Spatially correlated categorical time series: A case study in forest health

Thomas Kneib & Ludwig Fahrmeir
Department of Statistics
Ludwig-Maximilians-University Munich

1. Survey and data
2. Regression models for ordinal responses
3. Geoadditive mixed models
4. Mixed model based inference
5. Software
6. Results

LMU

9.3.2006

SFB 386

Survey and Data

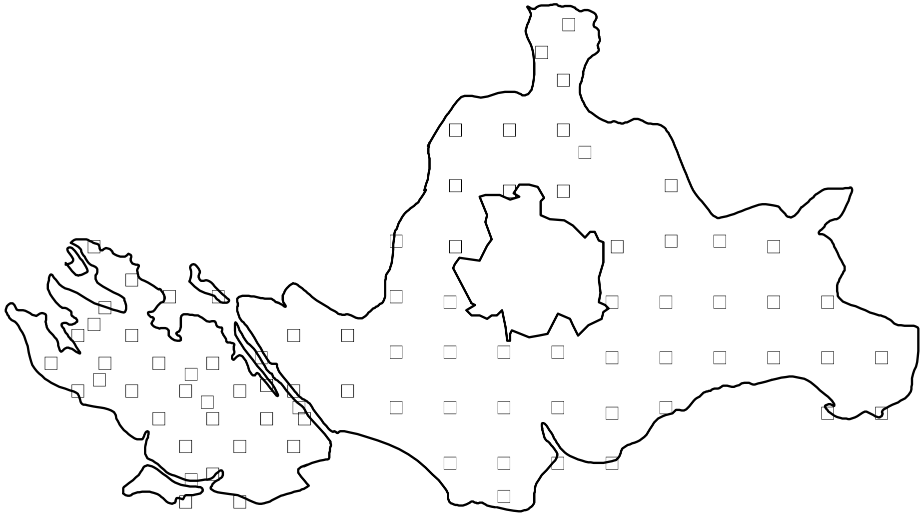
- Aim of the study: Identify factors influencing the health status of trees.

- 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.

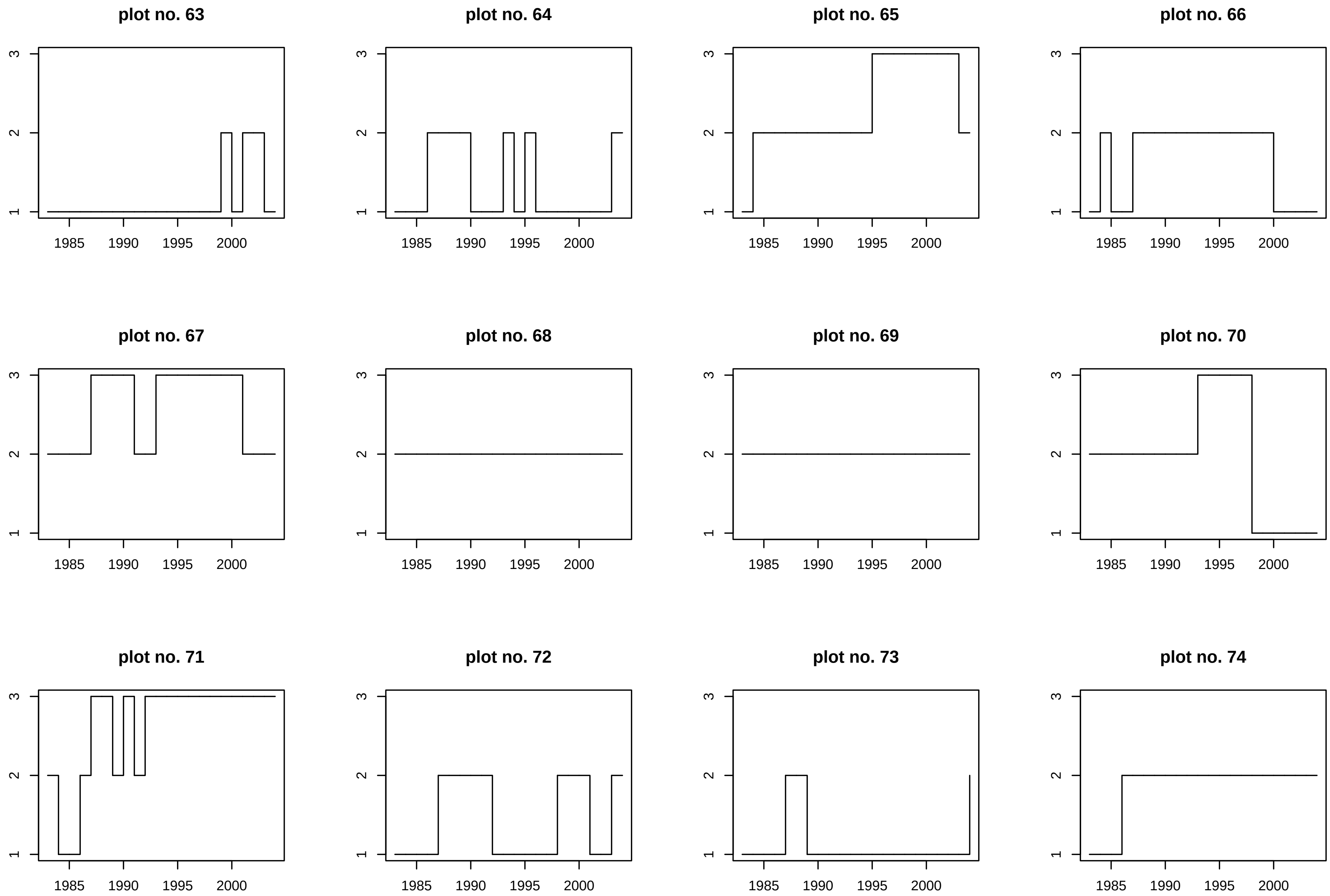
- Response: defoliation degree at plot i in year t, measured in three ordered categories:

