

# Modelling Geoadditive Regression Data

Thomas Kneib

Department of Statistics, Ludwig-Maximilians-University Munich

joint work with

Stefan Lang (University of Innsbruck) & Ludwig Fahrmeir (University of Munich)



29.11.2007



# Outline

- Geoadditive Regression: An Application to Car Insurance Data.
- Bayesian Inference in Structured Additive Regression.
- Spatio-Temporal Regression: Forest Health Data.

# Structured Additive Regression

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

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

or

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

- General idea of structured additive regression: 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.

- Example: Car insurance data from two insurance companies in Belgium.
- Sample of approximately 160.000 policyholders.
- Aims: Separate **risk analyses for claim size and claim frequency** to predict risk premium from covariates.
- Variables of primary interest: Claim size  $y_i$  or claim frequency  $h_i$  of policyholders.
- **Covariates:**
  - vage* vehicles age
  - page* policyholders age
  - hp* vehicles horsepower
  - bm* bonus-malus score
  - s* district in Belgium
  - v* Vector of categorical covariates

- **Geoadditive models:**

- Gaussian model for log-costs  $\log(y)$ :

$$\log(y) \sim N(\eta, \sigma^2)$$

with

$$\eta = f_1(vage) + f_2(page) + f_3(bm) + f_4(hp) + f_{spat}(s) + v'\zeta.$$

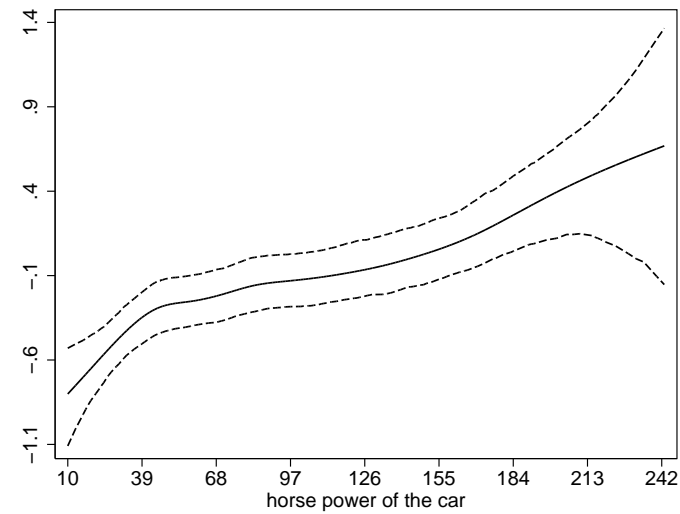
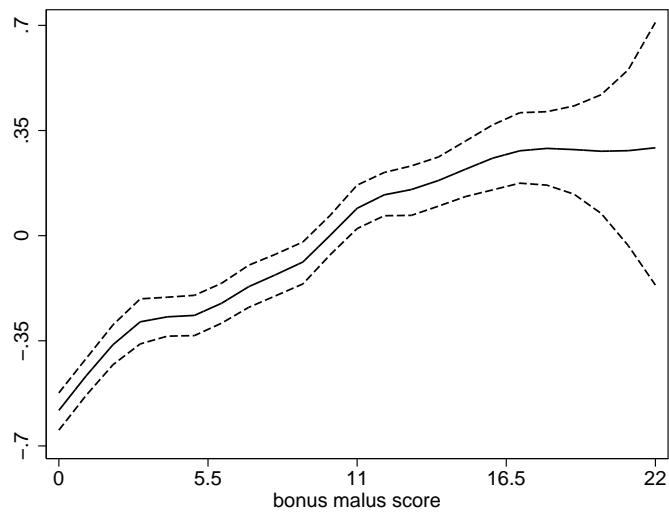
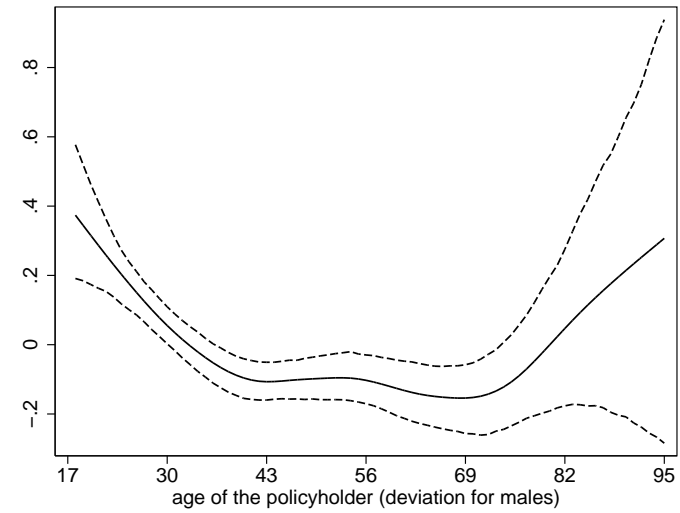
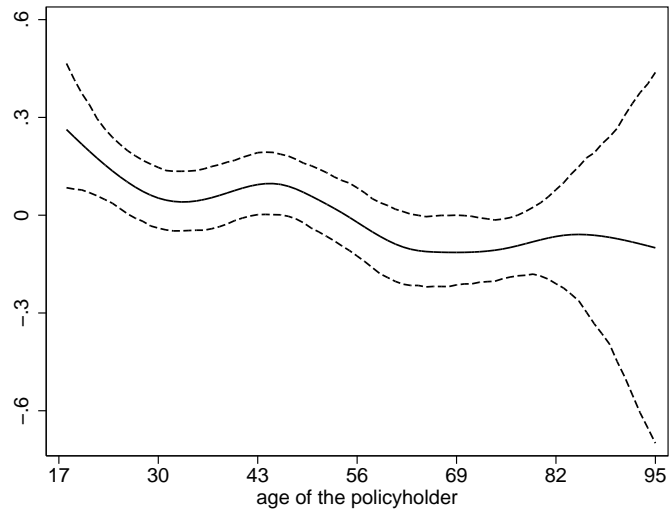
- Poisson model for frequencies  $h_i$ :

