BayesX: Analysing Geoadditive Regression Data

Thomas Kneib
Department of Statistics, Ludwig-Maximilians-University Munich

Joint work with
Andreas Brezger, Ludwig Fahrmeir & Stefan Lang

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

17.8.2007

Spatio-Temporal Regression Data

- Regression in a general sense:
  - Linear models and generalised linear models,
  - Multivariate (categorical) generalised linear models,
  - Regression models for duration times (Cox-type models, AFT models).

- Common structure: Model a quantity of interest in terms of categorical and continuous covariates, e.g.

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

  or

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

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

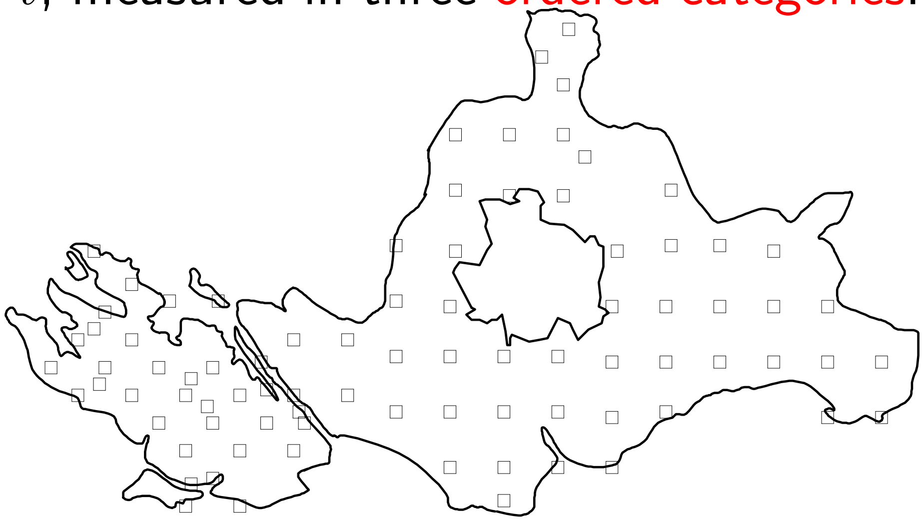
- Spatio-temporal regression models should allow
  - to account for spatial and temporal correlations,
  - for time- and space-varying effects,
  - for non-linear effects of continuous covariates,
  - for flexible interactions,
  - to account for unobserved heterogeneity.

$\Rightarrow$ Geoadditive regression models.

Example: Forest Health Data

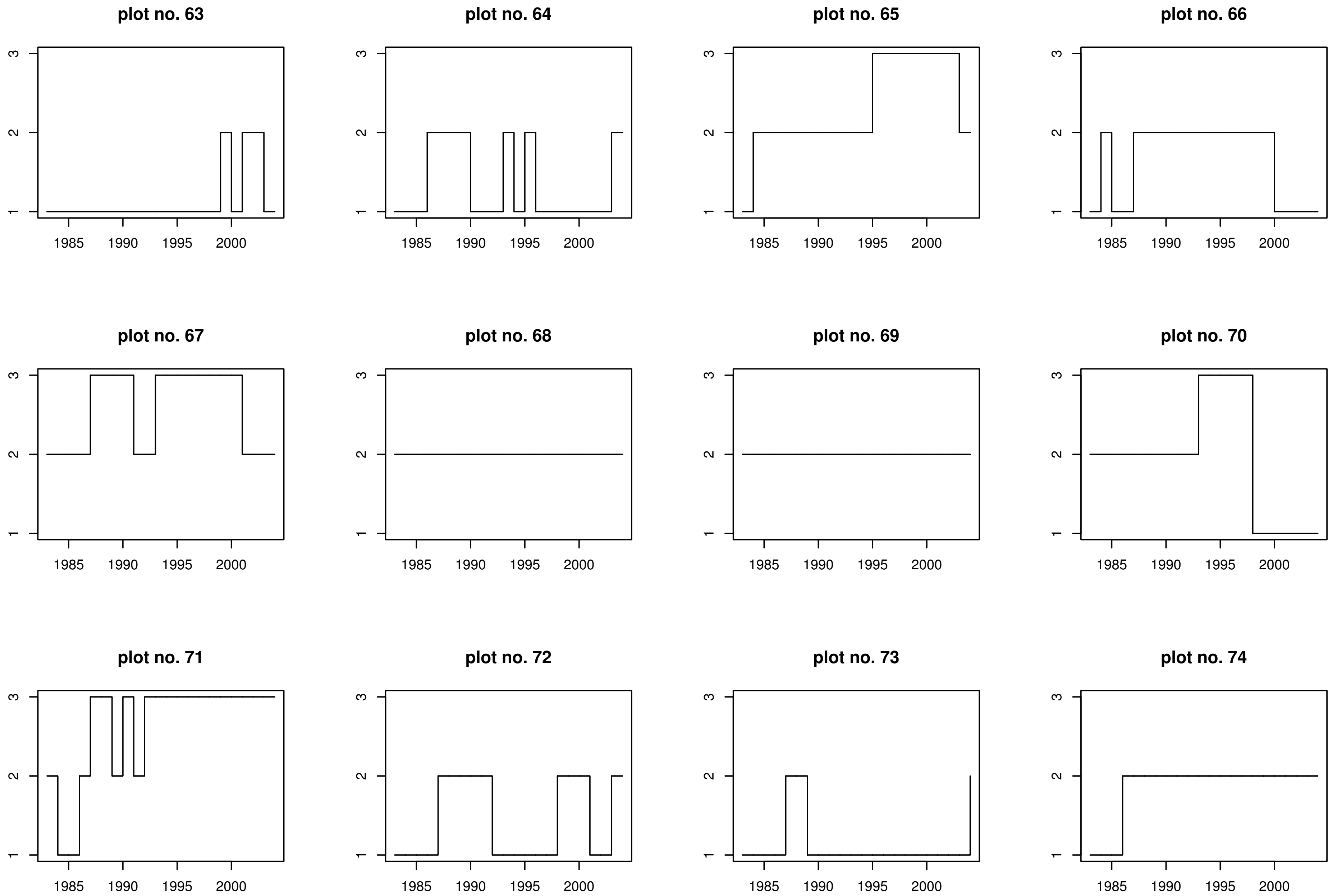
- Aim of the study: Identify factors influencing the health status of trees.
- Database: Yearly visual forest health inventories carried out from 1983 to 2004 in a northern Bavarian forest district.
- 83 observation plots of beeches within a 15 km times 10 km area.
- Response: defoliation degree at plot i in year t, measured in three ordered categories:

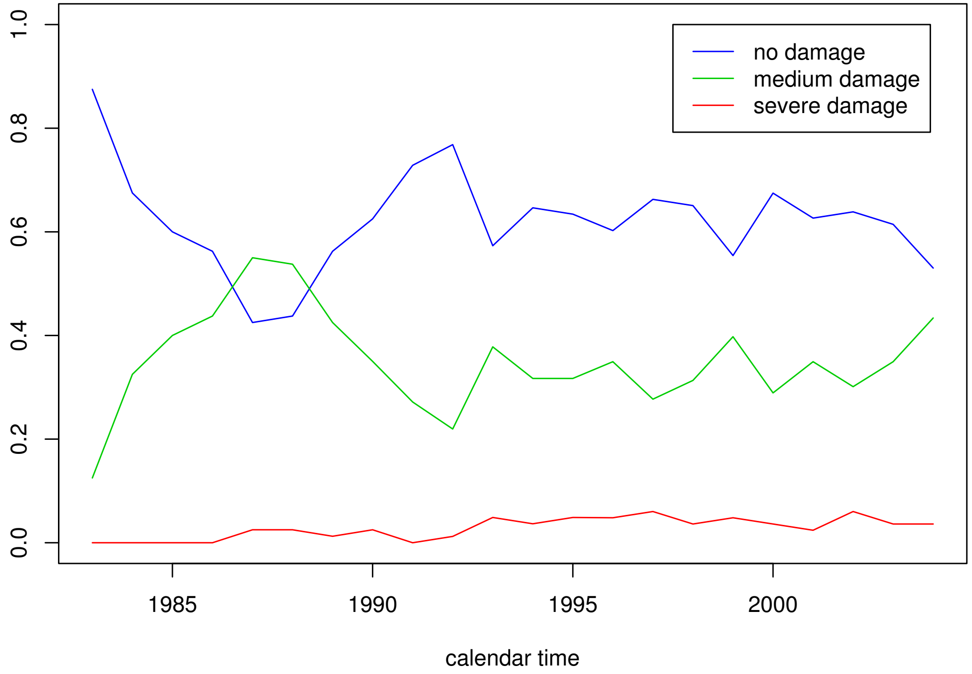
$$y_{it} = 1 \quad \text{no defoliation,}$$
$$y_{it} = 2 \quad \text{defoliation 25\% or less,}$$
$$y_{it} = 3 \quad \text{defoliation above 25\%.}$$



