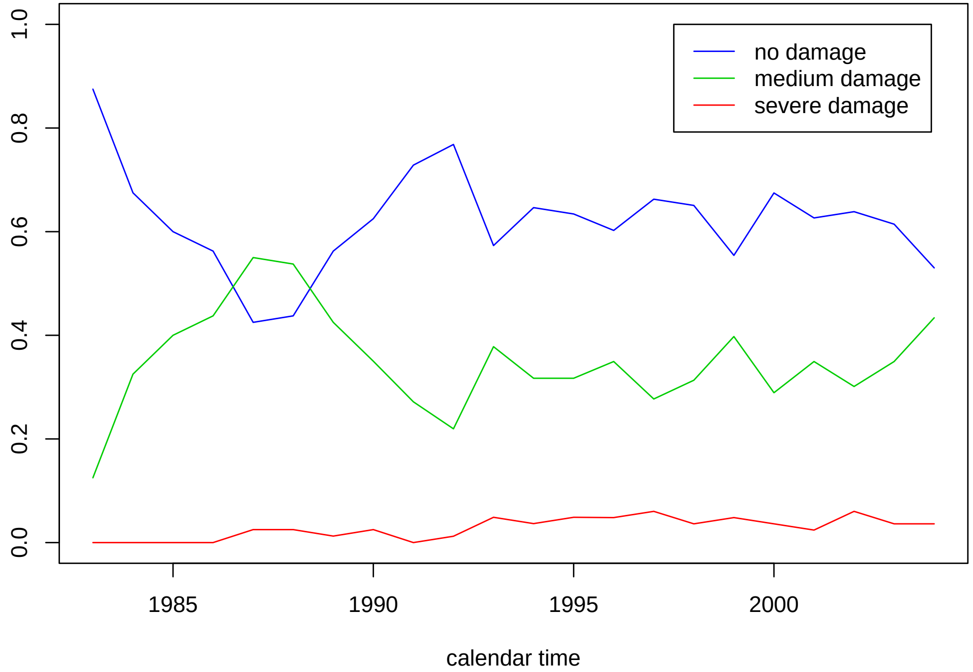
Example I: Forest health data

- Yearly forest health inventories carried out from 1983 to 2004.
- 83 beeches within a 15 km times 10 km area.
- Response: defoliation degree of beech i in year t, measured in three ordered categories:

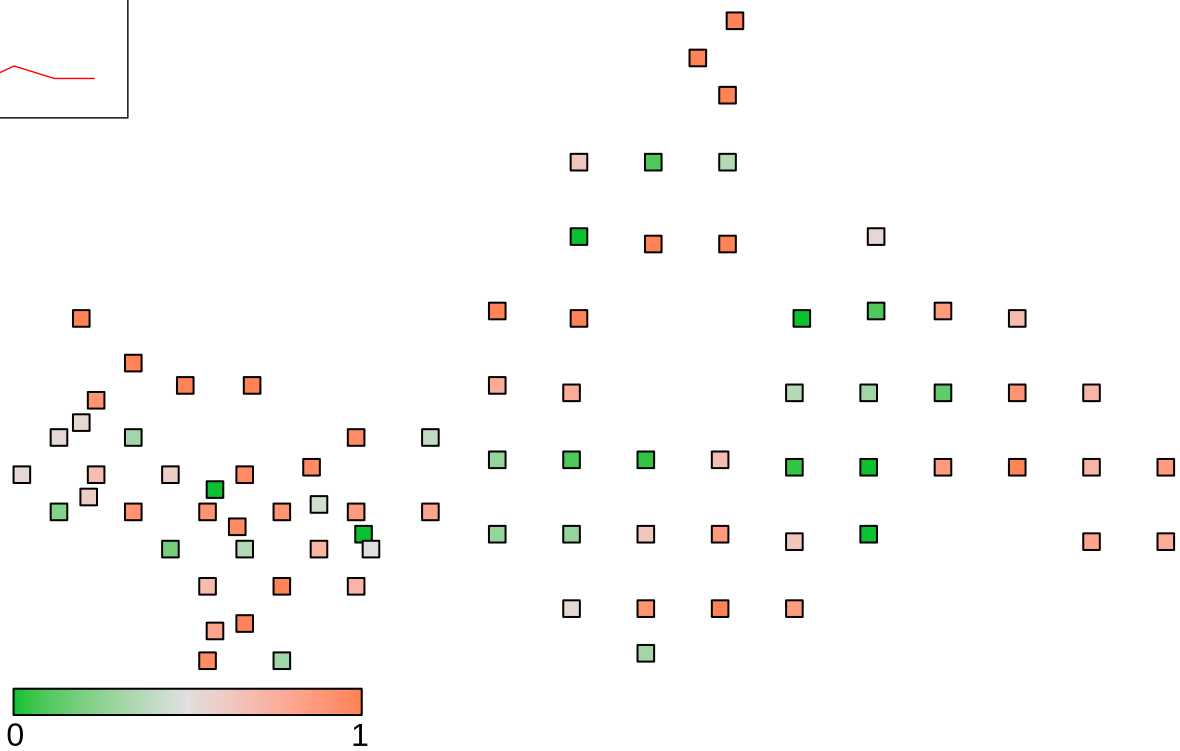
  $y_{it} = 1$ no defoliation,
  $y_{it} = 2$ defoliation 25% or less,
  $y_{it} = 3$ defoliation above 25%.

- Covariates:

  t calendar time,
  s_i site of the beech,
  a_{it} age of the tree in years,
  u_{it} further (mostly categorical) covariates.



Empirical time trends.