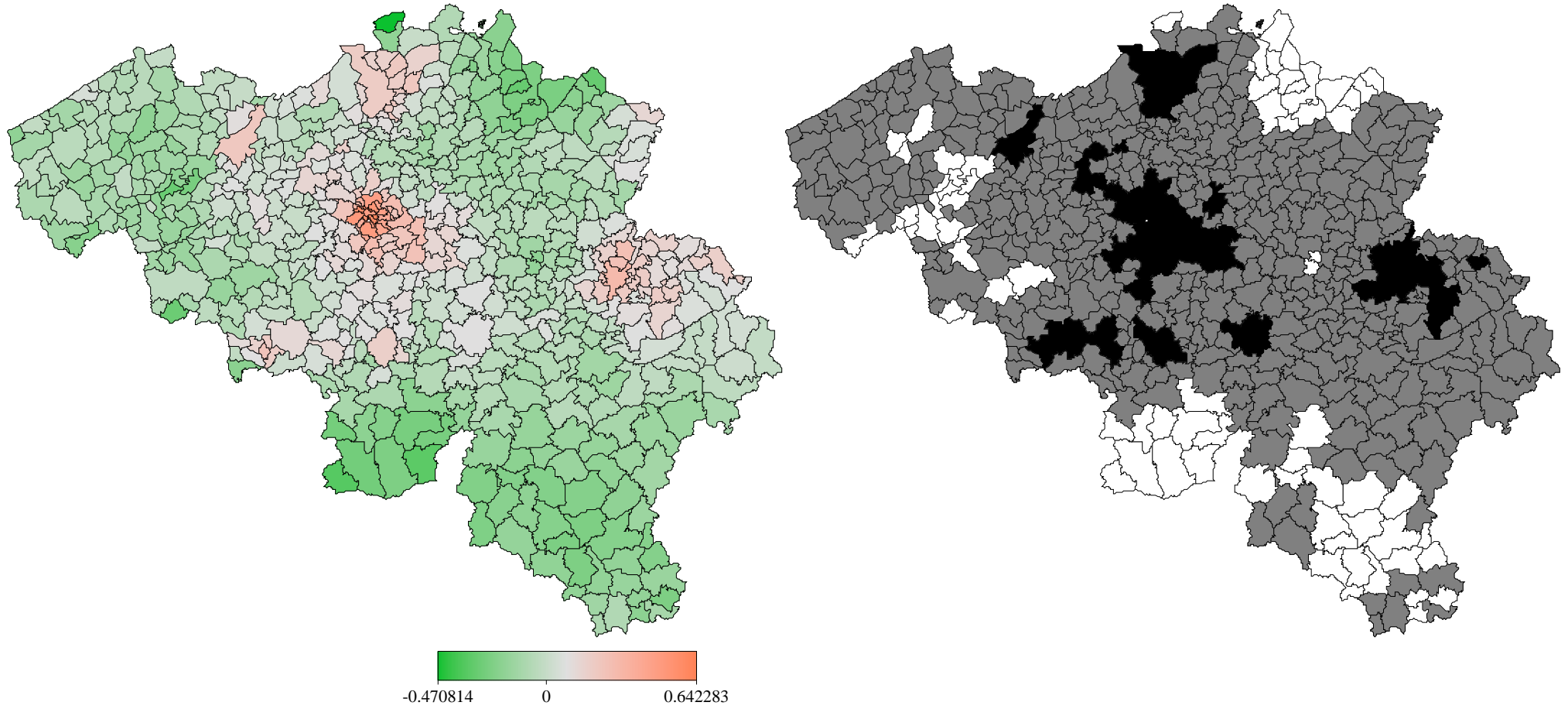
$$h \sim Po(\exp(\eta))$$

with

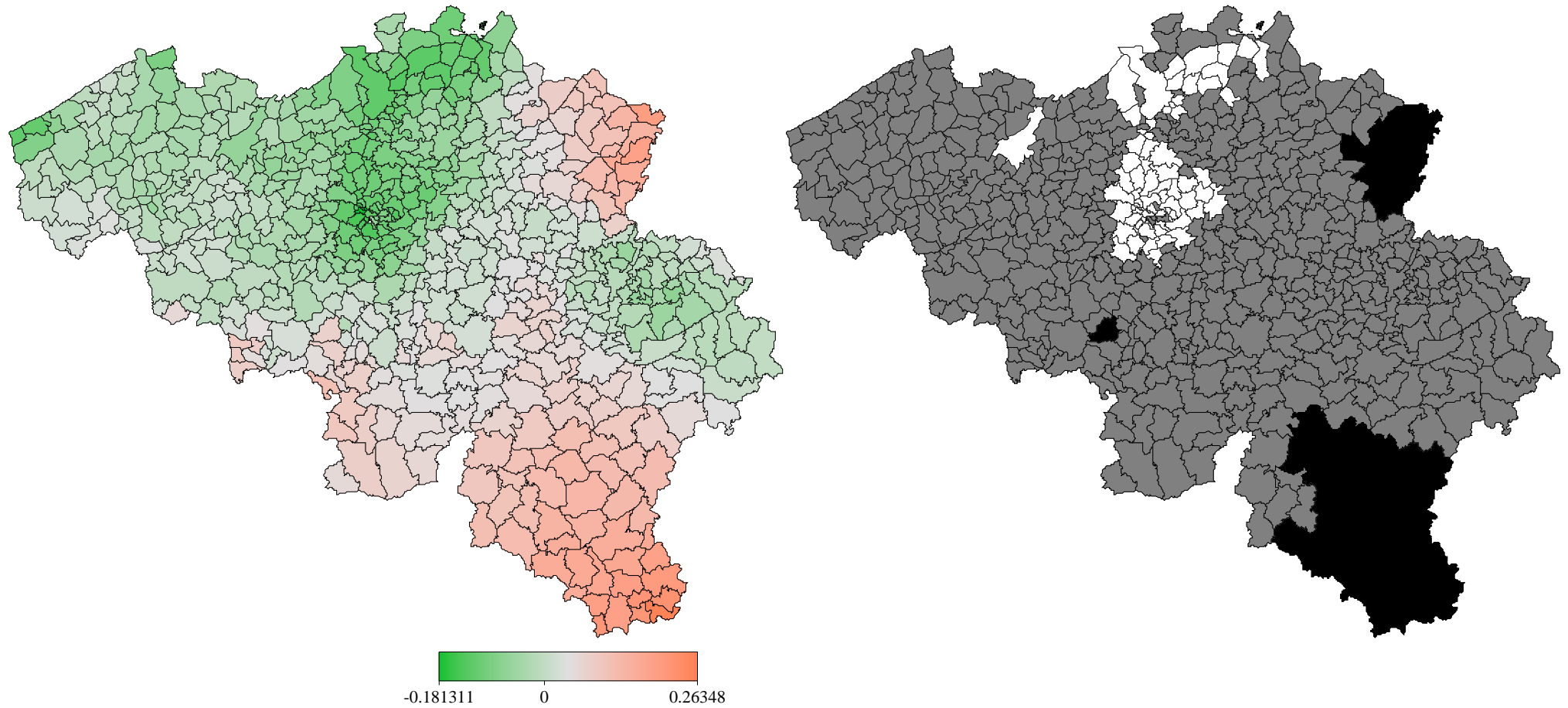
$$\eta = f_1(vage) + f_2(page) + f_3(page)sex + f_3(bm) + f_4(hp) + f_{spat}(s) + v'\zeta.$$