- Covariates:

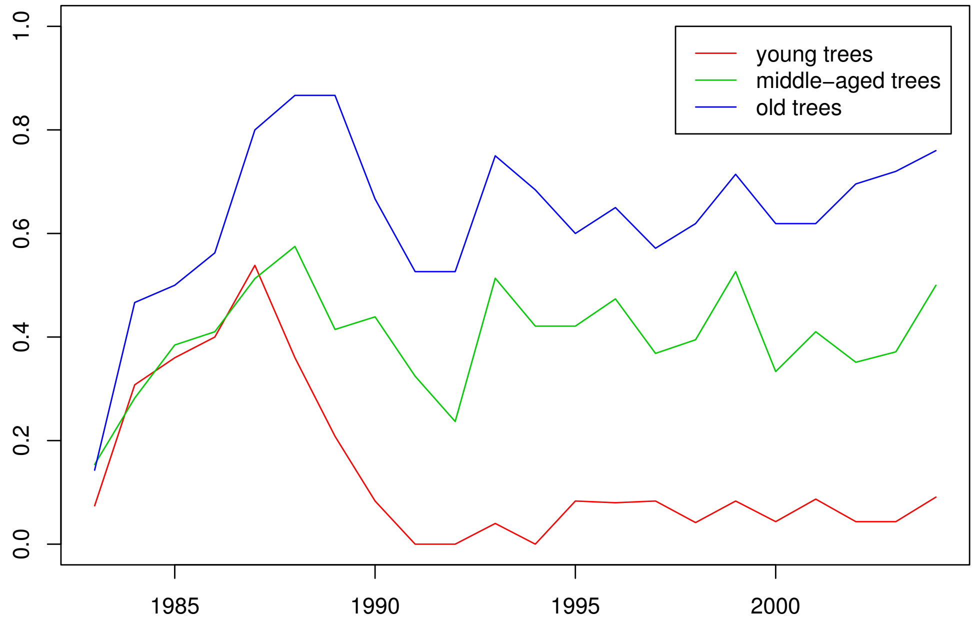
Continuous:	average age of trees at the observation plot
	elevation above sea level in meters
	inclination of slope in percent
	depth of soil layer in centimeters
	pH-value in 0-2cm depth
	density of forest canopy in percent
Categorical	thickness of humus layer in 5 ordered categories
	level of soil moisture
	base saturation in 4 ordered categories
Binary	type of stand
	application of fertilisation

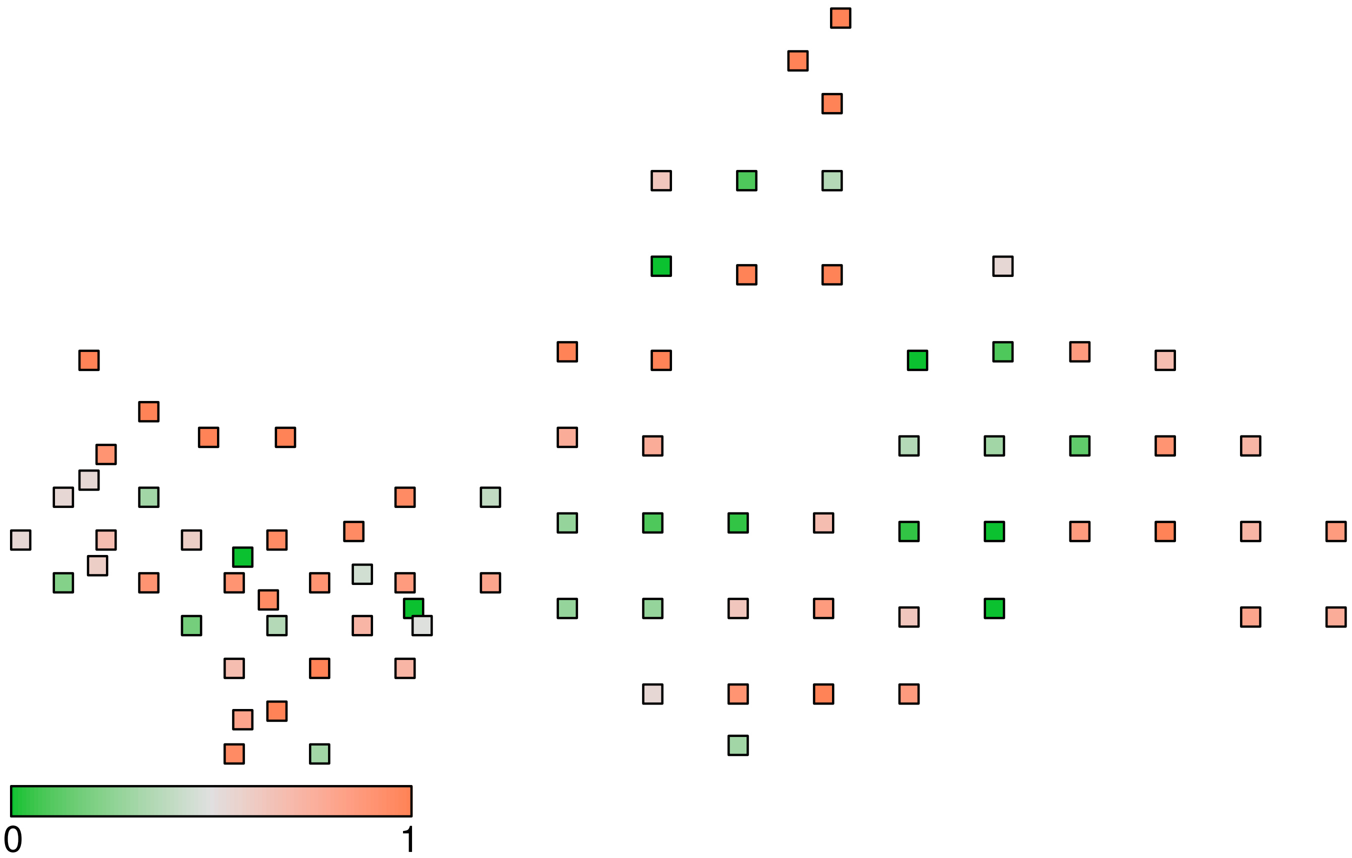

plot no. 63
plot no. 64
plot no. 65
plot no. 66
plot no. 67
plot no. 68
plot no. 69
plot no. 70
plot no. 71
plot no. 72
plot no. 73
plot no. 74
1
2
3
1985
1990
1995
2000



Empirical time trends.

Trends for different ages.





Percentage of time points for which a tree was classified to be damaged.

- We need a regression model that can simultaneously deal with the following issues:
  - A spatially aligned set of time series.

    ⇒ Both spatial and temporal correlations have to be considered.
  - Decide whether unobserved heterogeneity is spatially structured or not.
  - Non-linear effects of continuous covariates (e.g. age).
  - A possibly time-varying effect of age (i.e. an interaction between age and calendar time).
  - A categorical response variable.

Regression models for ordinal responses

- Defoliation degree is measured in three ordered categories.
- Derive regression models for ordinal responses based on latent variables:

$$D = x'\beta + \varepsilon.$$

- D can be considered an unobserved, continuous measure of defoliation.
- Link D to the categorical response Y based on ordered thresholds