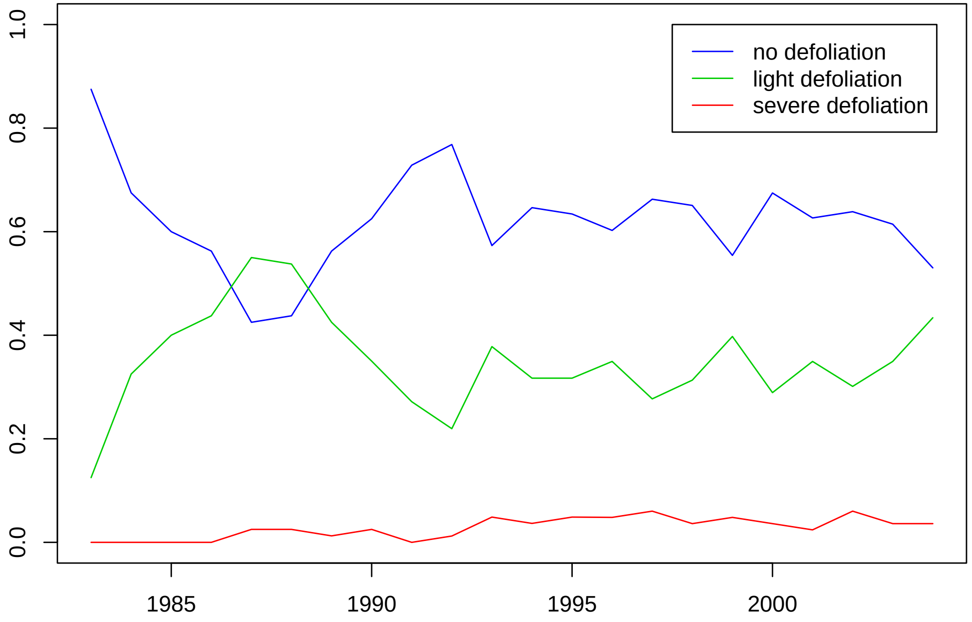
$y_{it} = 1$ no defoliation,
$y_{it} = 2$ defoliation 25% or less,
$y_{it} = 3$ defoliation above 25%.



- Covariates:

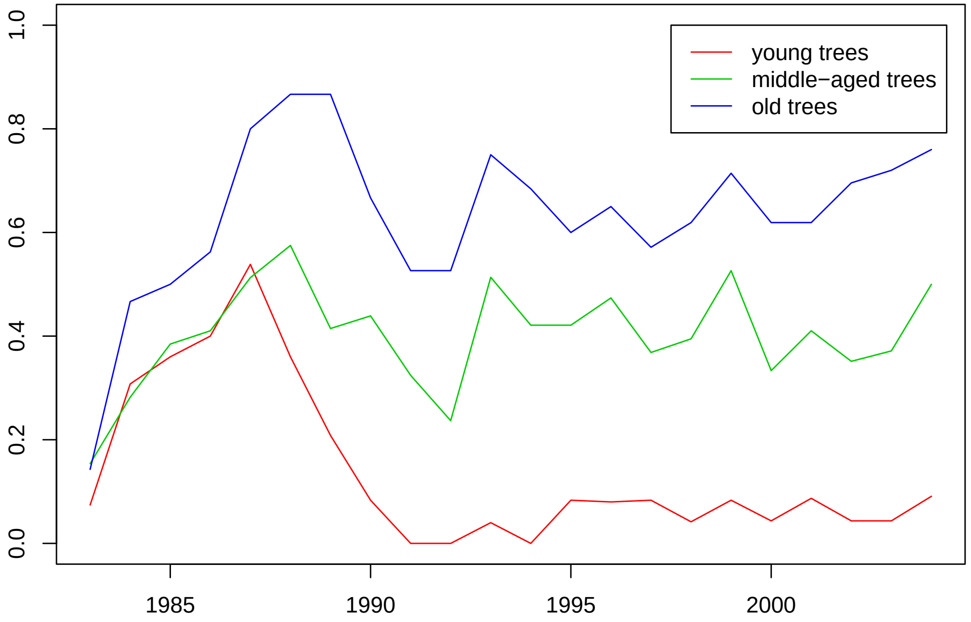
Continuous:	average age of trees at the observation plot
	elevation above sea level in meters
	inclination of slope in percent
	depth of soil layer in centimeters
	pH-value in 0-2cm depth
	density of forest canopy in percent
Categorical	thickness of humus layer in 5 ordered categories
	level of soil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