- Results for claim frequency:





- Spatial effect for claim size:

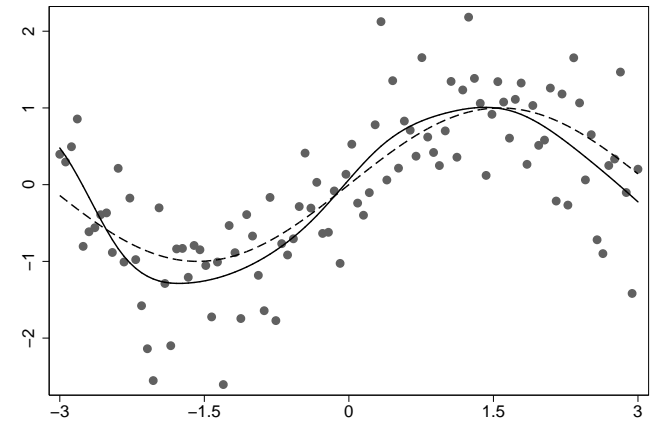
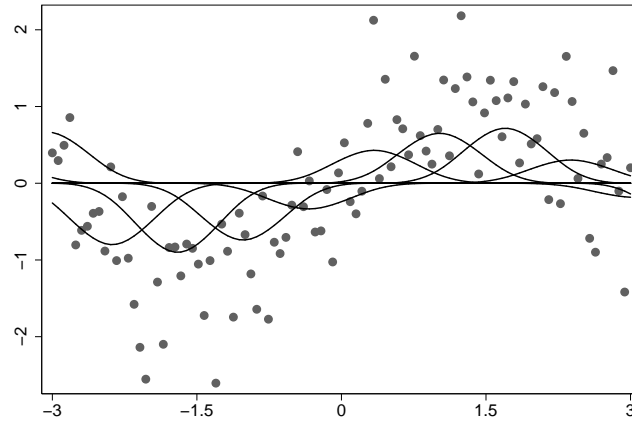
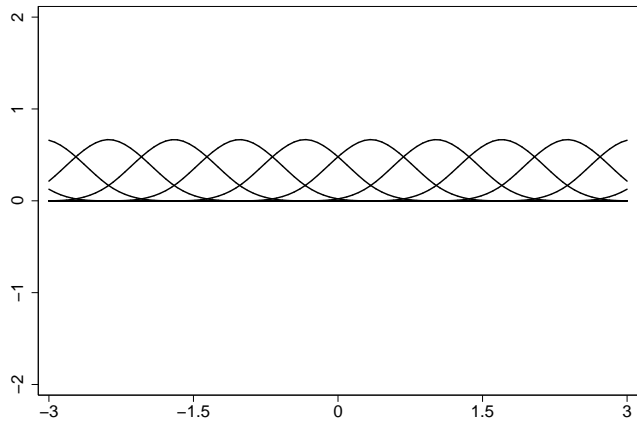


## Model Components and Priors

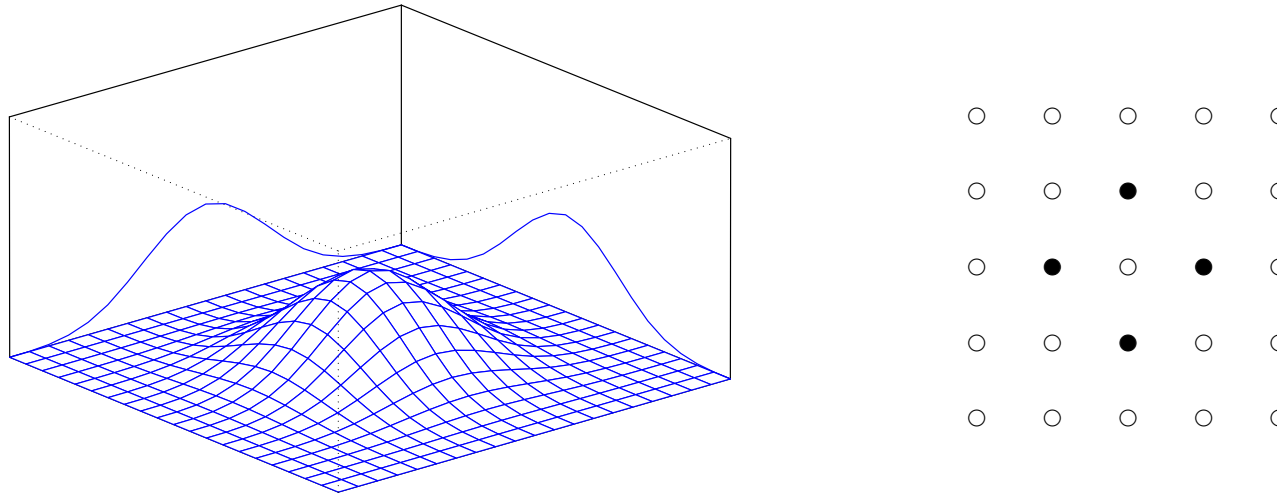
- **Penalised splines.**
  - Approximate  $f(x) = \sum \xi_j B_j(x)$  by a weighted sum of **B-spline basis** functions.
  - Employ a large number of basis functions to enable flexibility.
  - **Penalise differences** between parameters of adjacent basis functions to ensure smoothness

$$\frac{1}{2\tau^2} \sum (\xi_j - \xi_{j-1})^2 \quad (\text{first order differences})$$

$$\frac{1}{2\tau^2} \sum (\xi_j - 2\xi_{j-1} + \xi_{j-2})^2 \quad (\text{second order differences})$$



- **Bivariate** penalised splines.