$$-\infty = \theta^{(0)} < \theta^{(1)} < \theta^{(2)} < \theta^{(3)} = \infty$$

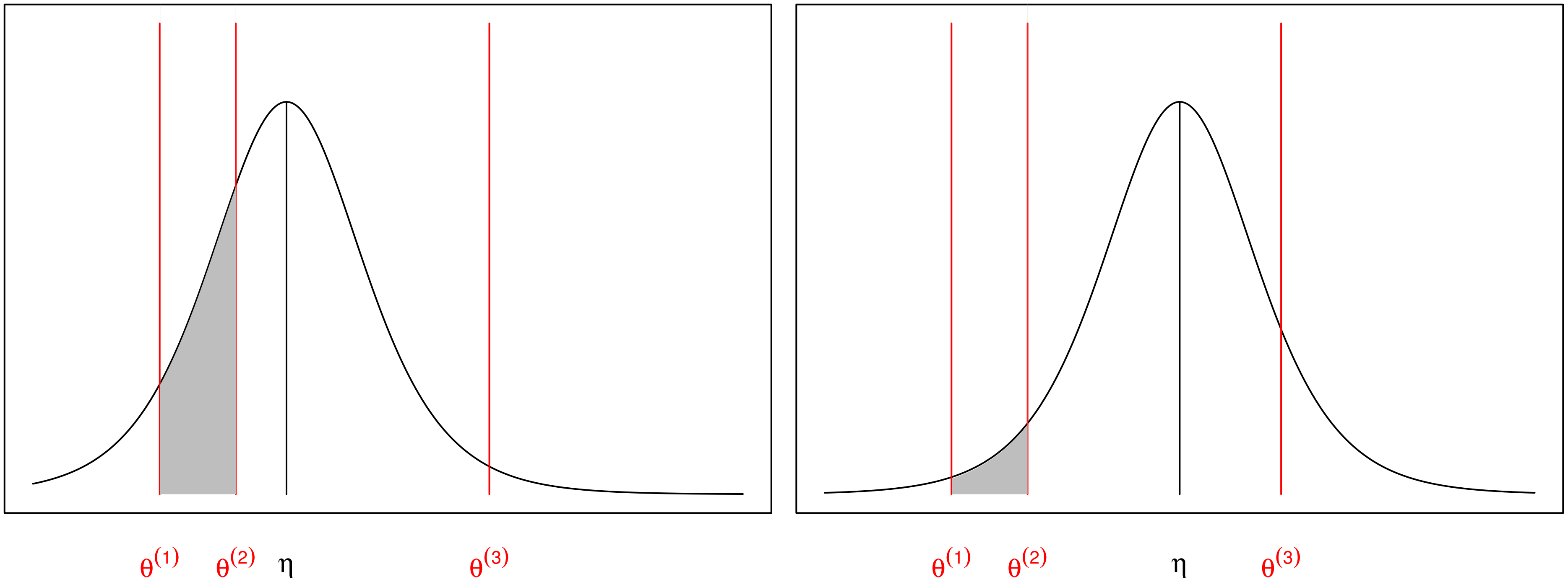
via

$$Y = r \quad \Leftrightarrow \quad \theta^{(r-1)} < D \leq \theta^{(r)}.$$

- Defines cumulative probabilities in terms of the cdf F of the latent error term ε:

$$P(Y \leq r) = P(D \leq \theta^{(r)}) = P(x'\beta + \varepsilon \leq \theta^{(r)}) = F(\theta^{(r)} - x'\beta).$$

- Intuitive interpretation:



- The thresholds slice the density $f = F'$.

- Suitable model in our application:

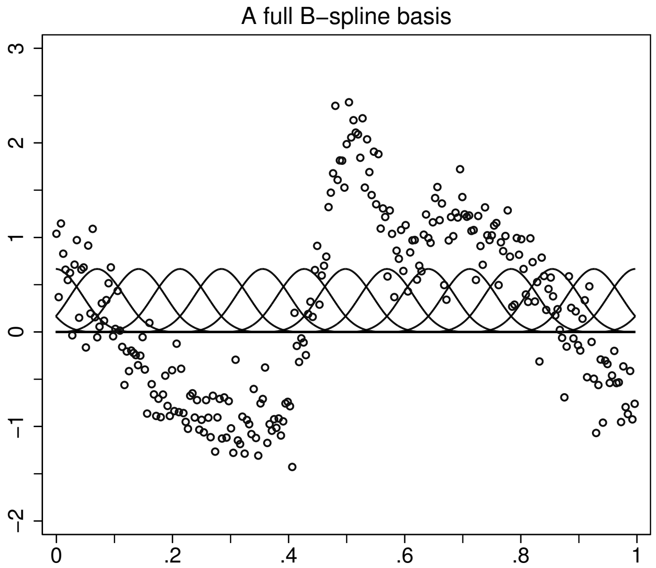
D_{it}	$=$	$f_1(age_{it})$	nonlinear effects of age,
		$+f_2(inc_i)$	inclination of slope, and
		$+f_3(can_{it})$	canopy density.
		$+f_{time}(t)$	nonlinear time trend.
		$+f_4(t, age_{it})$	interaction between age and calendar time.
		$+f_{spat}(s_i)$	structured and
		$+b_i$	unstructured spatial random effects.
		$+x'_{it}\gamma$	usual parametric effects.
		$+\varepsilon_{it}$	error term.

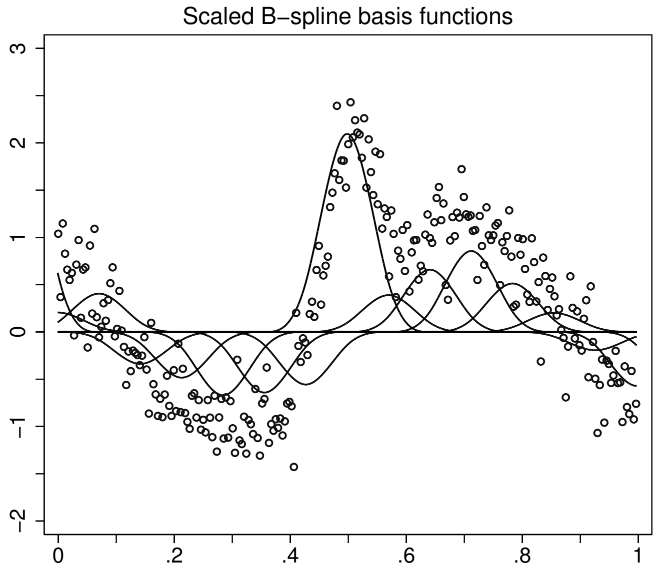
Penalised Splines

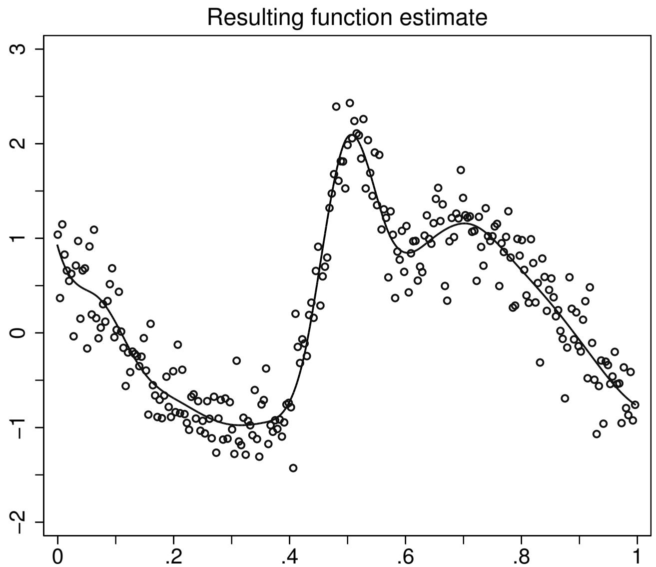
- Aim: Model nonparametric trend functions and nonparametric covariate effects.

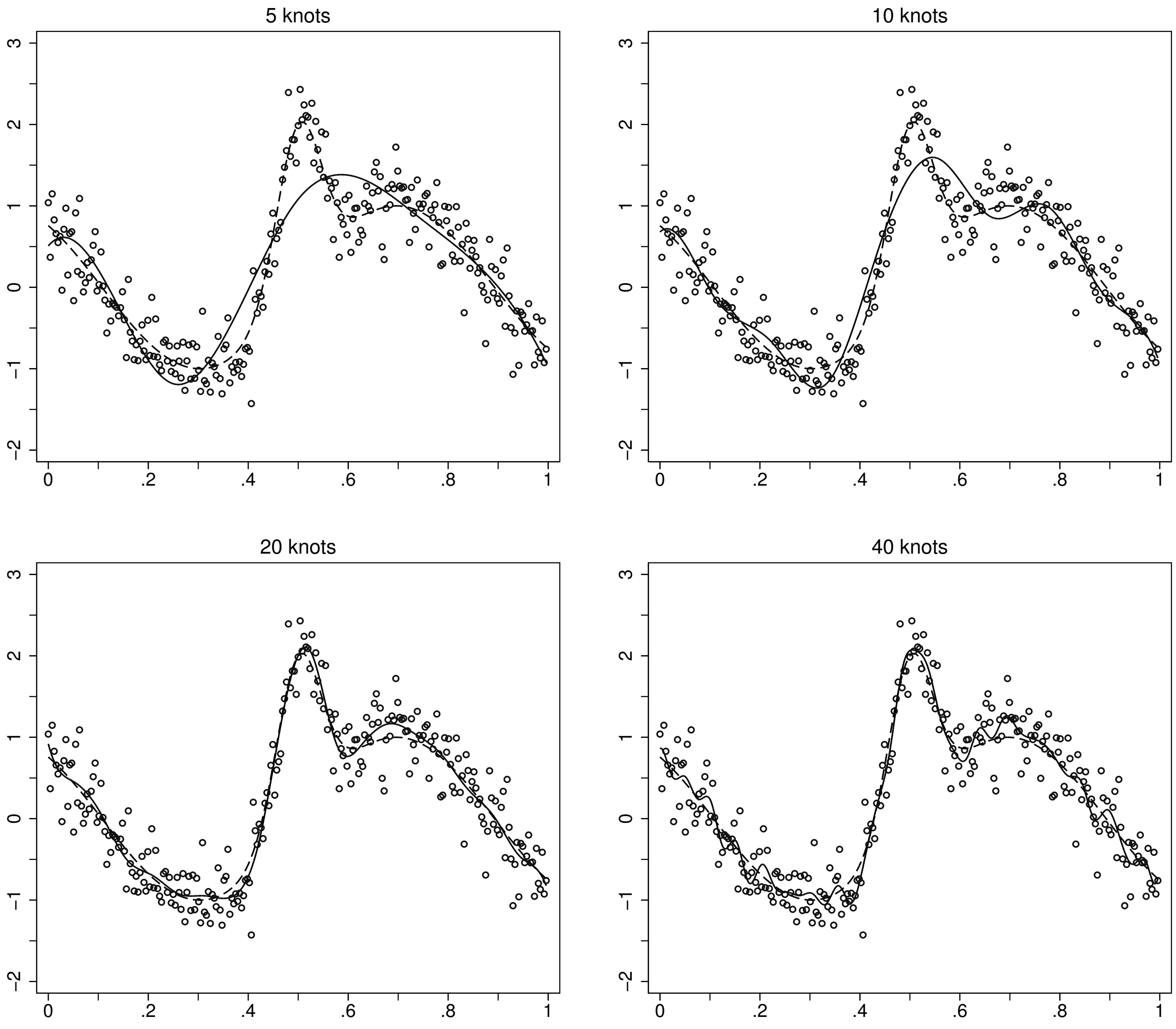
- Idea: Approximate $f(x)$ (or $f(t)$) by a weighted sum of B-spline basis functions:

$$f(x) = \sum_j \gamma_j B_j(x)$$


A full B−spline basis
3
2
1
0
−1
−2
0
.2
.4
.6
.8
1


Scaled B−spline basis functions
3
2
1
0
−1
−2
0
.2
.4
.6
.8
1


Resulting function estimate
3
2
1
0
−1
−2
0
.2
.4
.6
.8
1


5 knots
10 knots
20 knots
40 knots
0
.2
.4
.6
.8
1
−2
−1
0
1
2
3

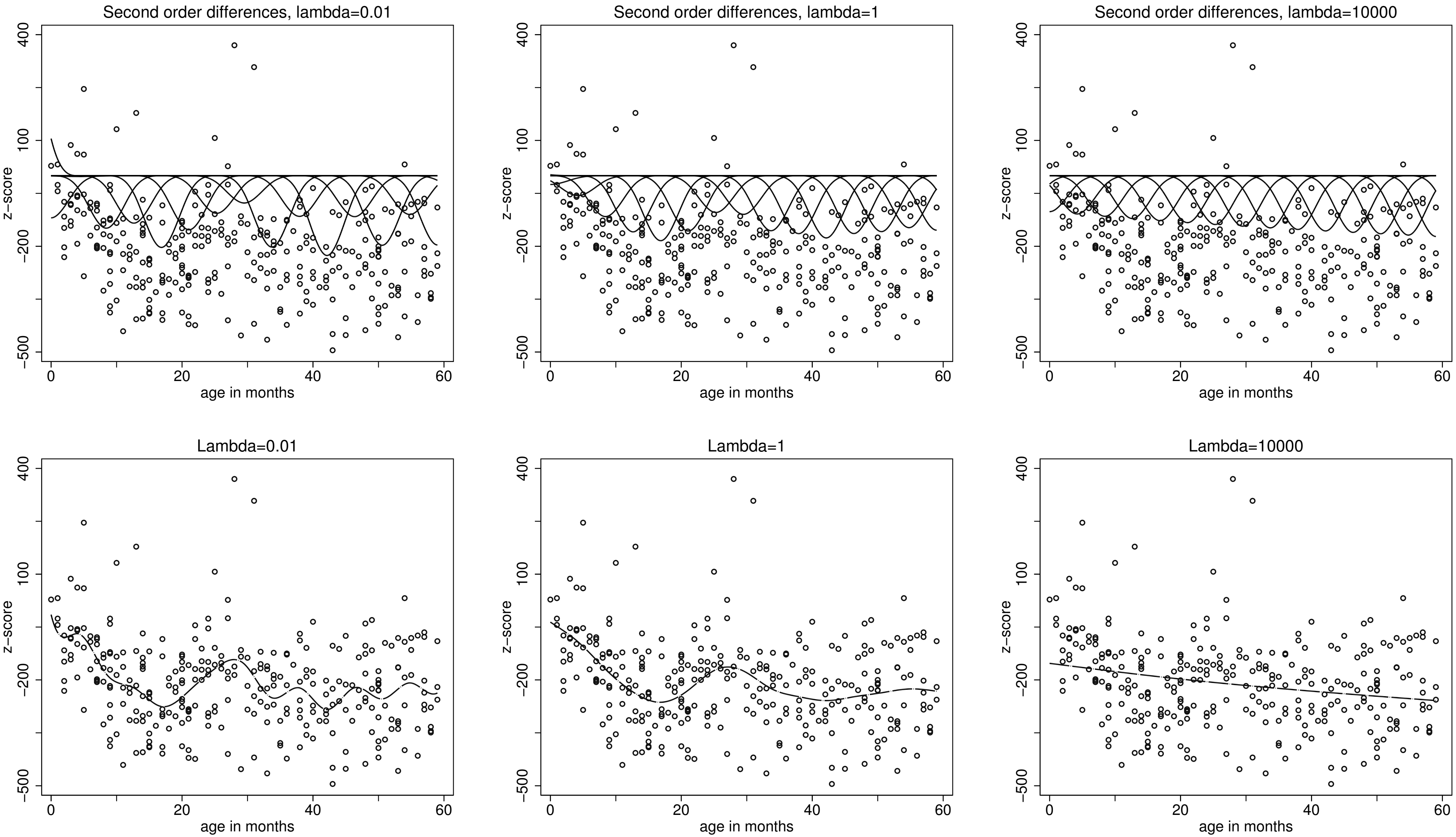
- The number of basis functions has significant impact on the function estimate.
- Employ a large number of basis functions to enable flexibility.
- Penalise differences between parameters of adjacent basis functions to ensure smoothness:

$$Pen(\gamma|\tau^2) = \frac{1}{2\tau^2}\sum_{j=2}^{p}(\gamma_j - \gamma_{j-1})^2 \qquad \text{first order differences}$$

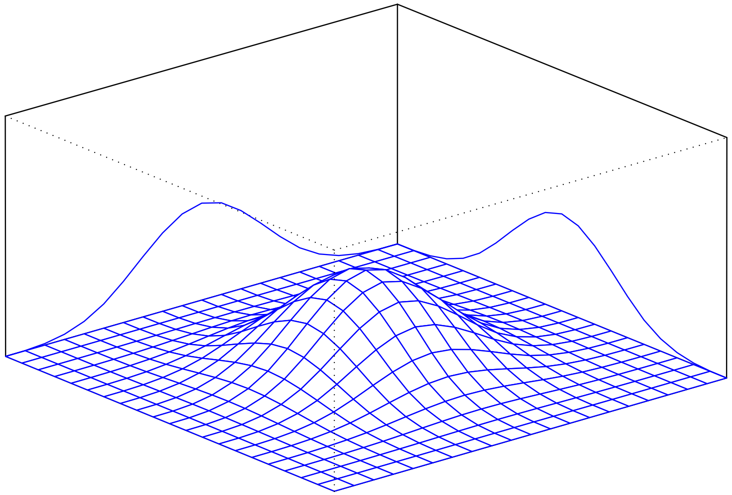
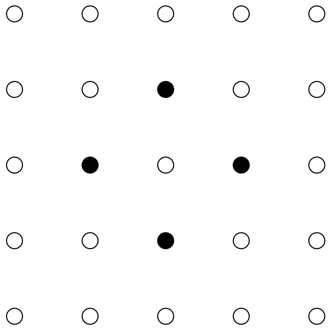
$$Pen(\gamma|\tau^2) = \frac{1}{2\tau^2}\sum_{j=3}^{p}(\gamma_j - 2\gamma_{j-1} + \gamma_{j-2})^2 \qquad \text{second order differences}$$

$\Rightarrow$ Penalised maximum likelihood estimation with smoothing parameter τ^2.

- A penalty term based on k-th order differences is an approximation to the integrated squared k-th derivative.
- Key question: Automatic selection of the smoothing parameter τ^2.