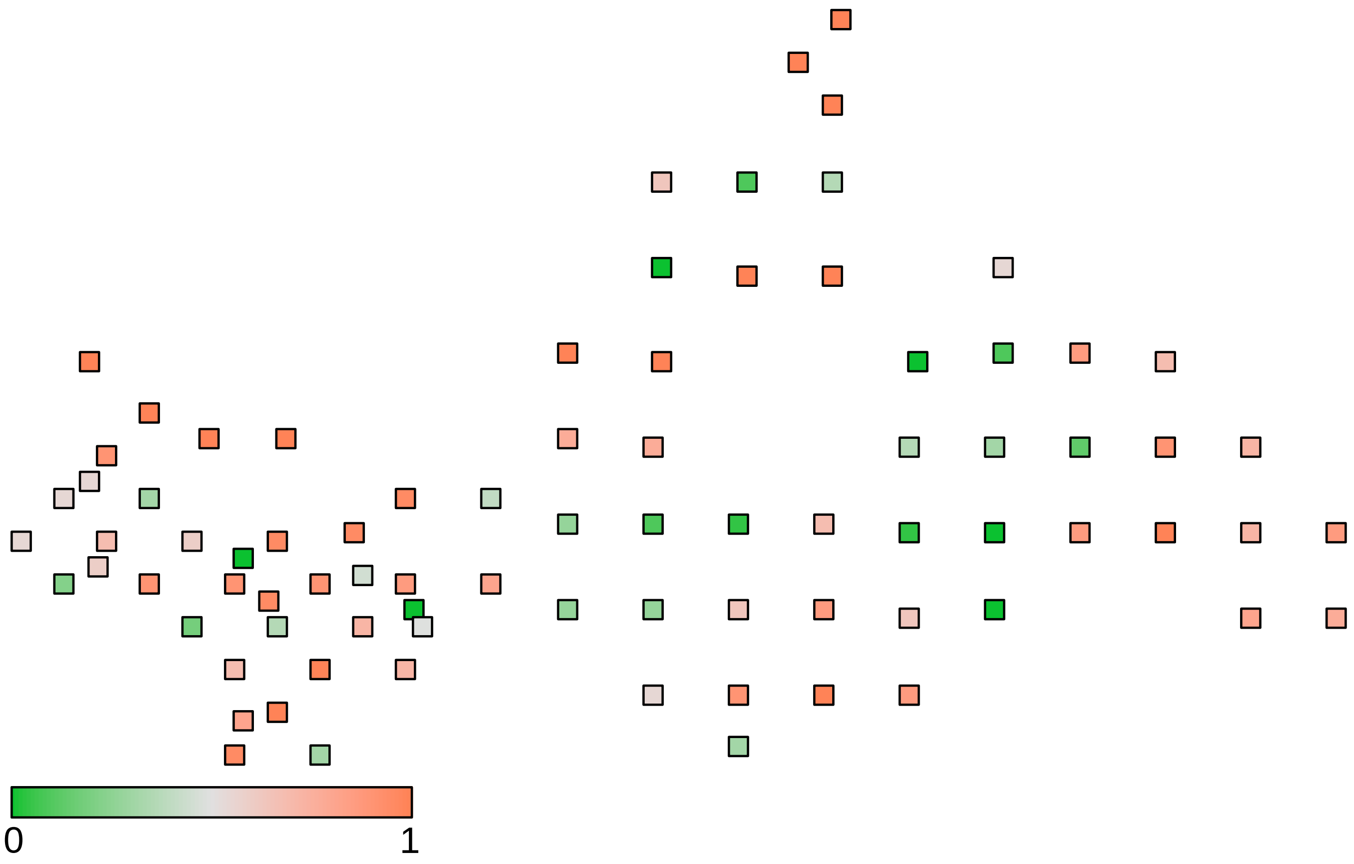

plot no. 63
plot no. 64
plot no. 65
plot no. 66
plot no. 67
plot no. 68
plot no. 69
plot no. 70
plot no. 71
plot no. 72
plot no. 73
plot no. 74
1
2
3
1985
1990
1995
2000



Empirical time trends.

Trends for different ages.




0
1

- We need a 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

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

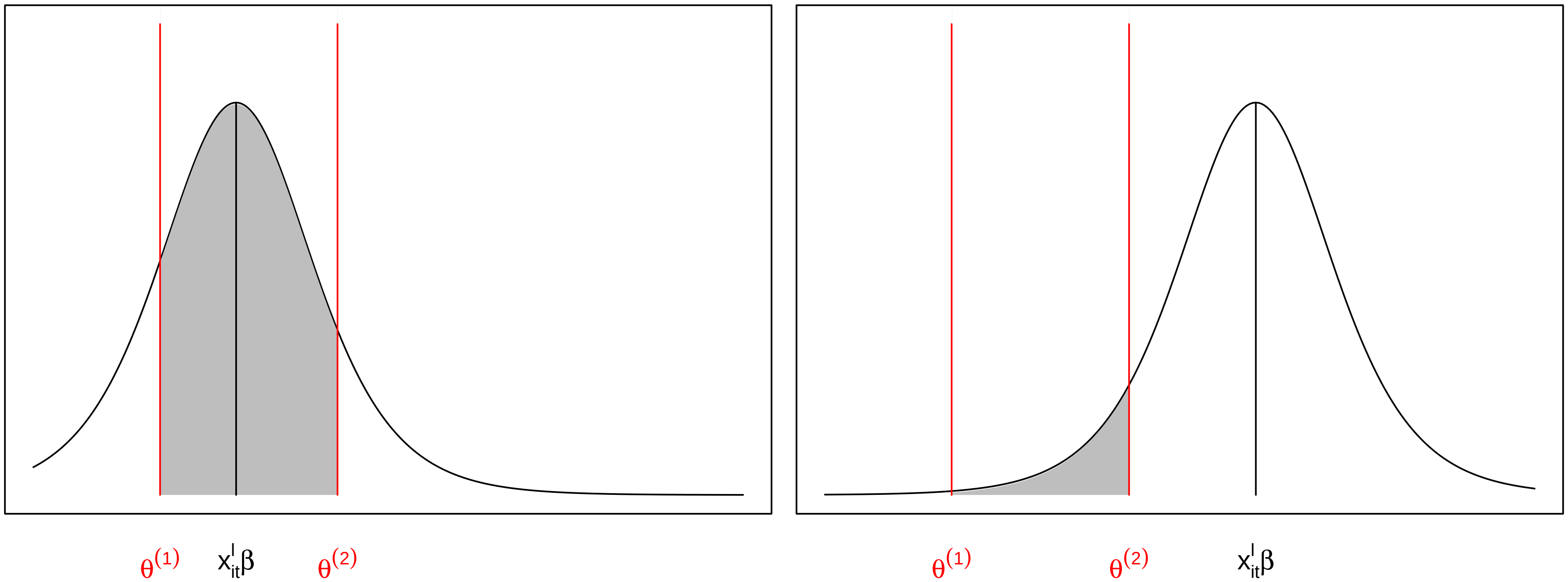
via

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

- 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 $f = F'$.