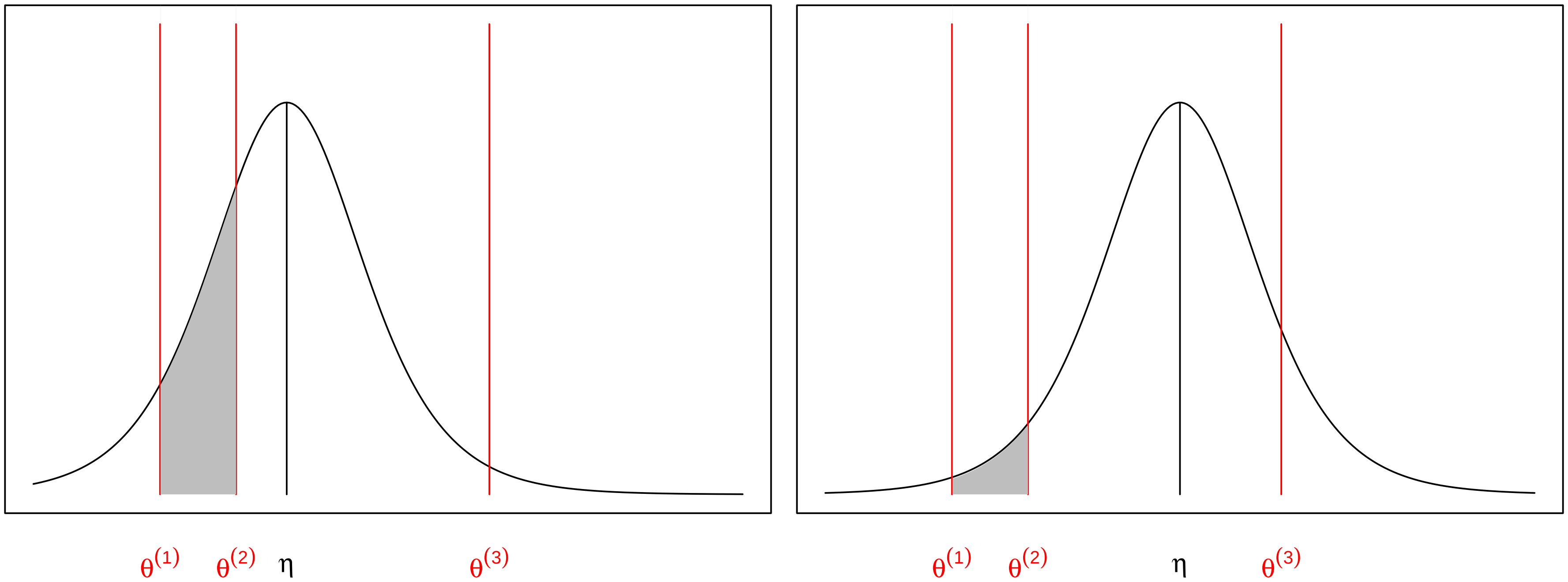
Empirical spatial effect.

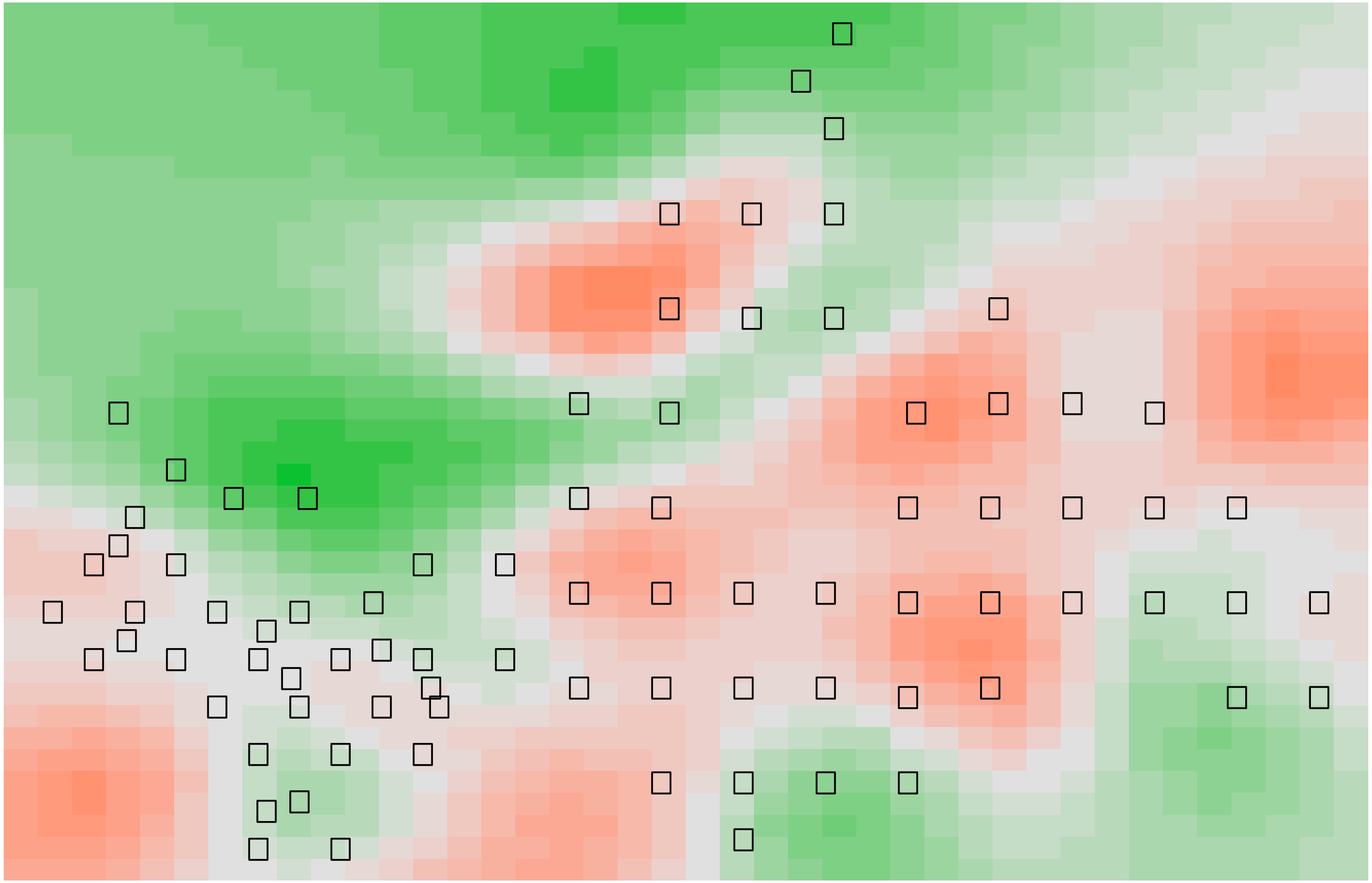
- Cumulative probit model:

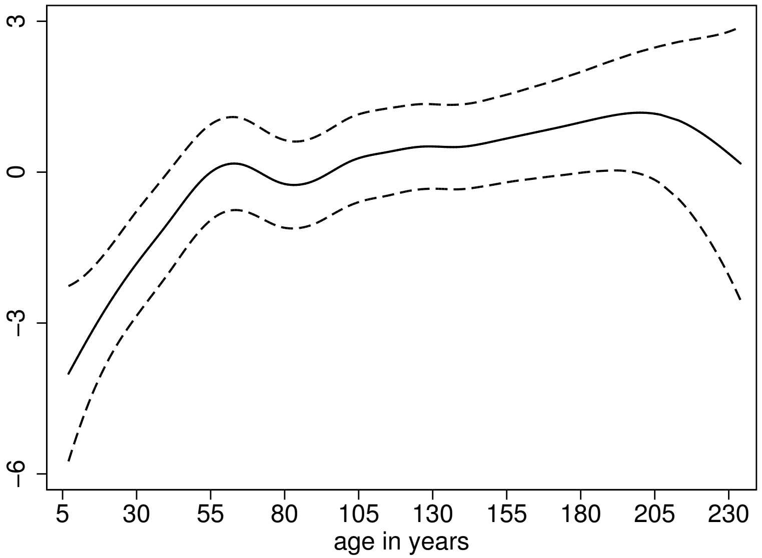
$$P(y_{it} \leq r) = \Phi\left(\theta^{(r)} - \eta_{it}\right)$$