Second order differences, lambda=0.01
Second order differences, lambda=1
Second order differences, lambda=10000
Lambda=0.01
Lambda=1
Lambda=10000
z−score
age in months
400
100
−200
−500
0
20
40
60

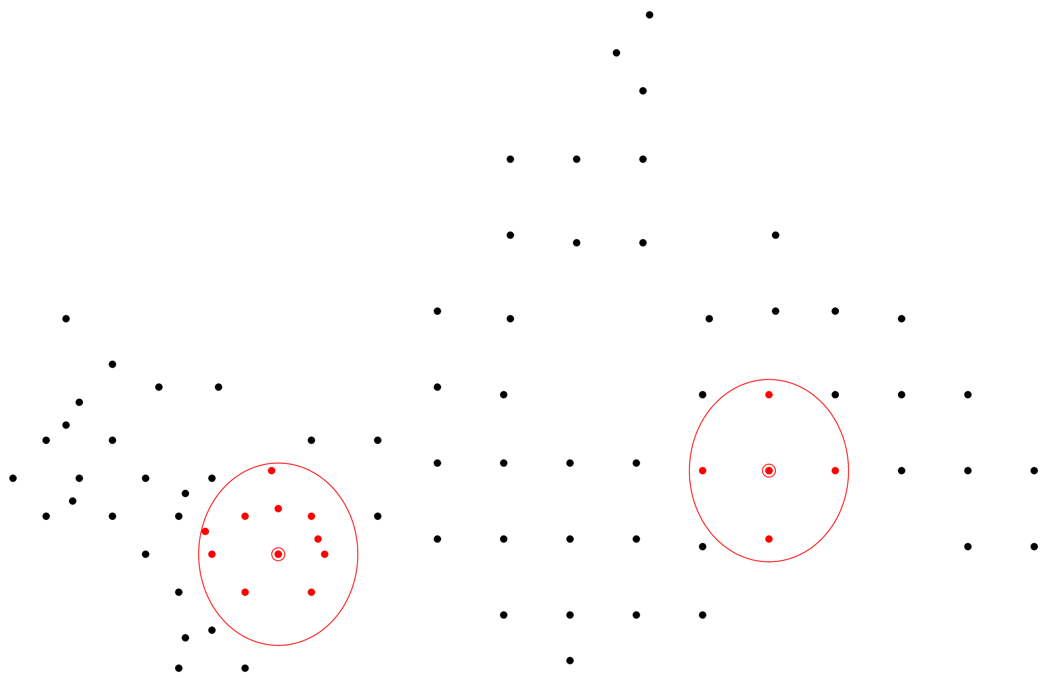
- Extension to bivariate penalised splines:
  - Bivariate basis functions based on tensor product B-splines.
  - Extend penalisation to neighbours on a grid.

⇒ Modelling of interaction surfaces (and spatial effects).

Spatial Modelling

- Markov random fields: Structured spatial effect.
- Bivariate extension of a first order random walk on the real line.
- Define two observation plots as neighbours if their distance is less than 1.2km.

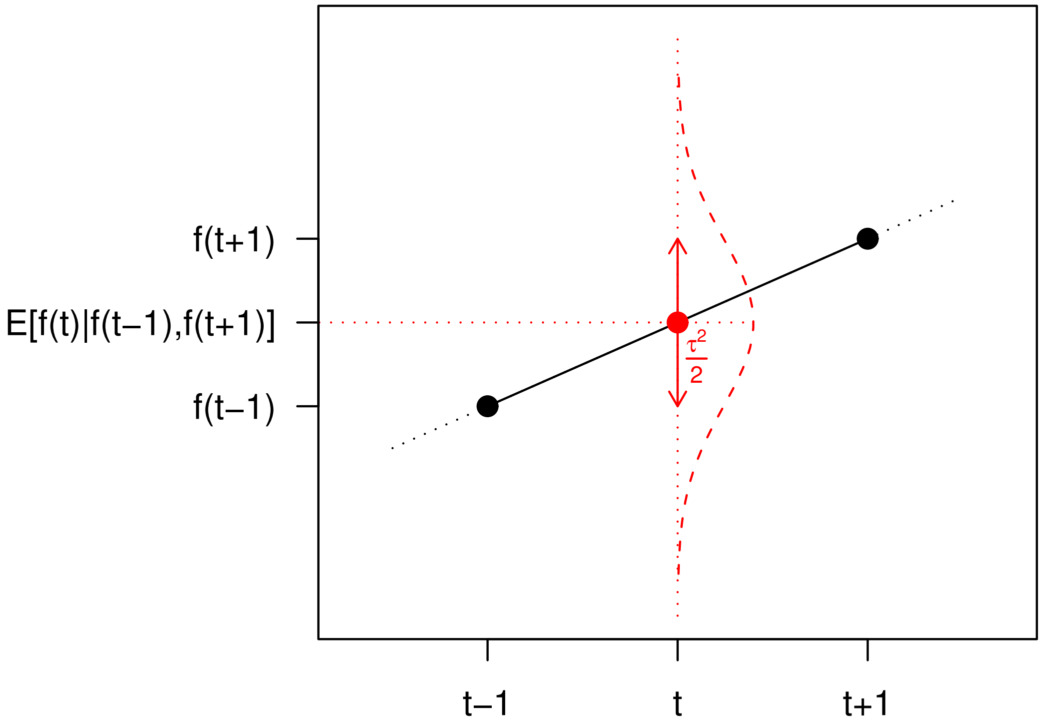


- Assume that the expected value of $\gamma_s = f_{spat}(s)$ is the average of the function evaluations of adjacent sites:

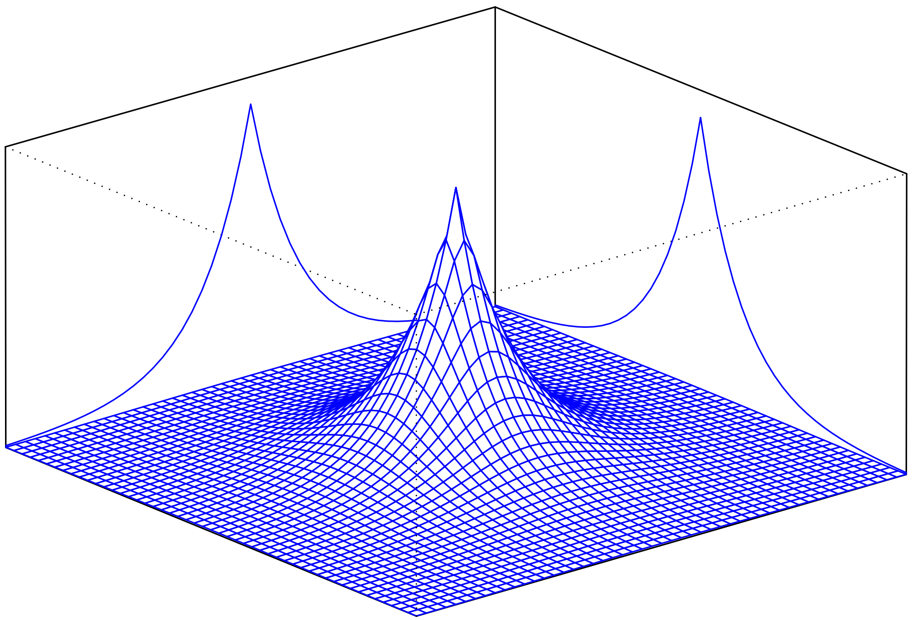
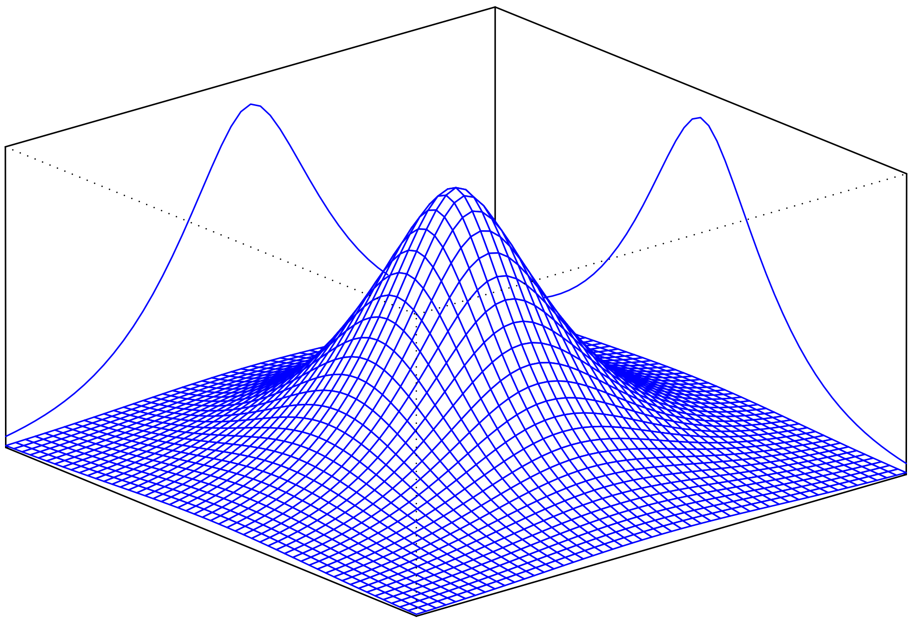
$$\gamma_s | \gamma_r, r \neq s \sim N\left(\frac{1}{N_s}\sum_{r\in\delta_s}\gamma_r, \frac{\tau^2}{N_s}\right)$$

where

δ_s	set of neighbors of plot s
N_s	no. of such neighbors.