- Three main concepts to account for the longitudinal structure:
  - Marginal models (define working correlations, short time series),
  - Autoregressive models (include lagged response variables as predictors, prediction),
  - Models with random effects

$$D_{it} = x_{it}'\beta + z_{it}'b_i + \varepsilon_{it}.$$

- In the forest health example:
  - Relatively long series.
  - Interest is on modelling the marginal expectation, not the conditional expectation.
  - In addition: spatial correlations, non-linear trends, further non-linear effects.

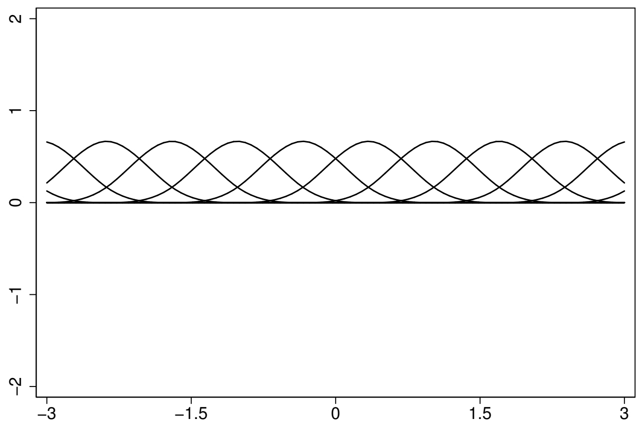
$\Rightarrow$ Extend mixed models to geoadditive mixed models.

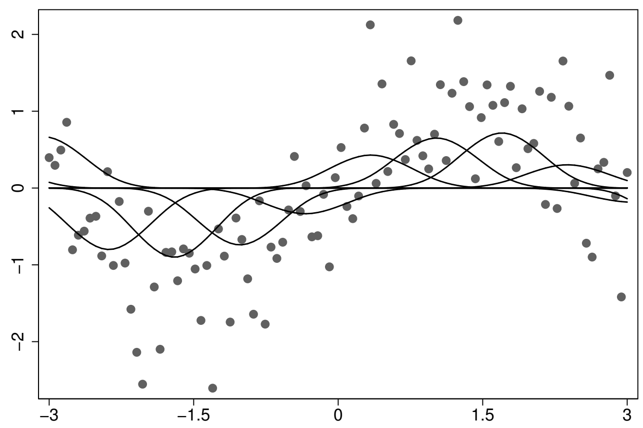
Geoadditive mixed models

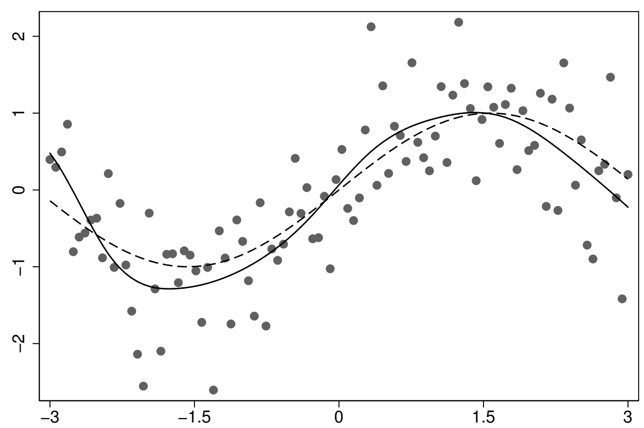
- Suitable model in our application:

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

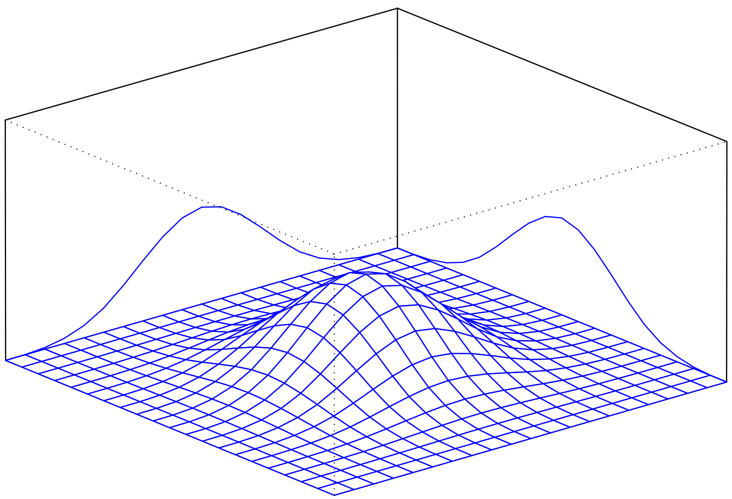
- Penalised splines: Nonlinear covariate effects, nonlinear time trends.
  - Approximate $f(x)$ by a weighted sum of B-spline basis functions.
  - Employ a large number of basis functions to enable flexibility.
  - Penalise differences between parameters of adjacent basis functions to ensure smoothness.