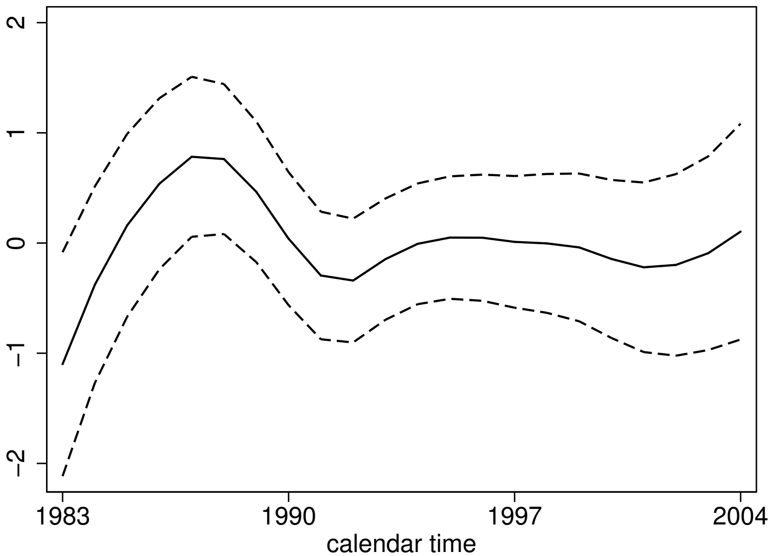
with standard normal cdf Φ, thresholds $-\infty = \theta^{(0)} < \theta^{(1)} < \theta^{(2)} < \theta^{(3)} = \infty$ and

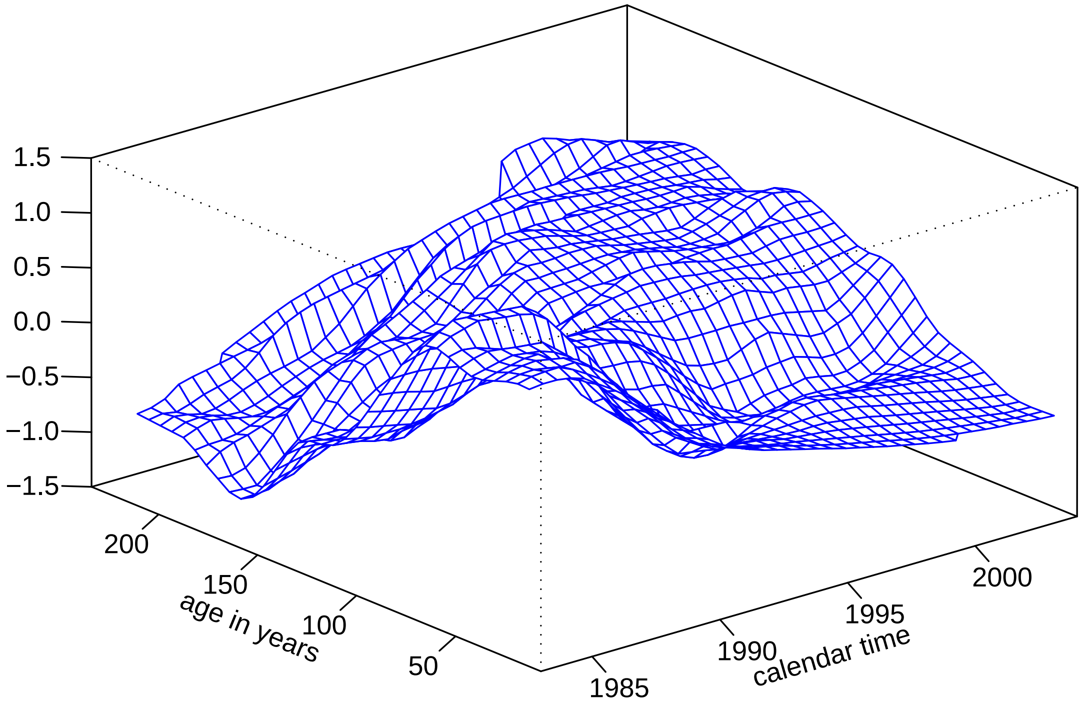
$$\eta_{it} = f_1(t) + f_2(age_{it}) + f_3(t, age_{it}) + f_{spat}(s_i) + u_{it}'\gamma$$






3
0
−3
−6
5 30 55 80 105 130 155 180 205 230
age in years


2
1
0
−1
−2
1983 1990 1997 2004
calendar time

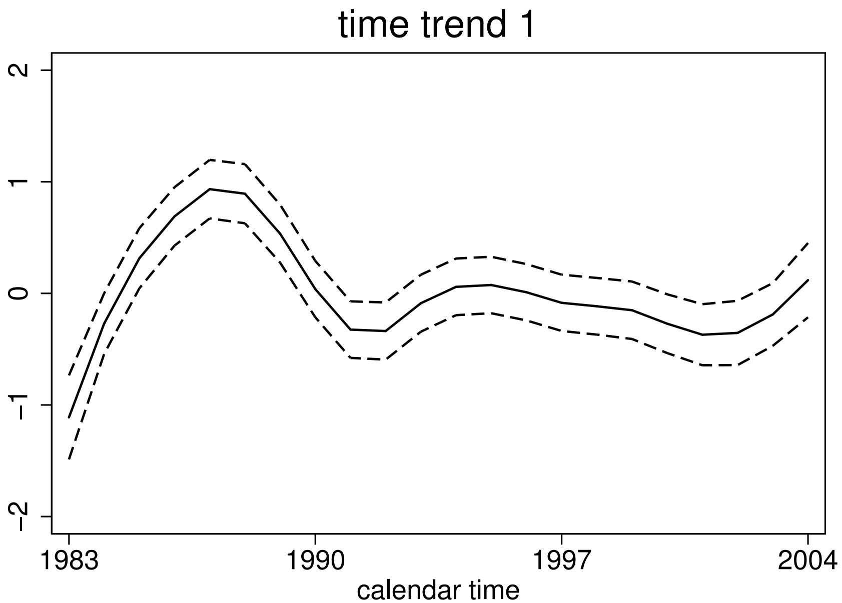

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
200
150
100
50
age in years
1985
1990
1995
2000
calendar time