- Kriging: Structured spatial effect.
- Assume a zero mean stationary Gaussian process for the spatial effect $\gamma_s = f_{spat}(s)$.
- Correlation of two sites is defined by an intrinsic correlation function.
- Can be interpreted as a basis function approach with radial basis functions.

- I.i.d. random effects: Unstructured spatial effect

$$\gamma_s \text{ i.i.d. } N(0, \tau^2).$$

- Also accounts for longitudinal structure of the data.

- Requires multiple measurements per observation plot.

Bayesian Inference

- Each term in the geoadditive predictor is associated with a vector of regression coefficients with improper multivariate Gaussian prior:

$$p(\gamma|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\gamma' K \gamma\right).$$

- The log-prior can be interpreted as a penalty term.

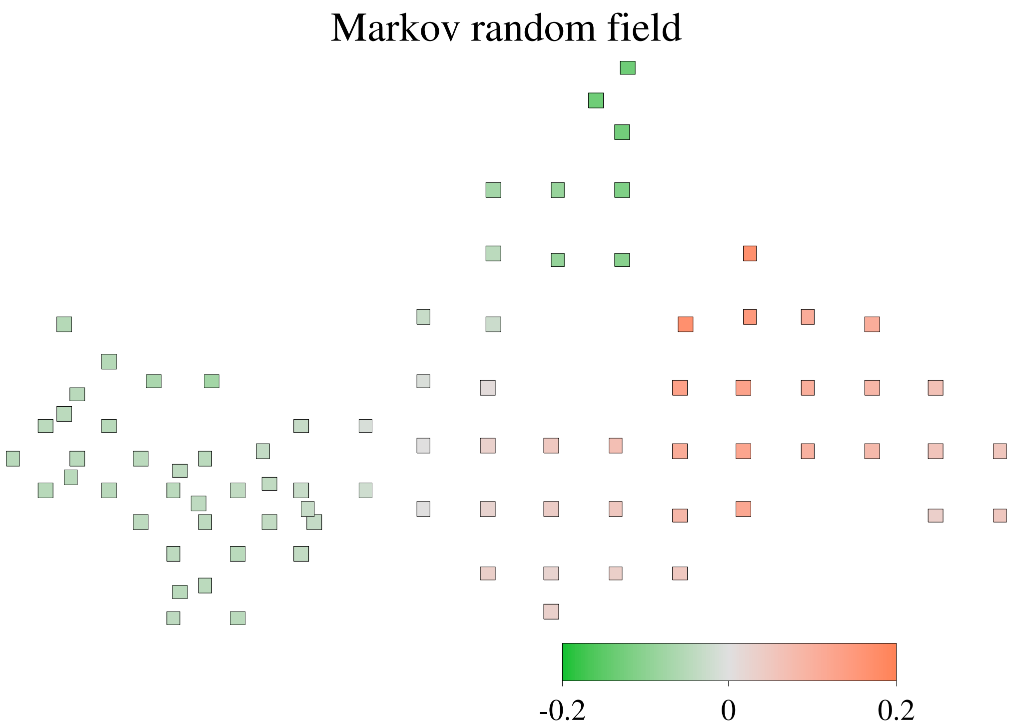
- The precision matrix K acts as a penalty matrix that ensures smoothness of the corresponding estimates.

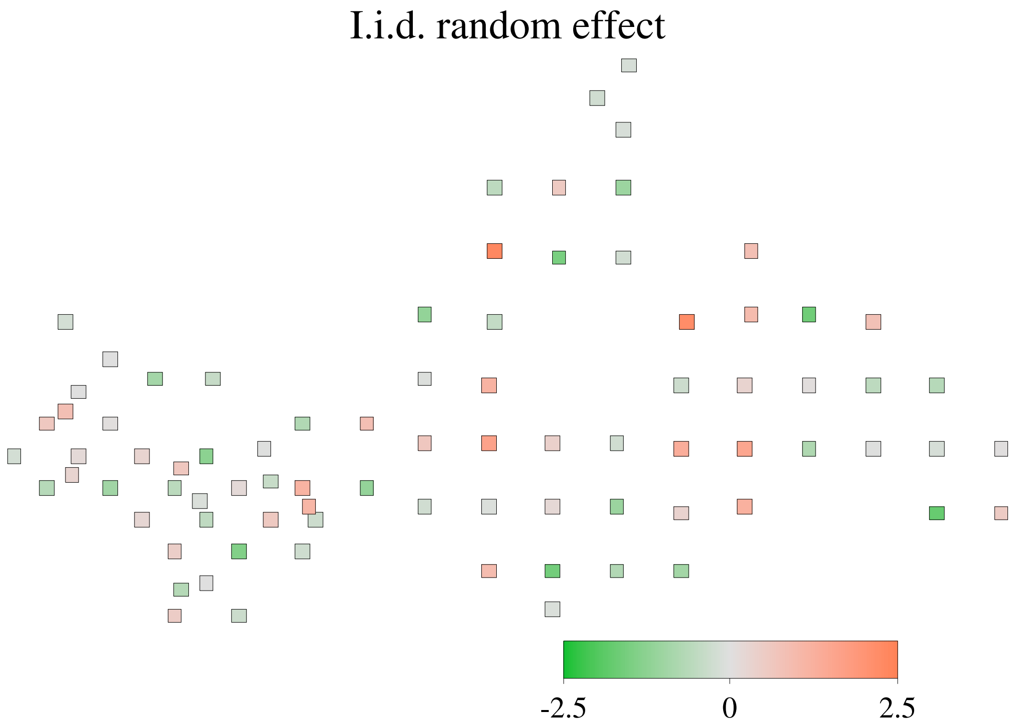
- The variance τ^2 can be interpreted as a smoothing parameter and controls the trade-off between smoothness and fidelity to the data:
  - τ^2 small $\Rightarrow$ smooth estimates.
  - τ^2 large $\Rightarrow$ wiggly estimates.

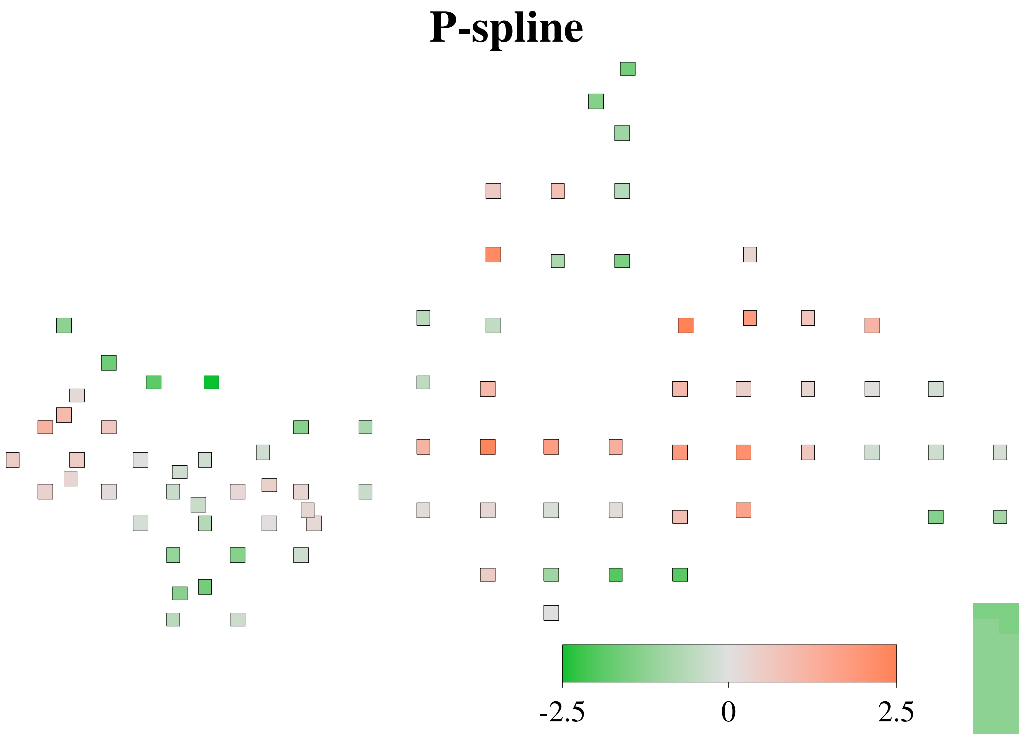
- Fully Bayesian inference:
  - All parameters (including the variance parameters τ^2) are assigned suitable prior distributions.
  - Estimation is based on MCMC simulation techniques.
  - Usual estimates: Posterior expectation, posterior median (easily obtained from the samples).