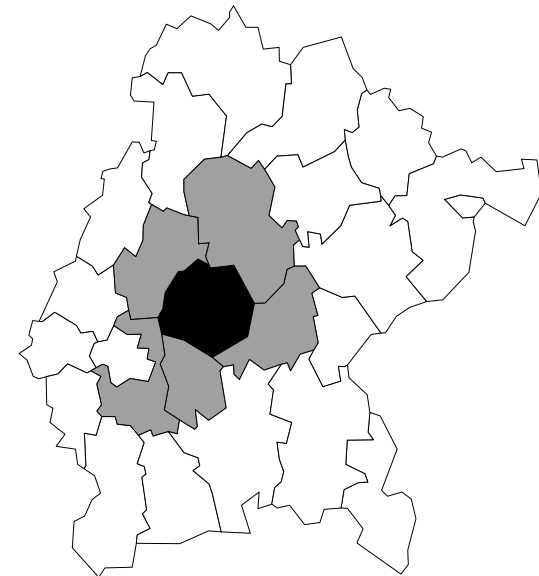
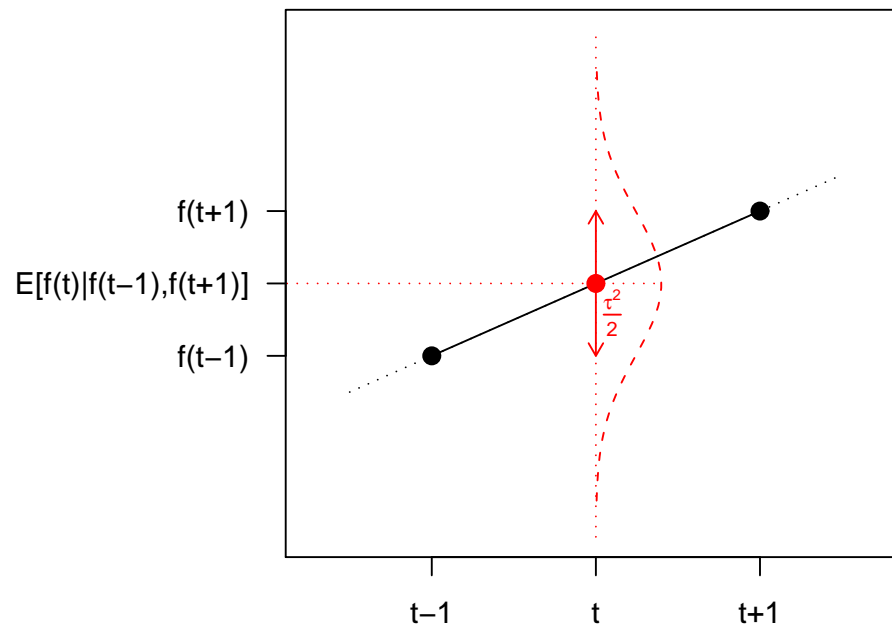
- **Varying coefficient models.**

- Effect of covariate  $x$  varies smoothly over the domain of a second covariate  $z$ :

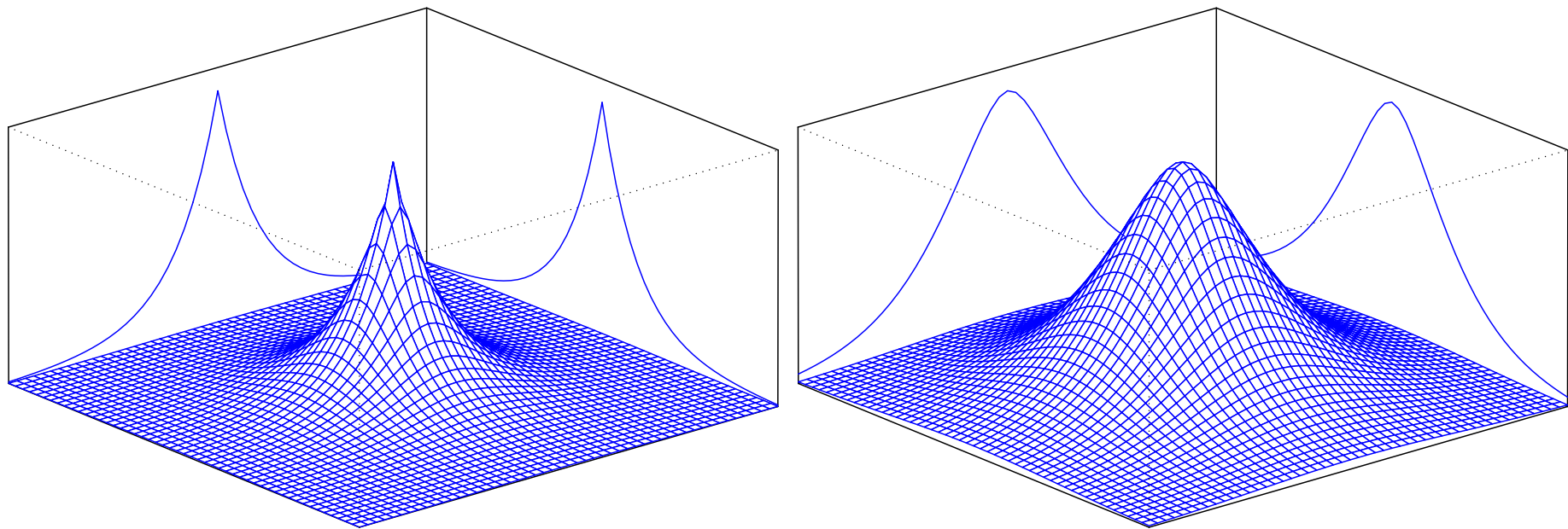
$$f(x, z) = x \cdot g(z)$$

- 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**.



- All effects can be cast into one **general framework**.
- All vectors of function evaluations  $f_j$  can be expressed as

$$f_j = Z_j \xi_j$$

with design matrix  $Z_j$  and regression coefficients  $\xi_j$ .

- **Generic form of the prior** for  $\xi_j$ :

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

- $K_j \geq 0$  acts as a **penalty matrix**,  $\text{rank}(K_j) = k_j \leq d_j = \dim(\xi_j)$ .
- $\tau_j^2 \geq 0$  can be interpreted as a **variance** or (inverse) **smoothness parameter**.

# Bayesian Inference

- Fully Bayesian inference:

- All parameters (including the variance parameters  $\tau^2$ ) are assigned suitable prior distributions.
- Typically, estimation is based on MCMC simulation techniques.
- Usual estimates: Posterior expectation, posterior median (easily obtained from the samples).

- Empirical Bayes inference:

- Differentiate between parameters of primary interest (regression coefficients) and hyperparameters (variances).
- Assign priors only to the former.
- Estimate the hyperparameters by maximising their marginal posterior.
- Plugging these estimates into the joint posterior and maximising with respect to the parameters of primary interest yields posterior mode estimates.

- MCMC-based inference:
  - Assign **inverse gamma prior** to  $\tau_j^2$ :

$$p(\tau_j^2) \propto \frac{1}{(\tau_j^2)^{a_j+1}} \exp\left(-\frac{b_j}{\tau_j^2}\right).$$

Proper for  $a_j > 0, b_j > 0$       Common choice:  $a_j = b_j = \varepsilon$  small.

Improper for  $b_j = 0, a_j = -1$       Flat prior for variance  $\tau_j^2$ ,

$b_j = 0, a_j = -\frac{1}{2}$       Flat prior for standard deviation  $\tau_j$ .