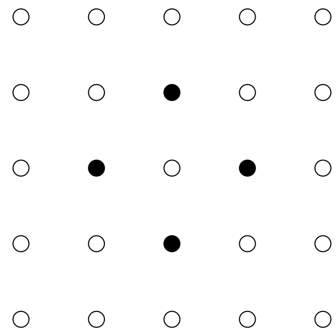




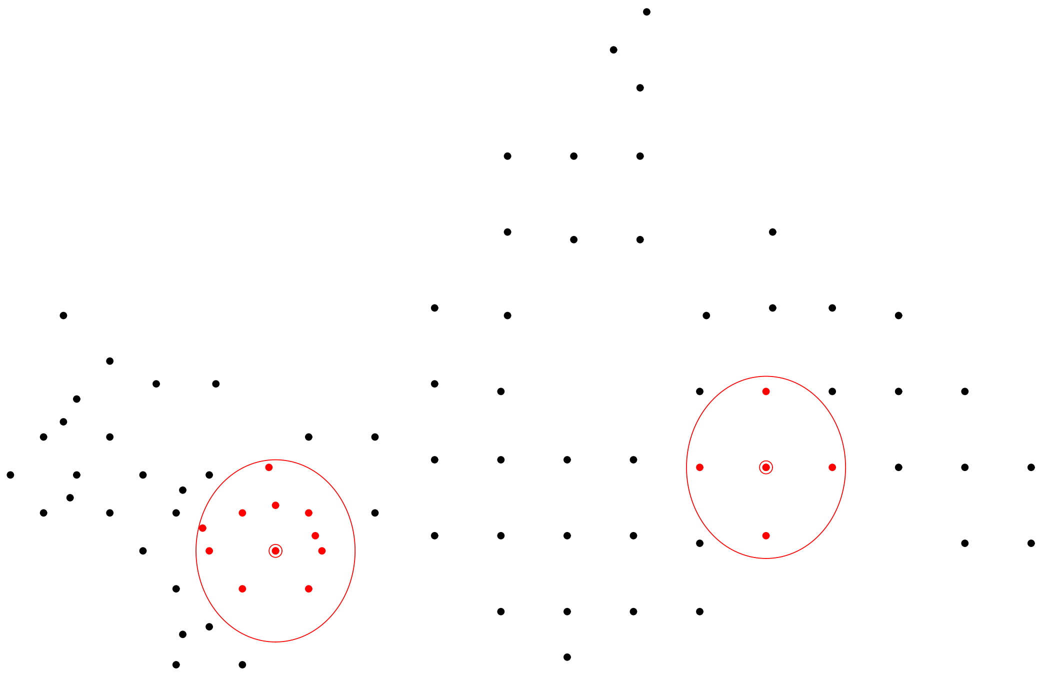


- Bivariate penalised splines: Interaction surfaces, structured spatial effect.
  - Bivariate basis functions based on tensor product B-splines.
  - Extend penalisation to neighbours on a grid.

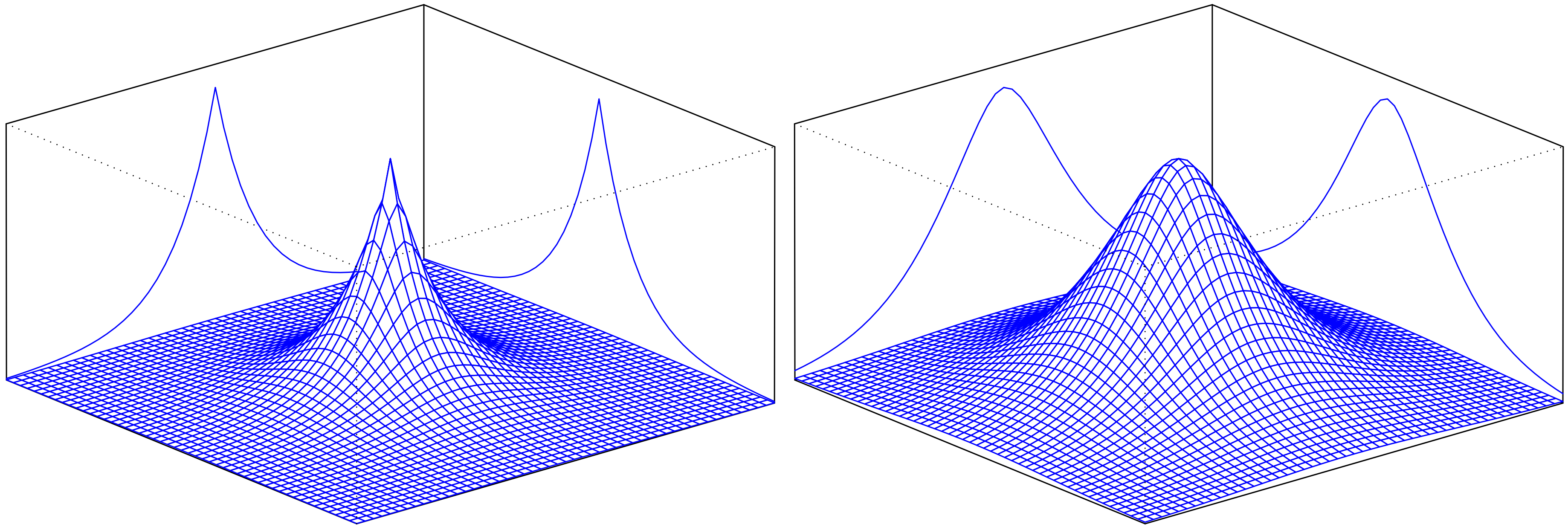




- Markov random fields: Structured spatial effect.
  - Bivariate extension of a first order random walk on the real line.
  - Define two observation plots as neighbours if their distance is less than 1.2km.
  - Assume that the expected value of $f_{spat}(s)$ is the average of the function evaluations of adjacent sites.



- Stationary Gaussian random fields: Structured spatial effect.
  - 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.



Mixed model based inference

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

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

⇒ Reparametrize the model to a proper mixed model.

- Obtain empirical Bayes estimates via iterating
  - Penalized maximum likelihood for regression coefficients.
  - Restricted Maximum / Marginal likelihood for variance parameters.