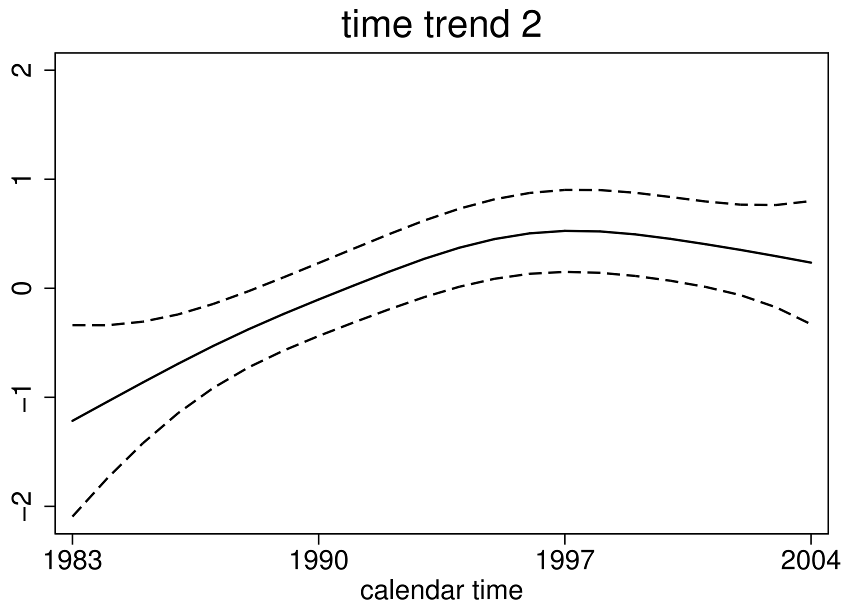
- Category-specific trends:

$$P(y_{it} \leq r) = \Phi\left[\theta^{(r)} - f_1^{(r)}(t) - f_2(age_{it}) - f_{spat}(s_i) - u_{it}'\gamma\right]$$

- More complicated constraints:

$$-\infty < \theta^{(1)} - f_1^{(1)}(t) < \theta^{(2)} - f_1^{(2)}(t) < \infty \qquad \text{for all } t.$$

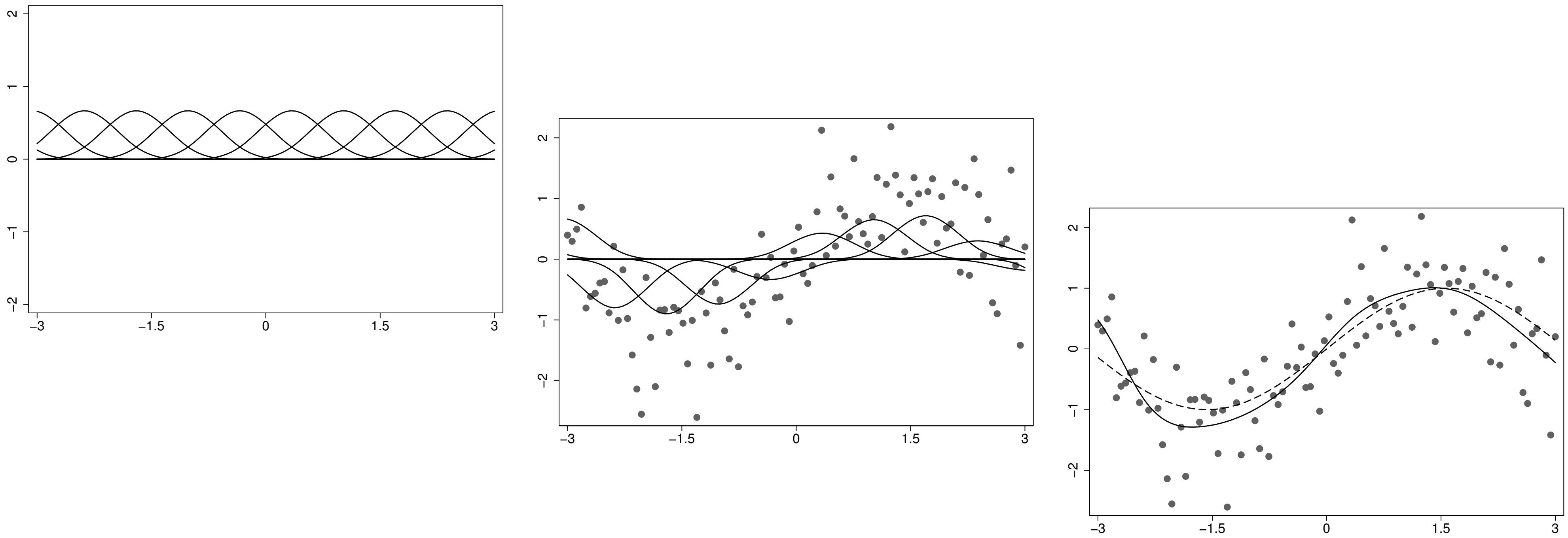




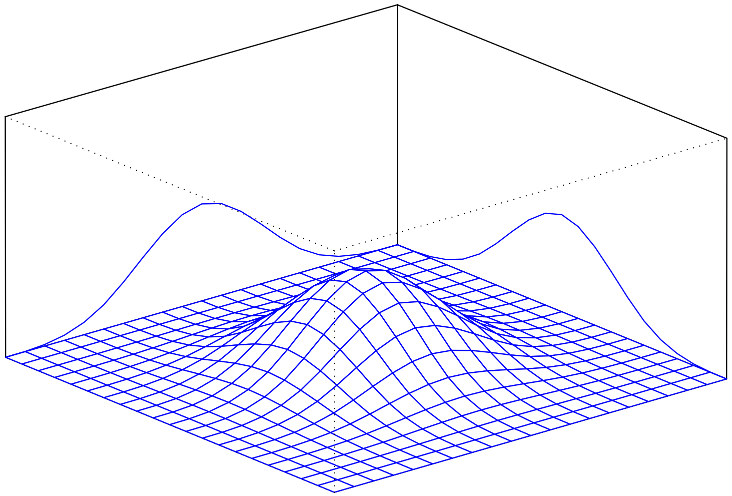
Structured additive regression

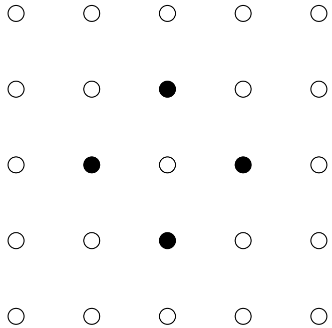
- General Idea: Replace usual parametric predictor with a flexible semiparametric predictor containing
  - Nonparametric effects of time scales and continuous covariates,
  - Spatial effects,
  - Interaction surfaces,
  - Varying coefficient terms (continuous and spatial effect modifiers),
  - Random intercepts and random slopes.
- All effects can be cast into one general framework.

- Penalised splines.
  - Approximate $f(x)$ by a weighted sum of B-spline basis functions.
  - Employ a large number of basis functions to enable flexibility.
  - Penalise differences between parameters of adjacent basis functions to ensure smoothness.



- Bivariate penalised splines.