- **Conditions for proper posteriors** in structured additive regression are available.
- **Gibbs sampler** for  $\tau_j^2 | \cdot$ :

Sample from an inverse Gamma distribution with parameters

$$a'_j = a_j + \frac{1}{2} \text{rank}(K_j) \quad \text{and} \quad b'_j = b_j + \frac{1}{2} \xi_j' K_j \xi_j.$$

- **Metropolis-Hastings** update for  $\xi_j | \cdot$ :

Propose new state from a multivariate Gaussian distribution with precision matrix and mean

$$P_j = Z_j' W Z_j + \frac{1}{\tau_j^2} K_j \quad \text{and} \quad m_j = P_j^{-1} Z_j' W (\tilde{y} - \eta_{-j}).$$

**IWLS-Proposal** with appropriately defined working weights  $W$  and working observations  $\tilde{y}$ .

- Efficient algorithms make use of the sparse matrix structure of  $P_j$  and  $K_j$ .

- Empirical Bayes inference.
  - Consider the variances  $\tau_j^2$  as **unknown constants** to be estimated from their marginal posterior.
  - Consider the regression coefficients  $\xi_j$  as **correlated random effects** with multivariate Gaussian distribution
    - ⇒ Use mixed model methodology for estimation.
- Problem: In most cases **partially improper random effects distribution**.
- Mixed model representation: Decompose

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

where

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

⇒  $\beta_j$  is a **fixed effect** and  $b_j$  is an **i.i.d. random effect**.

- This yields a **variance components model** with predictor

$$\eta = X\beta + Vb$$

where in turn

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

- Obtain **empirical Bayes estimates** / **penalized likelihood estimates** via iterating
  - Penalized maximum likelihood for the regression coefficients  $\beta$  and  $b$ .
  - Restricted Maximum / Marginal likelihood for the variance parameters in  $Q$ :

$$L(Q) = \int L(\beta, b, Q)p(b)d\beta db \rightarrow \max_Q.$$

- Involves a Laplace approximation to the marginal likelihood (corresponding to REML estimation of variances in Gaussian mixed models).

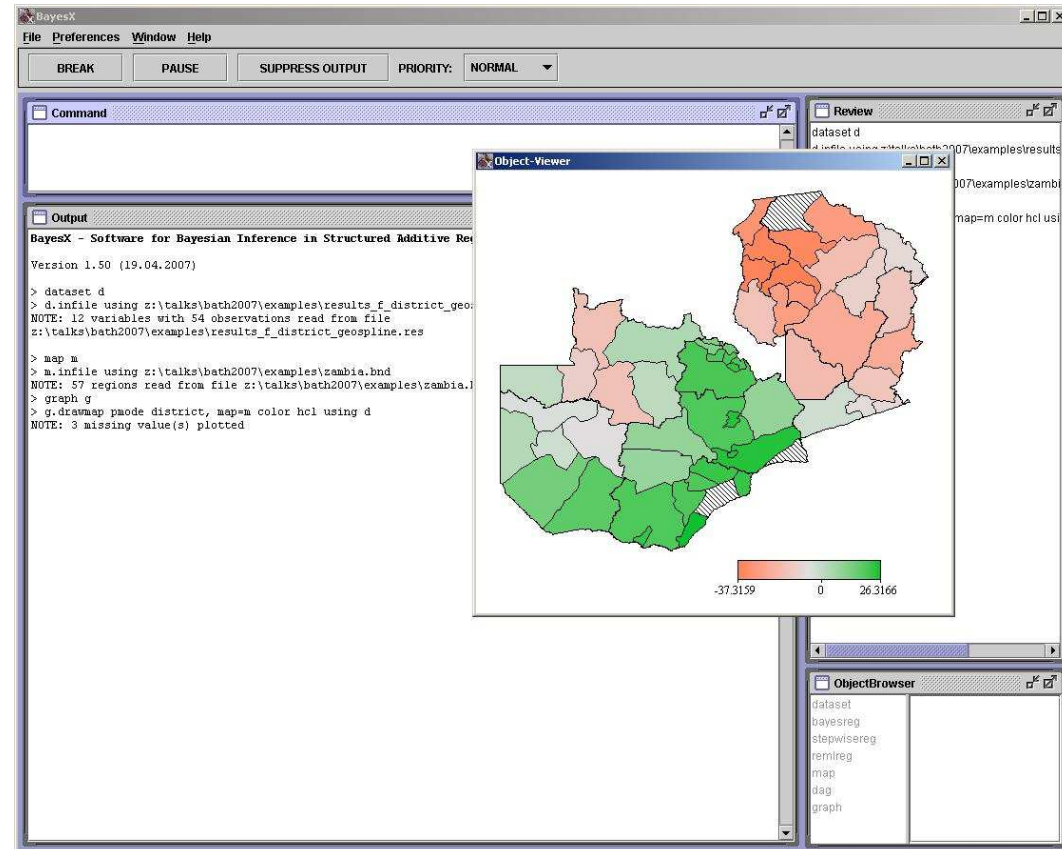
# BayesX

- BayesX is a software tool for estimating structured additive regression models.



- Available from

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



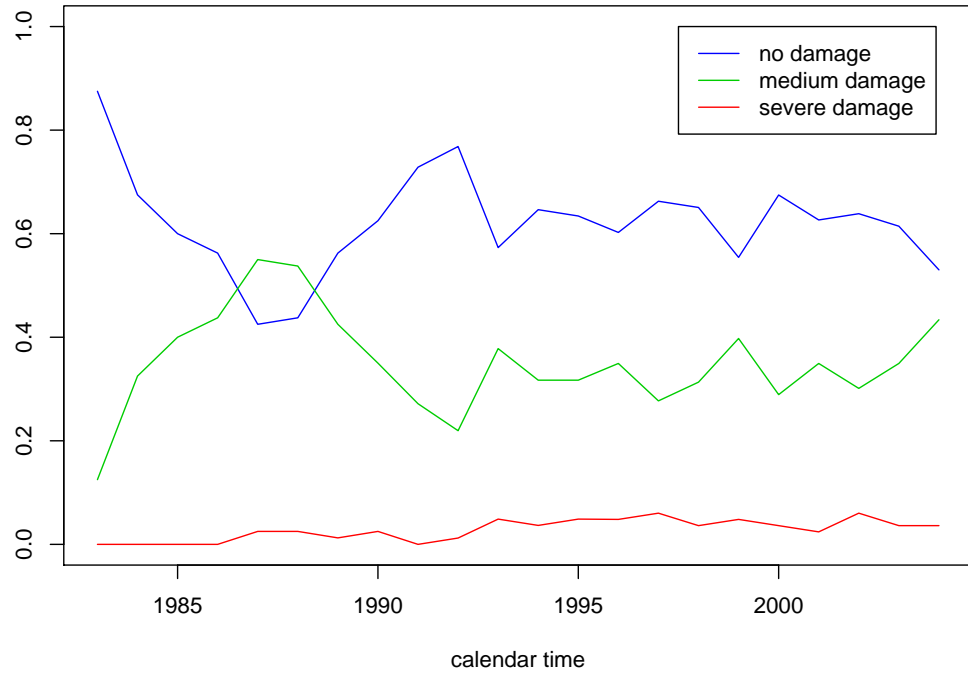
# Spatio-Temporal Regression: Forest Health Data