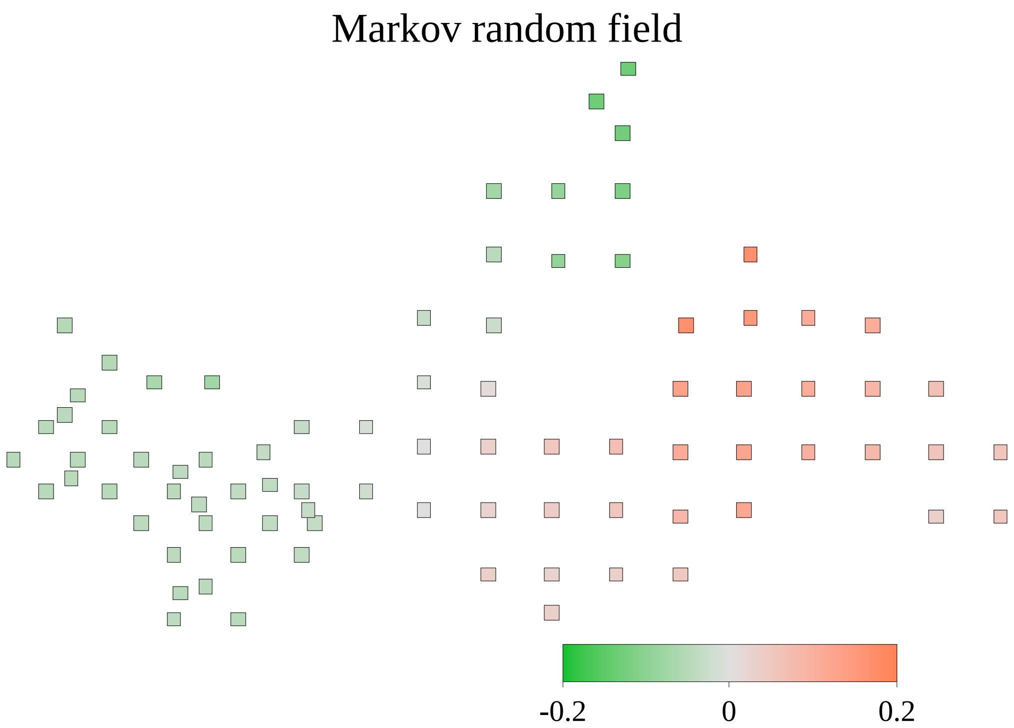
Software

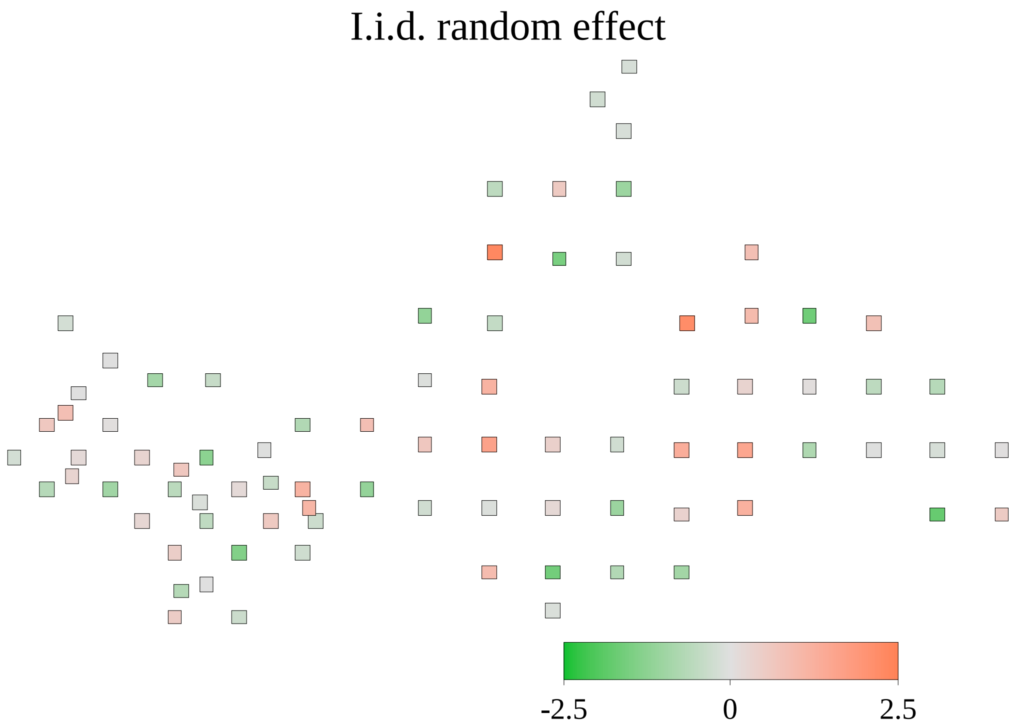
- Implemented in the software package BayesX.