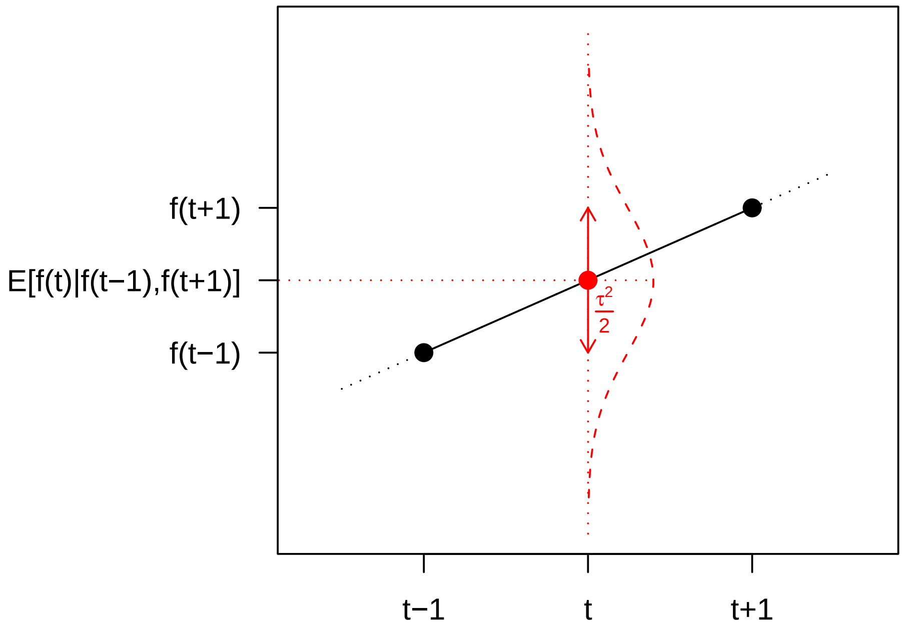


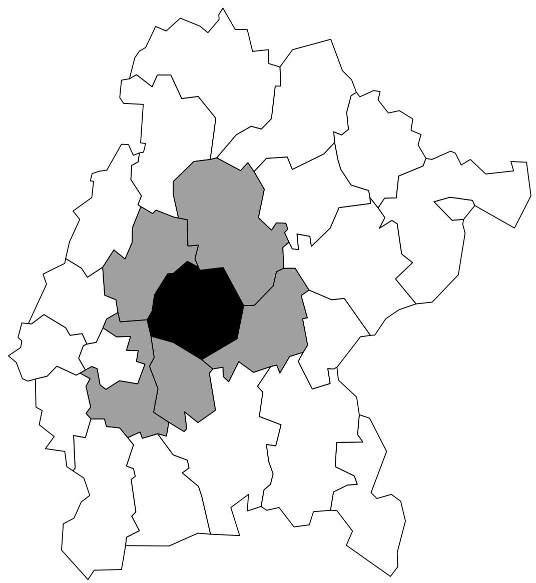
- Varying coefficient models.
  - Effect of covariate x varies smoothly over the domain of a second covariate z:

$$f(x, z) = x \cdot g(z)$$

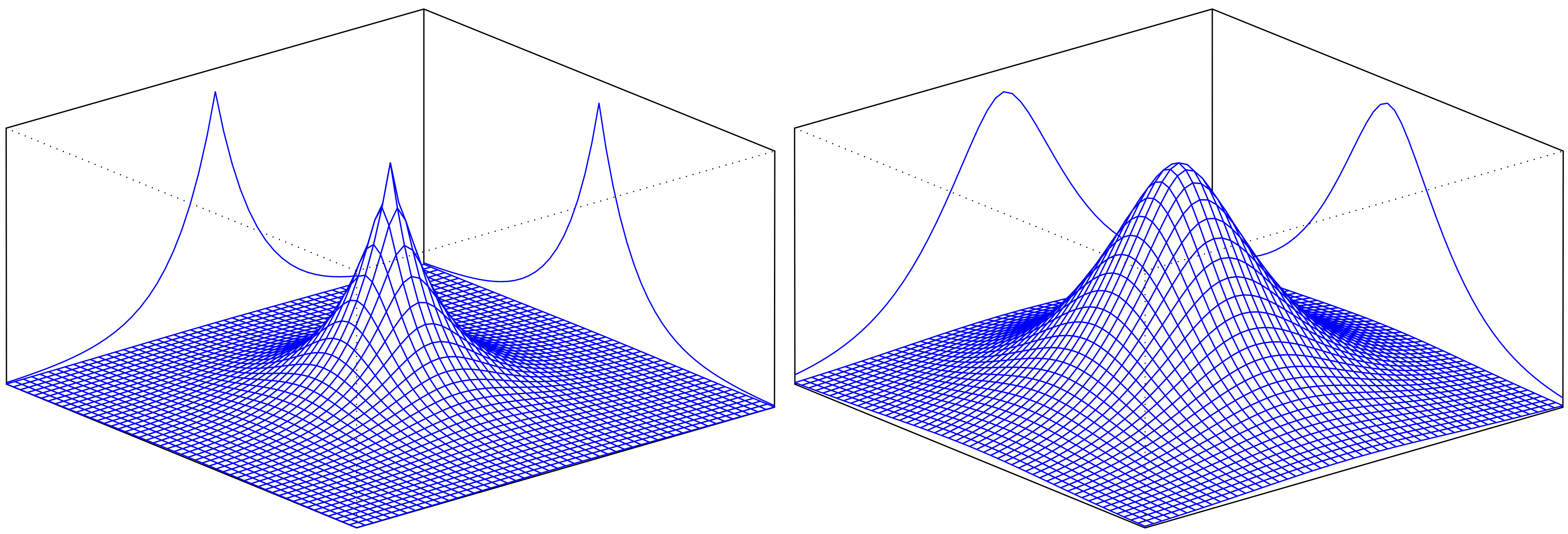
  - Spatial effect modifier $\Rightarrow$ Geographically weighted regression.

- Spatial effect for regional data: Markov random fields.
  - Bivariate extension of a first order random walk on the real line.
  - Define appropriate neighbourhoods for the regions.
  - Assume that the expected value of $f_{spat}(s)$ is the average of the function evaluations of adjacent sites.





- Spatial effect for point-referenced data: Stationary Gaussian random fields.
  - Well-known as Kriging in the geostatistics literature.
  - Spatial effect follows a zero mean stationary Gaussian stochastic process.
  - Correlation of two arbitrary sites is defined by an intrinsic correlation function.
  - Can be interpreted as a basis function approach with radial basis functions.



Mixed model based inference

- Each term in the predictor is associated with a vector of regression coefficients with multivariate Gaussian prior / random effects distribution:

$$p(\xi_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\xi_j' K_j \xi_j\right)$$

- K_j is a penalty matrix, τ_j^2 a smoothing parameter.

- In most cases K_j is rank-deficient.

$\Rightarrow$ Reparametrise the model to obtain a mixed model with proper distributions.