- 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:

$$\begin{aligned}y_{it} = 1 & \quad \text{no defoliation,} \\y_{it} = 2 & \quad \text{defoliation 25\% or less,} \\y_{it} = 3 & \quad \text{defoliation above 25\%}.\end{aligned}$$

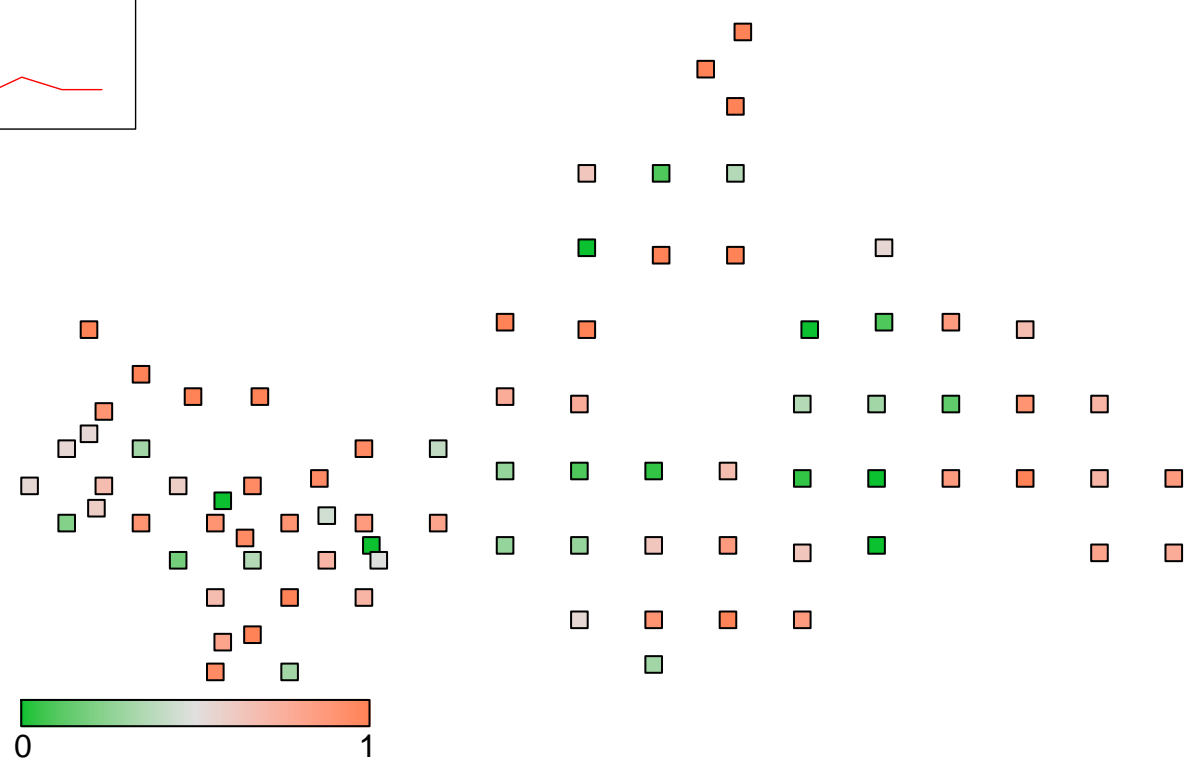
- Covariates:

$$\begin{aligned}t & \quad \text{calendar time,} \\s_i & \quad \text{site of the beech,} \\a_{it} & \quad \text{age of the tree in years,} \\u_{it} & \quad \text{further (mostly categorical) covariates.}\end{aligned}$$



Empirical time trends.

Empirical spatial effect.