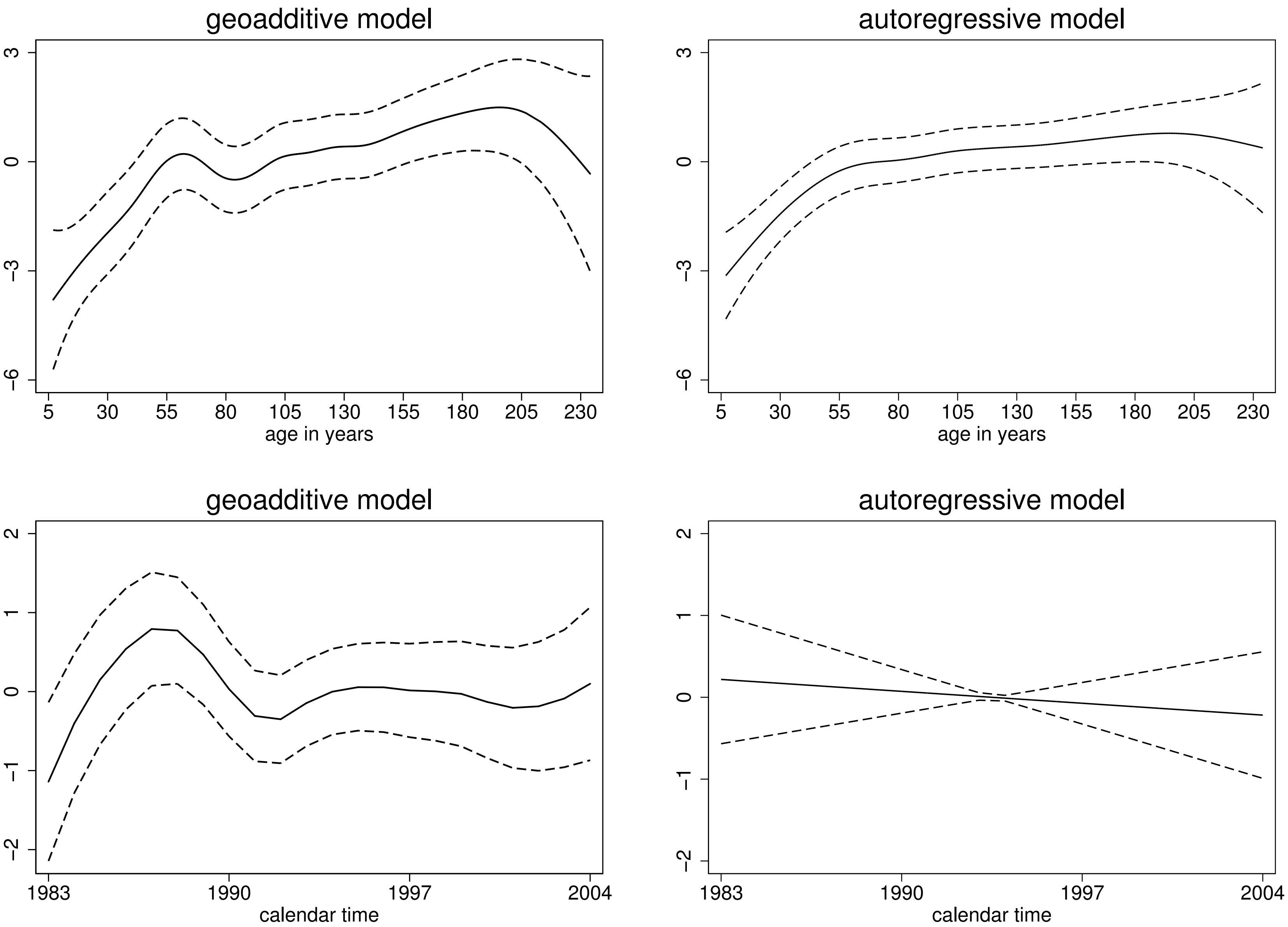
- Available from

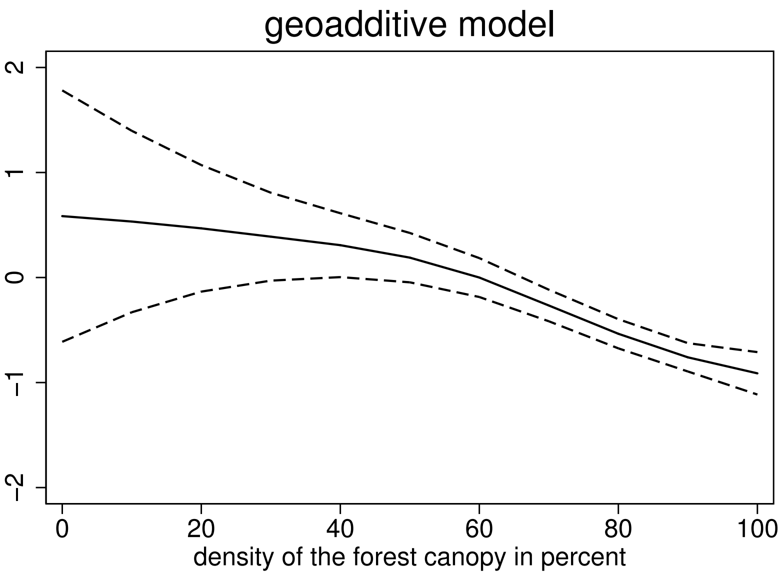
`http://www.stat.uni-muenchen.de/~bayesx`