- Decompose

$$\xi_j = X_j \beta_j + Z_j b_j,$$

where

$$p(\beta_j) \propto const \qquad \text{and} \qquad b_j \sim N(0, \tau_j^2 I).$$

$\Rightarrow$ β_j is a fixed effect and b_j is an i.i.d. random effect.

- This yields the variance components model

$$\eta = x'\beta + z'b,$$

where in turn

$$p(\beta) \propto const \qquad \text{and} \qquad b \sim N(0, Q).$$

- Obtain empirical Bayes estimates / penalised likelihood estimates via iterating
  - Penalised maximum likelihood for the regression coefficients β and b.
  - Restricted Maximum / Marginal likelihood for the variance parameters in Q:

$$L(Q) = \int L(\beta, b, Q)p(b)d\beta db \to \max_{Q} .$$

Software

- Implemented in the software package BayesX.

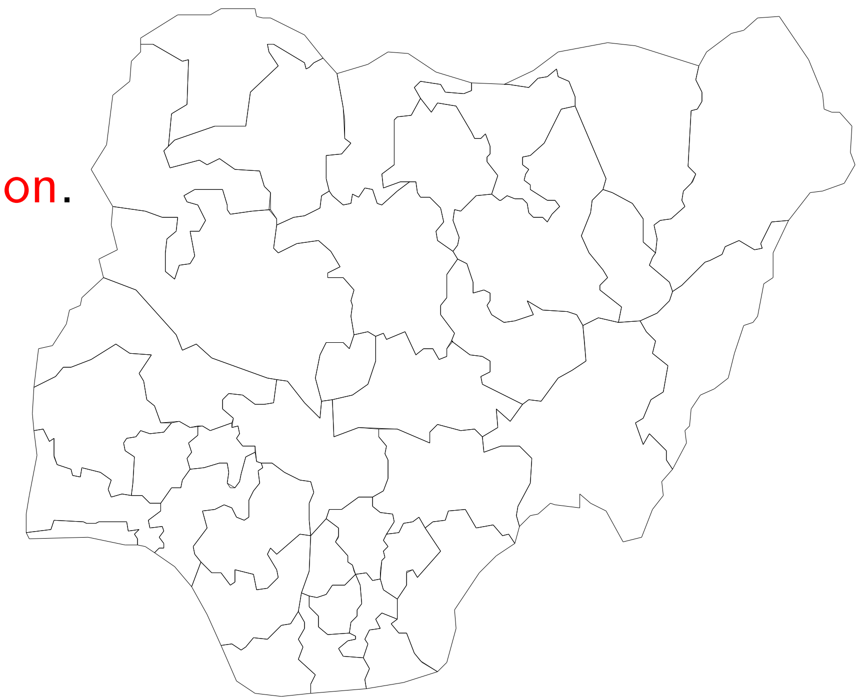
- Available from

`http://www.stat.uni-muenchen.de/~bayesx`

Childhood mortality in Nigeria

- Data from the 2003 Demographic and Health Survey (DHS) in Nigeria.
- Retrospective questionnaire on the health status of women in reproductive age and their children.
- Survival time of $n = 5323$ children.
- Numerous covariates including spatial information.
- Analysis based on the Cox model:

$$\lambda(t; u) = \lambda_0(t) \exp(u'\gamma).$$

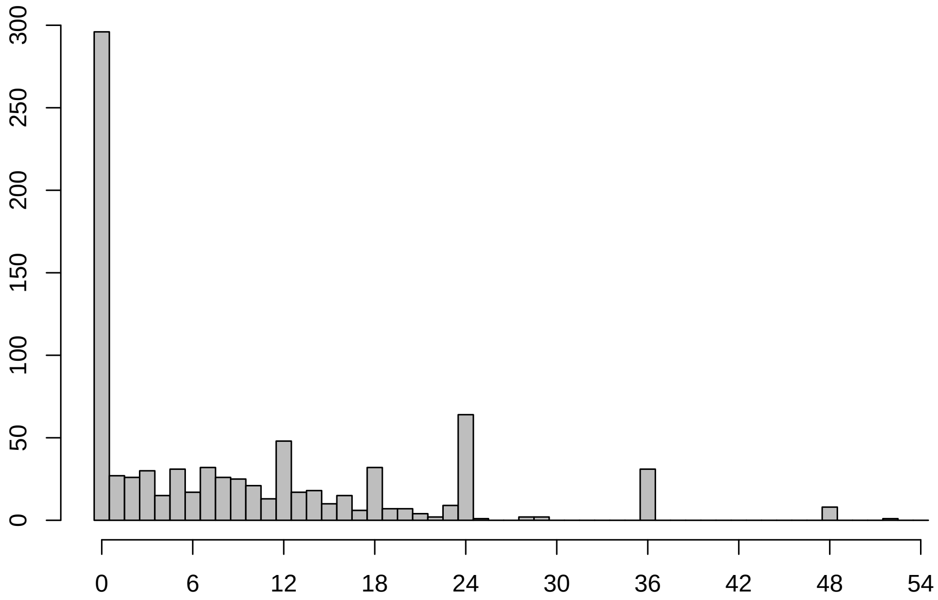


- Limitations of the classical Cox model:
  - Restricted to right censored observations.
  - Post-estimation of the baseline hazard.
  - Proportional hazards assumption.
  - Parametric form of the predictor.
  - No spatial correlations.

$\Rightarrow$ Geoadditive hazard regression.