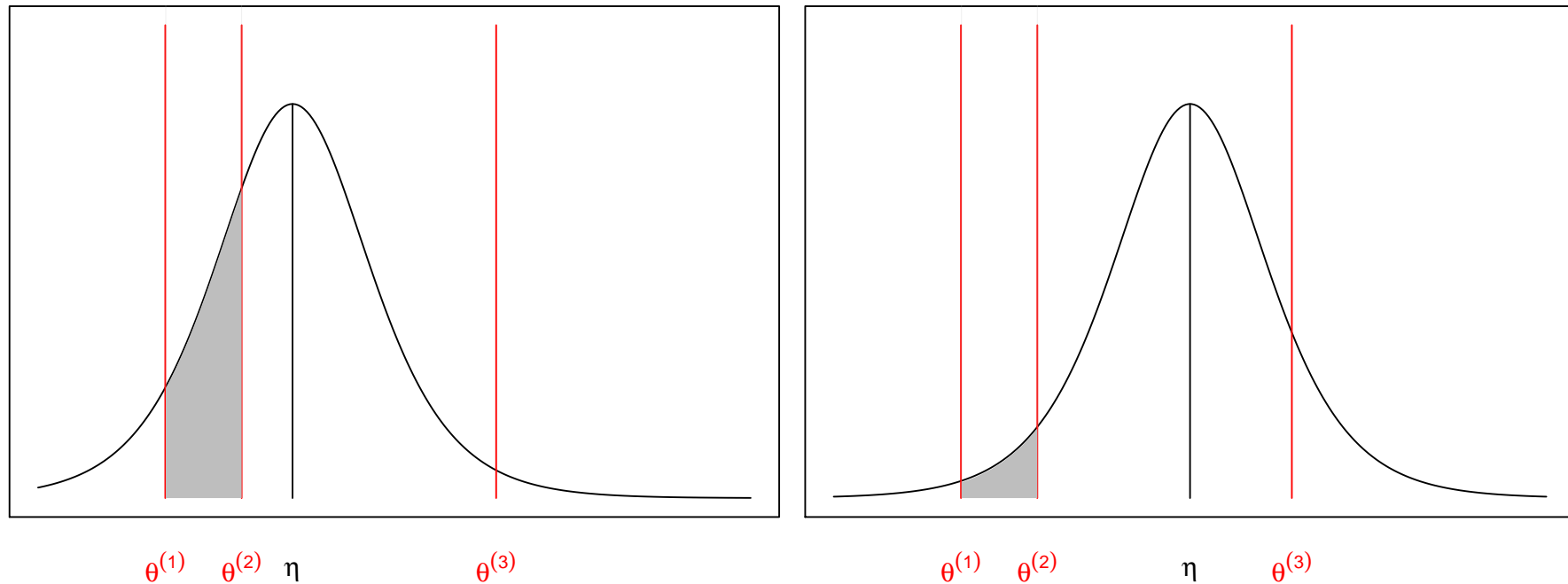


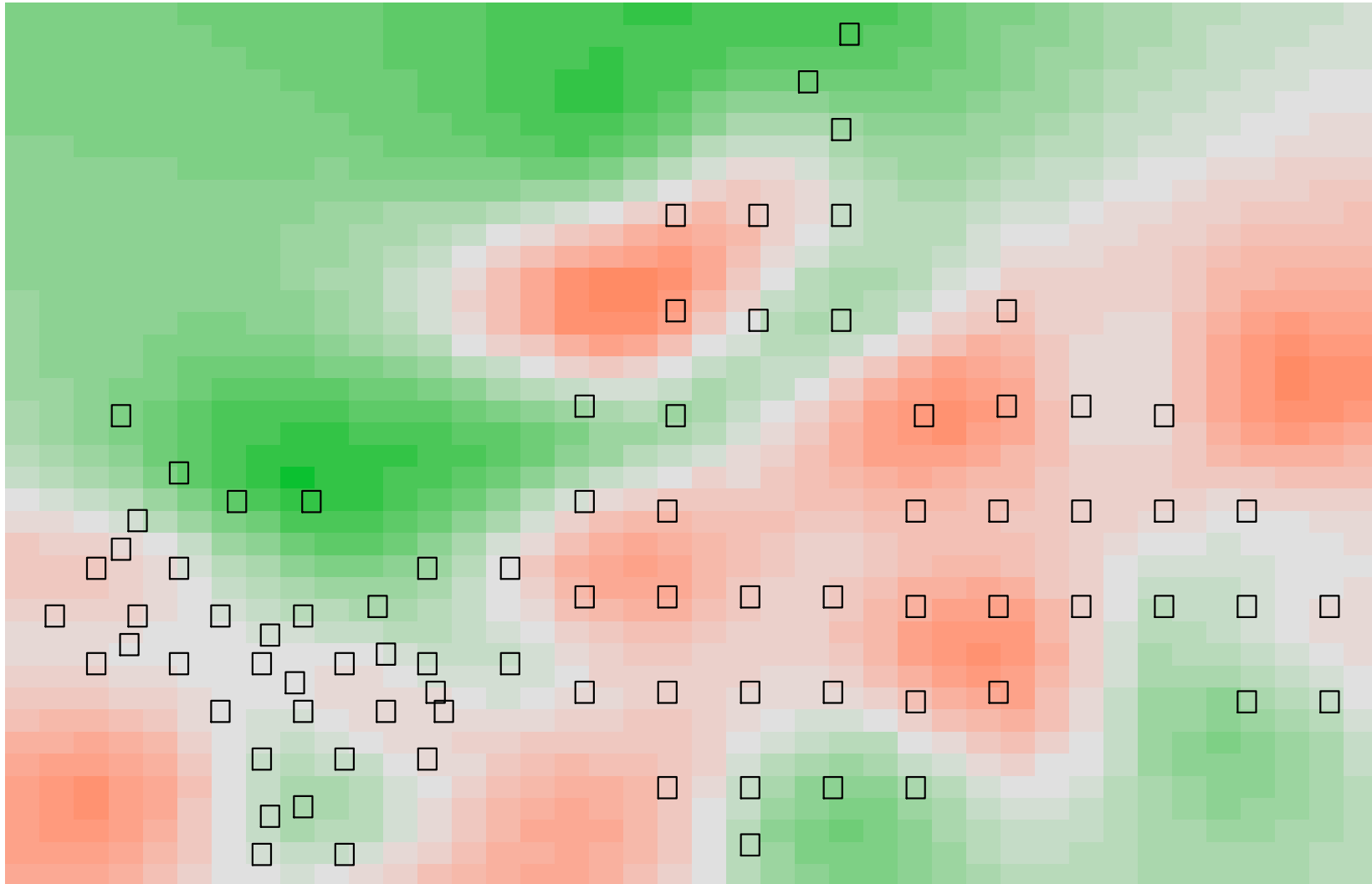
- Cumulative probit model:

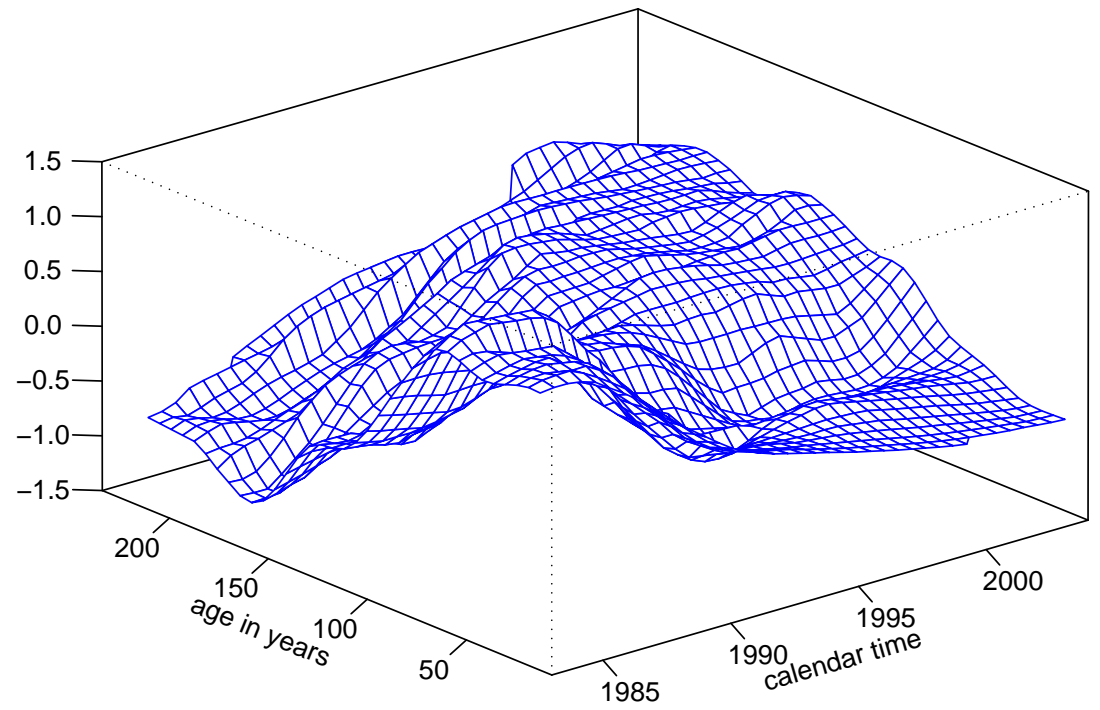
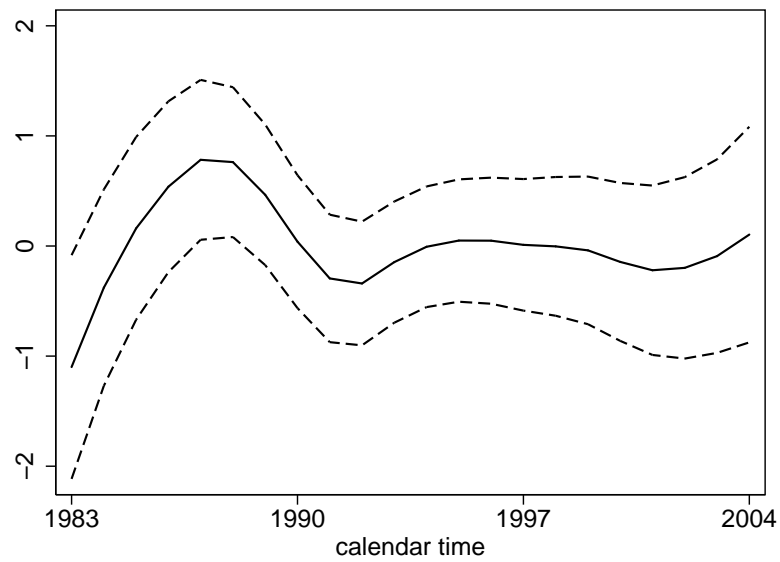
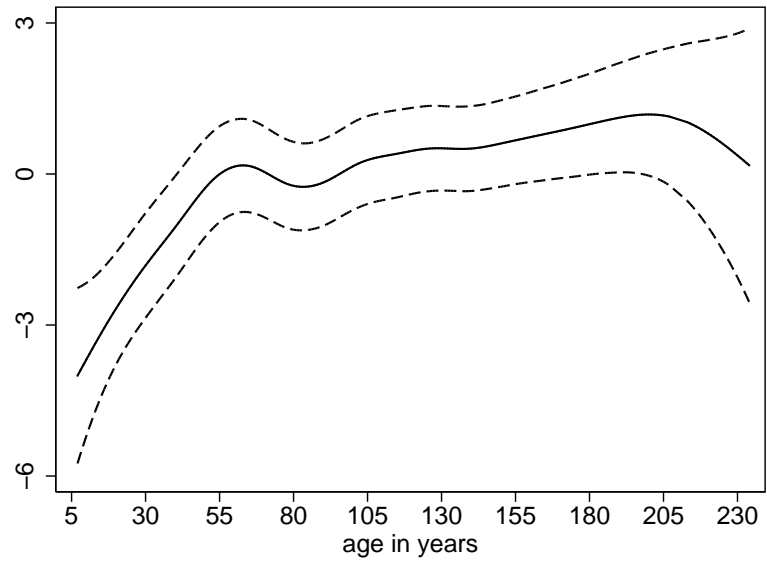
$$P(y_{it} \leq r) = \Phi \left( \theta^{(r)} - \eta_{it} \right)$$