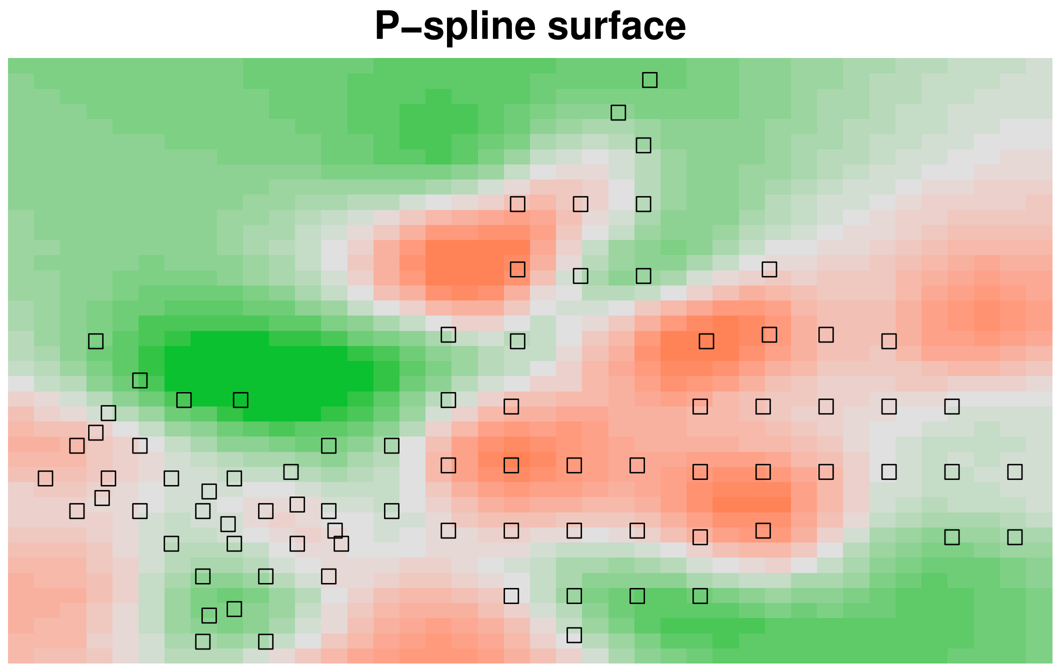
- Empirical Bayes inference:
  - Differentiate between parameters of primary interest (regression coefficients) and hyperparameters (variances).
  - Assign priors only to the former.
  - Estimate the hyperparameters by maximising their marginal posterior.
  - Plugging these estimates into the joint posterior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