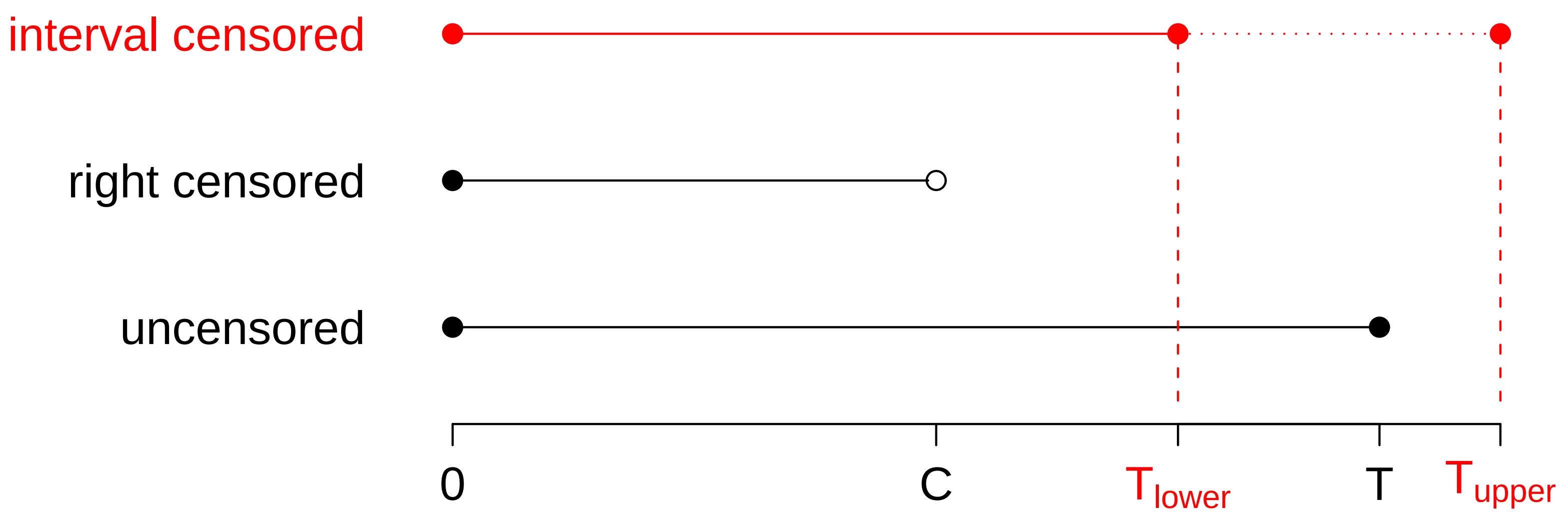
Interval censored survival times

- In theory, survival times should be available in days.
- Retrospective questionnaire ⇒ most uncensored survival times are rounded (Heaping).



- In contrast: censoring times are given in days.

⇒ Treat survival times as interval censored.


interval censored
right censored
uncensored
0
C
T_{lower}
T
T_{upper}

- Likelihood contributions:

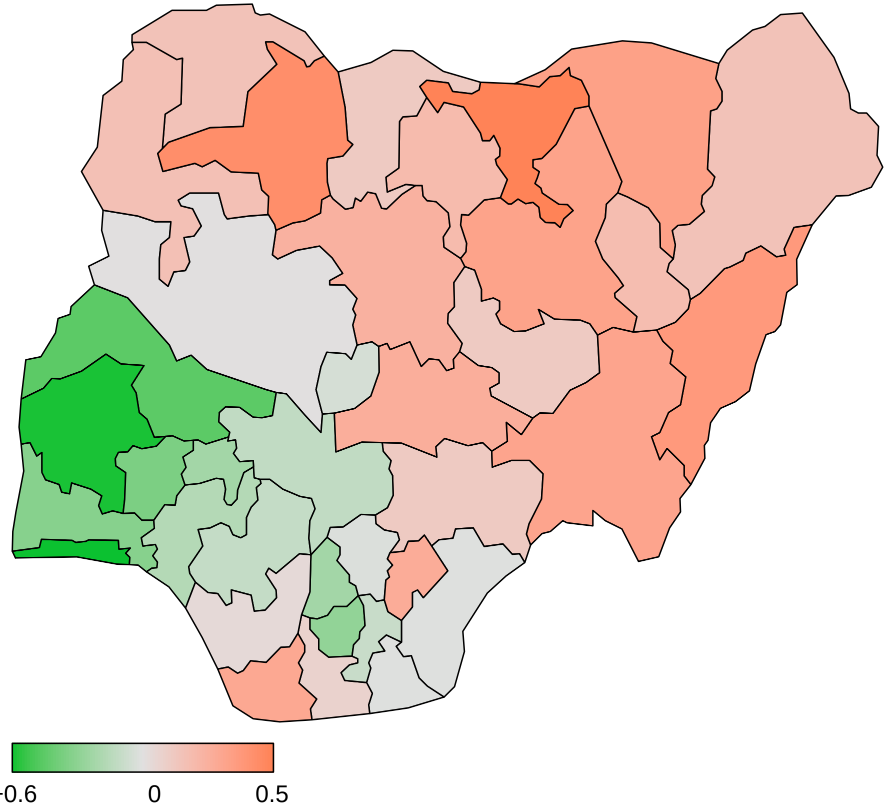
$$
\begin{aligned}
P(T > C) &= S(C) \\
&= \exp\left[-\int_0^C \lambda(t)dt\right].
\end{aligned}
$$

$$
\begin{aligned}
P(T \in [T_{lower}, T_{upper}]) &= S(T_{lower}) - S(T_{upper}) \\
&= \exp\left[-\int_0^{T_{lower}} \lambda(t)dt\right] - \exp\left[-\int_0^{T_{upper}} \lambda(t)dt\right].
\end{aligned}
$$

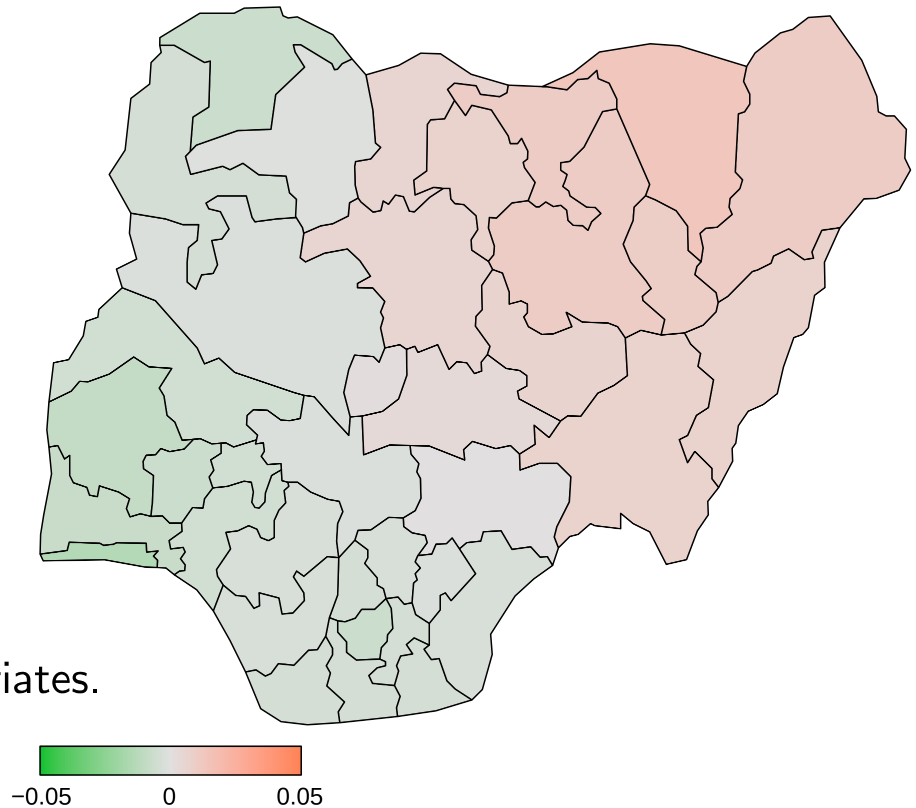
- Derivatives of the log-likelihood become much more complicated for interval censored survival times.

- Numerical integration techniques have to be used in both cases.