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

$$\eta_{it} = f_1(t) + f_2(\text{age}_{it}) + f_3(t, \text{age}_{it}) + f_{\text{spat}}(s_i) + u'_{it}\gamma$$





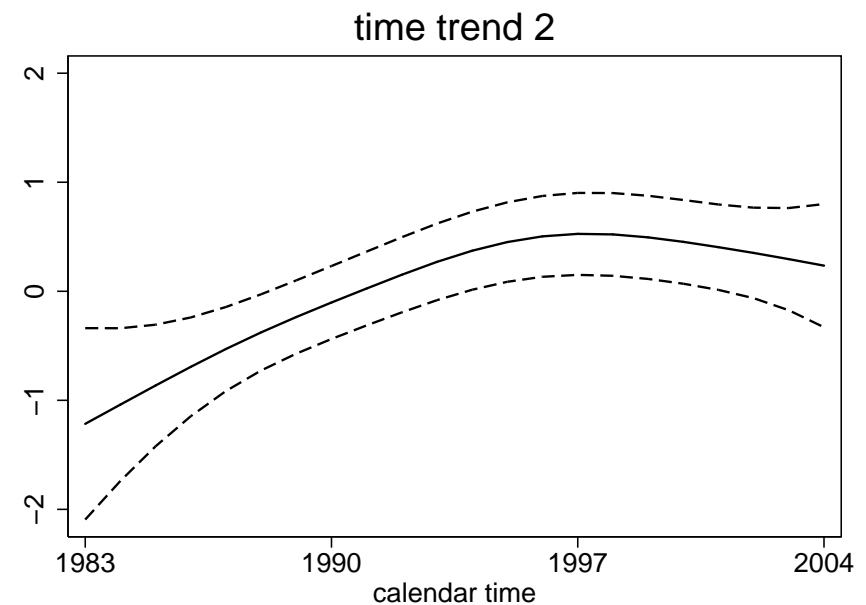
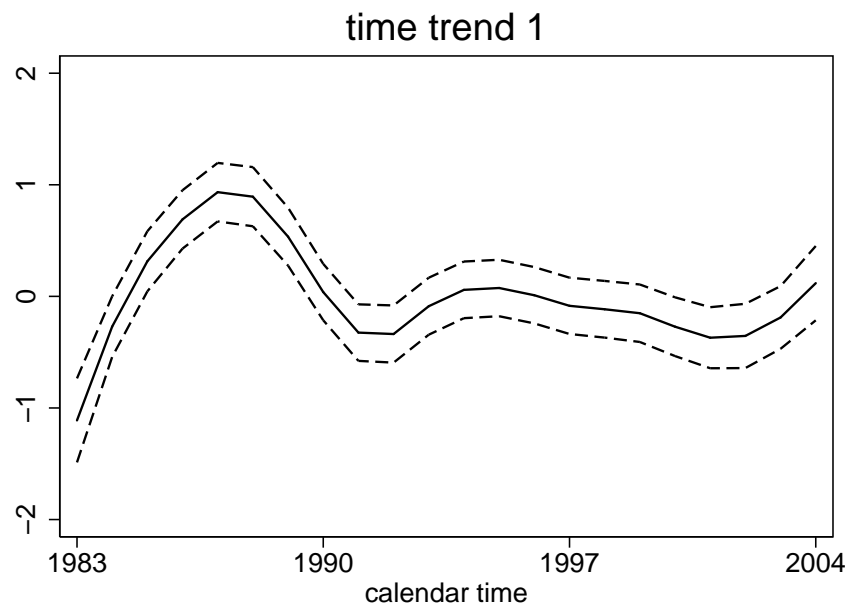


- Category-specific trends:

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

- More complicated constraints:

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



## Summary

- Flexible semiparametric regression models for geoadditive data structures.
- Fully automated Bayesian inferential procedures.
- Similar types of models are available for extended Cox-type hazard regression models:
  - Joint estimation of covariate effects and baseline hazard rate.
  - Time-varying effect to overcome proportional hazards.
  - Interval, left, and right censored survival times.
- A place called home:

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