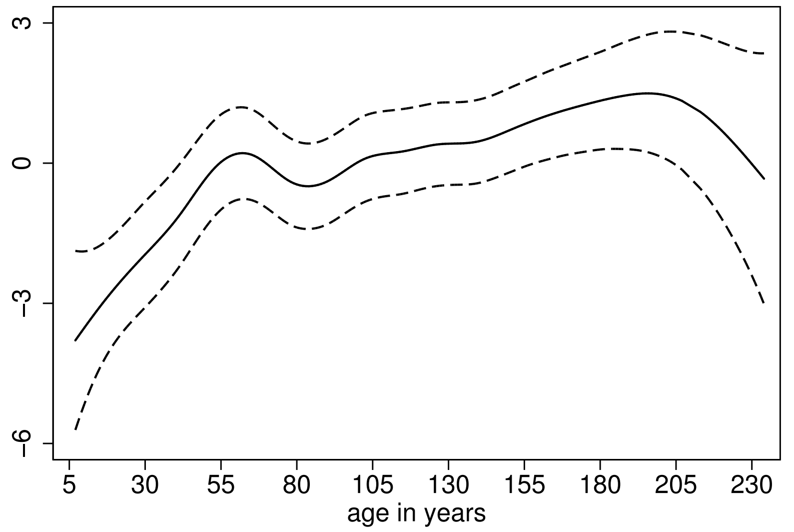
Results

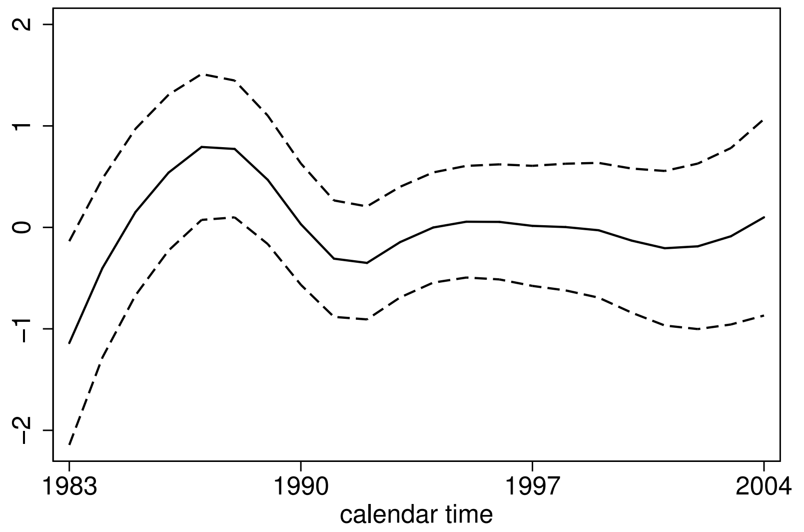


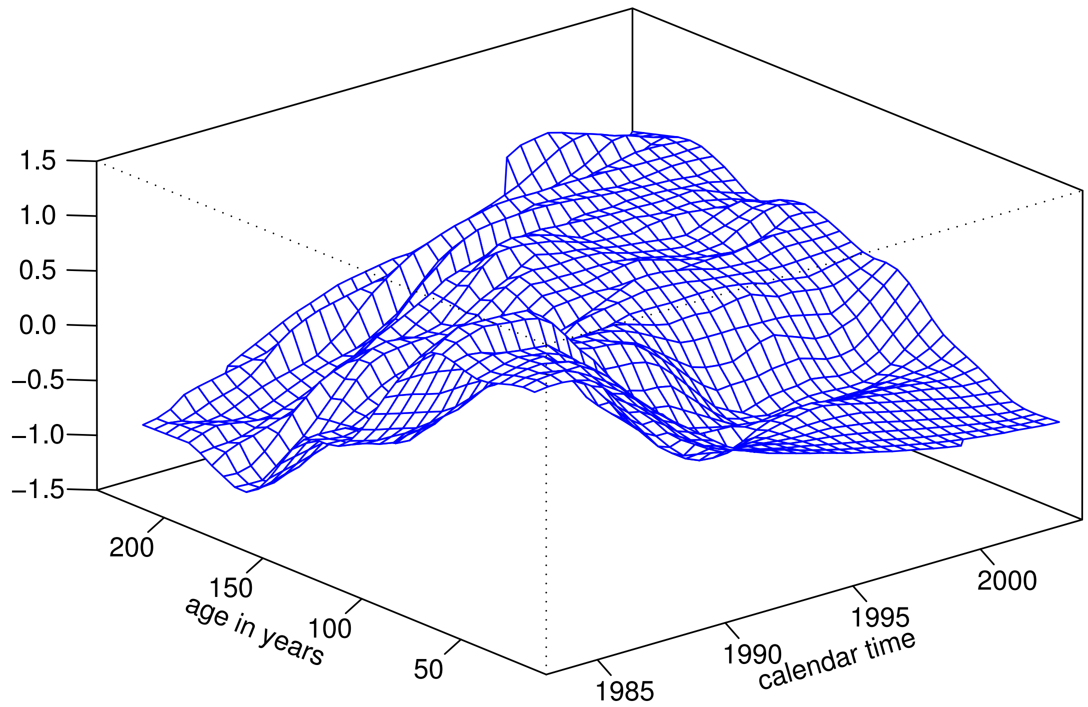


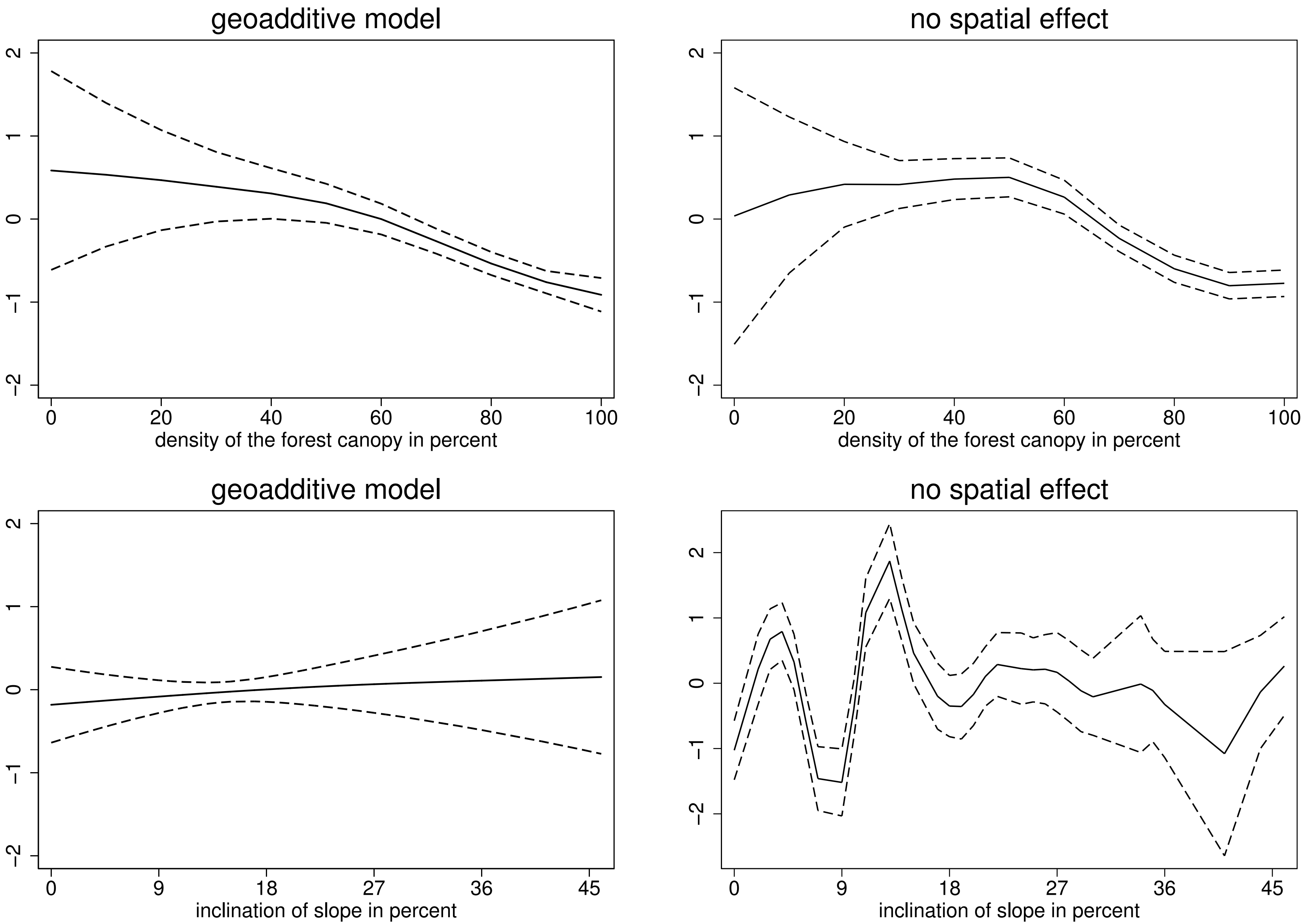

P-spline
-2.5
0
2.5


P−spline surface


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

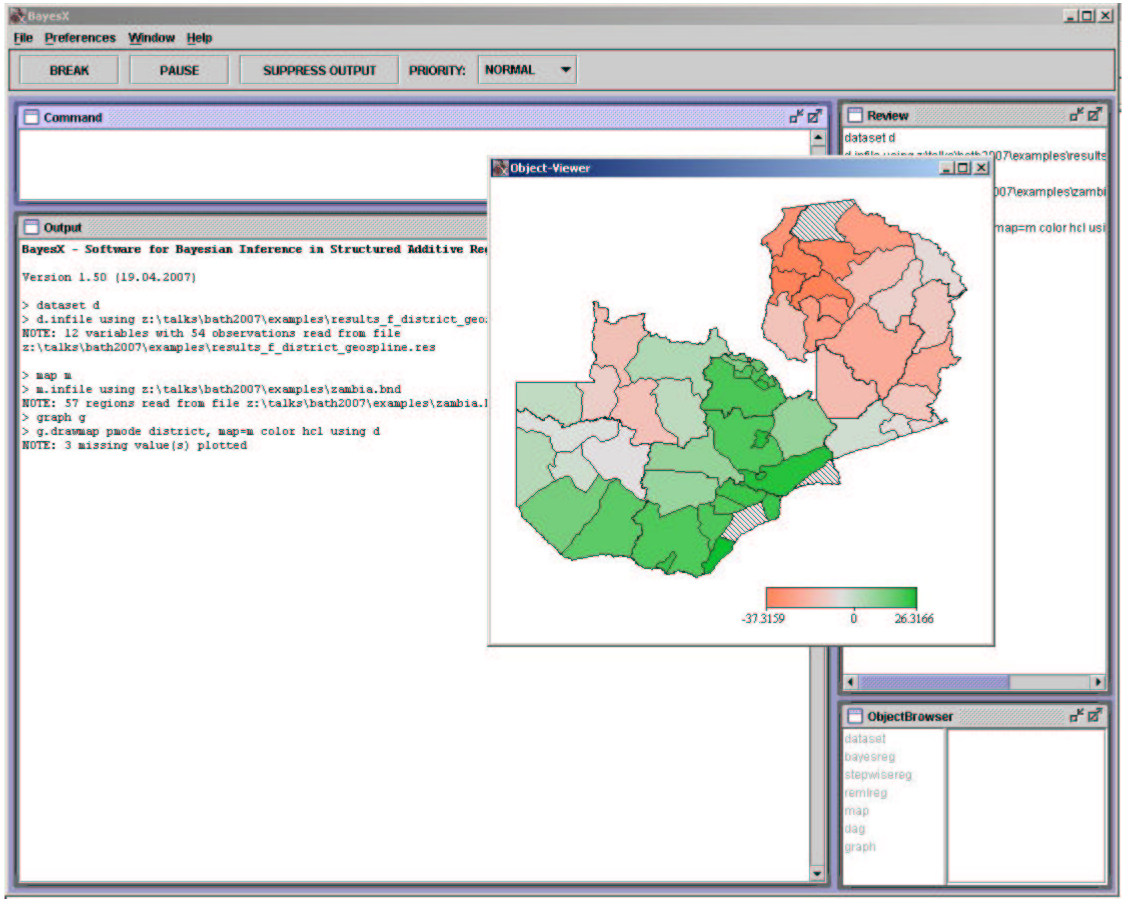

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
200
150
100
50
age in years
1985
1990
1995
2000
calendar time


geoadditive model
no spatial effect
density of the forest canopy in percent
density of the forest canopy in percent
geoadditive model
no spatial effect
inclination of slope in percent
inclination of slope in percent

- Summary:
  - Inclusion of any kind of spatial effect leads to a dramatically improved model fit.
  - The unstructured part dominates the structured spatial effect.
  - Temporal effects are present in the data.
  - Nonparametric effects allow for more realistic models and additional insight.
  - Inclusion of the spatial effect also improved interpretability of other effects.

BayesX

- BayesX is a software tool for estimating geoadditive regression models.



- Stand-alone software with Stata-like syntax.
- Developed by Andreas Brezger, Thomas Kneib and Stefan Lang with contributions of seven colleagues.
- Computationally demanding parts are implemented in C++.
- Graphical user interface and visualisation tools are implemented in Java.
- Currently, BayesX only runs under Windows, a Linux version as well as a connection to R are work in progress.
- More information:

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

- Inferential procedures:
  - Fully Bayesian inference based on MCMC.
  - Empirical Bayes inference based on mixed model methodology.
- Univariate response types:
  - Gaussian,
  - Bernoulli and Binomial,
  - Poisson and zero-inflated Poisson,
  - Gamma,
  - Negative Binomial.

- Categorical responses with ordered categories:
  - Ordinal as well as sequential models,
  - Logit and probit models,
  - Effects can be category-specific or constant over the categories.
- Categorical responses with unordered categories:
  - Multinomial logit and multinomial probit models,
  - Category-specific and globally-defined covariates,
  - Non-availability indicators can be defined to account for varying choice sets.

- Continuous survival times:
  - Cox-type hazard regression models,
  - Joint estimation of baseline hazard rate and covariate effects,
  - Time-varying effects and time-varying covariates,
  - Arbitrary combinations of right, left and interval censoring as well as left truncation.
- Multi-state models:
  - Describe the evolution of discrete phenomena in continuous time,
  - Model in terms of transition intensities, similar as in the Cox model.

Conclusions

- Take home message:

  BayesX is a user-friendly software that allows for the routine estimation of a broad class of geoadditive regression models.

- Geoadditive models can be estimated for various types of responses.

- Fully automated fit without the need for subjective judgements.

- Realistically complex models for complex data.

- Challenging task: Model choice and variable selection in geoadditive regression.

- More on the application:

  Kneib, T. & Fahrmeir, L. (2008): A Space-Time Study on Forest Health. In: Chandler, R. E. & Scott, M. (eds.): Statistical Methods for Trend Detection and Analysis in the Environmental Sciences, Wiley.

- A place called home:

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