- Piecewise constant time-varying covariates and left truncation can easily be included.



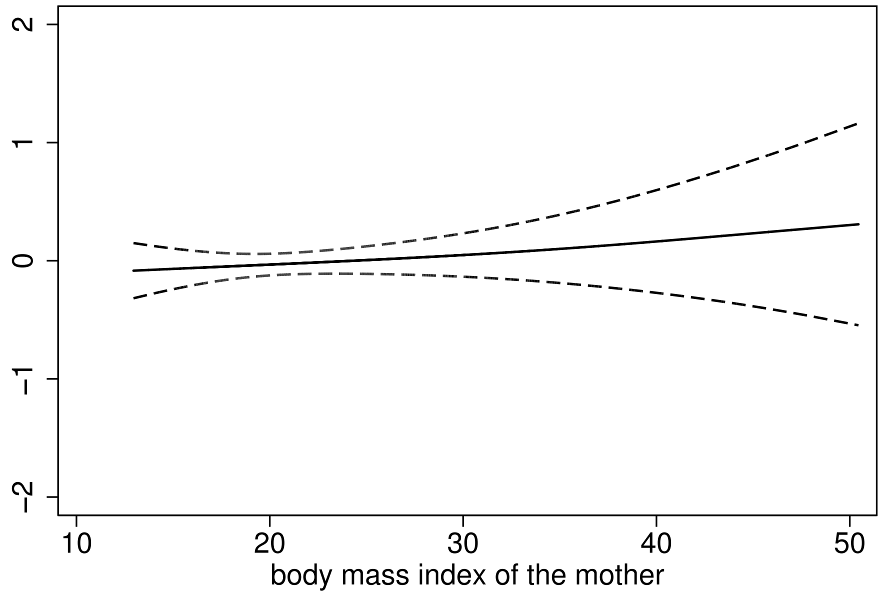
Spatial effect without covariates.

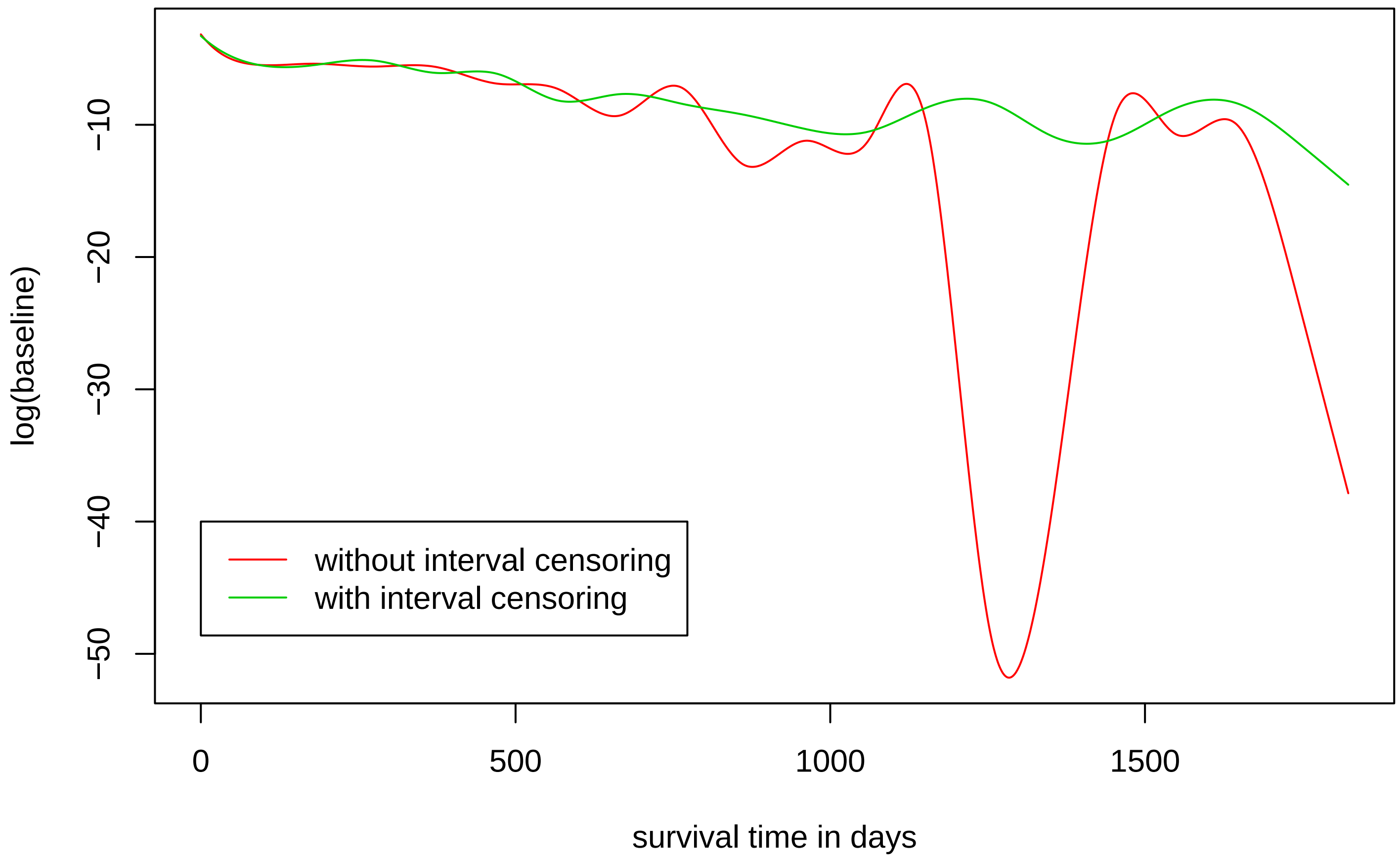


Spatial effect including covariates.

Age of the mother at birth.

Body mass index of the mother.





log(baseline)
−10
−20
−30
−40
−50
0
500
1000
1500
survival time in days
without interval censoring
with interval censoring

Discussion

- Empirical Bayesian treatment of complex geoadditive regression models:
  - Based on mixed model representation.
  - Applicable for a wide range of regression models.
  - Does not rely on MCMC simulation techniques.

    $\Rightarrow$ No questions on convergence and mixing of Markov chains, no hyperpriors.
  - Closely related to penalised likelihood estimation in a frequentist setting.
- Future work:
  - Extended modelling for categorical responses, e.g. based on correlated latent utilities.
  - Multi state models.
  - Interval censoring for multi state models.

References

- Fahrmeir, L., Kneib, T. & Lang, S. (2004): Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, 14, 715-745.

- Kneib, T. & Fahrmeir, L. (2005): Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, to appear.

- Kneib, T. (2005): Geoadditive hazard regression for interval censored survival times. SFB 386 Discussion Paper 447, University of Munich.

- Software and preprints:

`http://www.stat.uni-muenchen.de/~kneib`