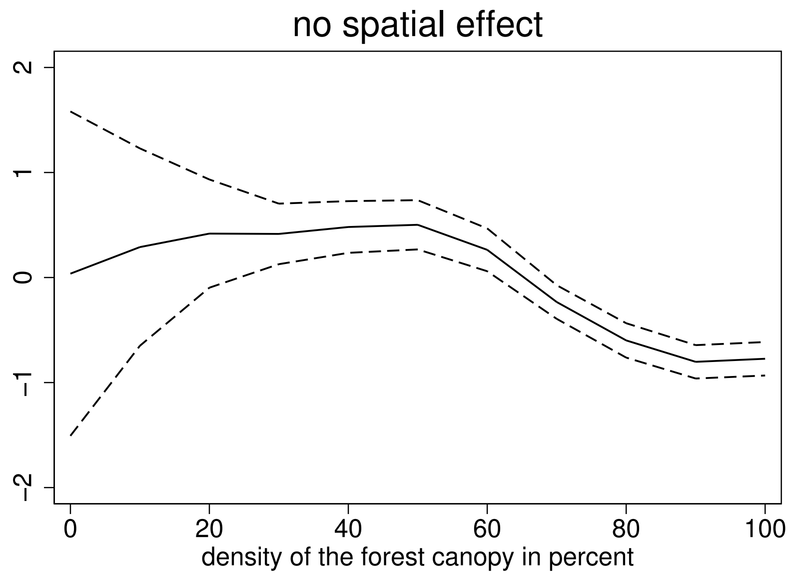
Results

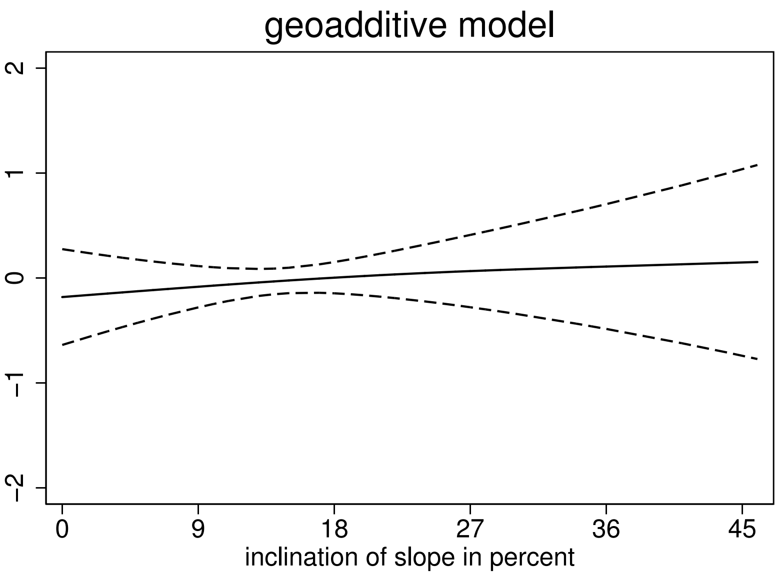


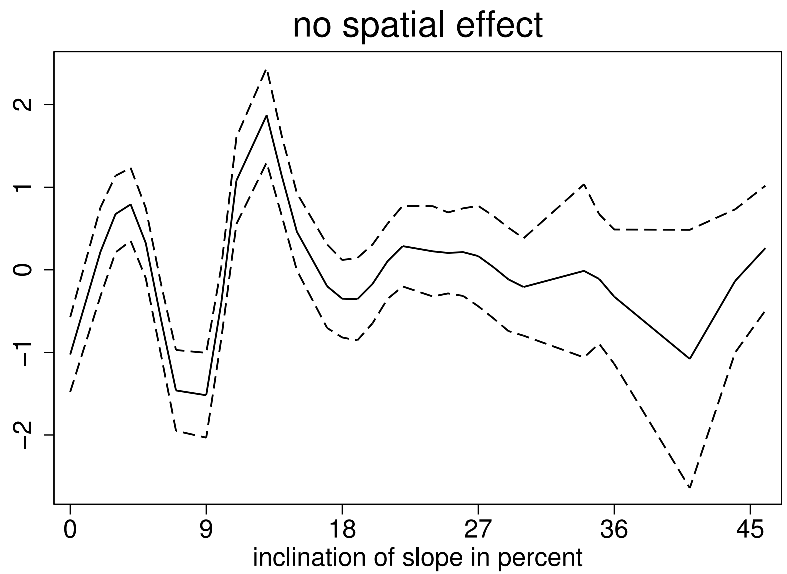


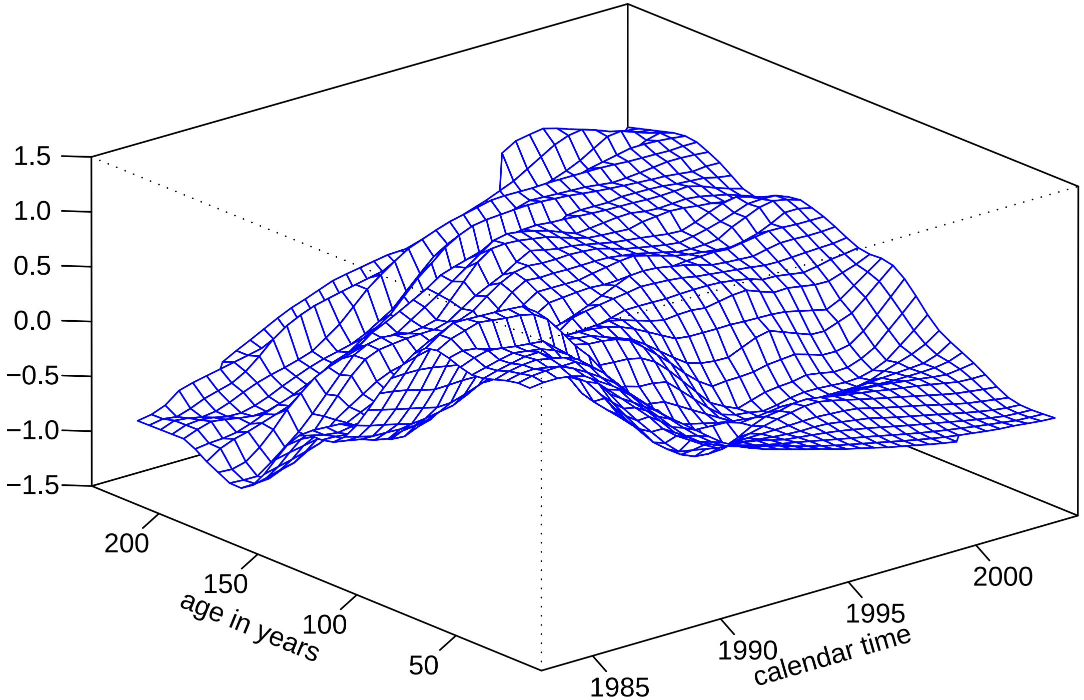

geoadditive model
autoregressive model
3
0
−3
−6
5 30 55 80 105 130 155 180 205 230
age in years
geoadditive model
autoregressive model
2
1
0
−1
−2
1983 1990 1997 2004
calendar time


geoadditive model
−2
−1
0
1
2
0
20
40
60
80
100
density of the forest canopy in percent


no spatial effect
−2
−1
0
1
2
0
20
40
60
80
100
density of the forest canopy in percent

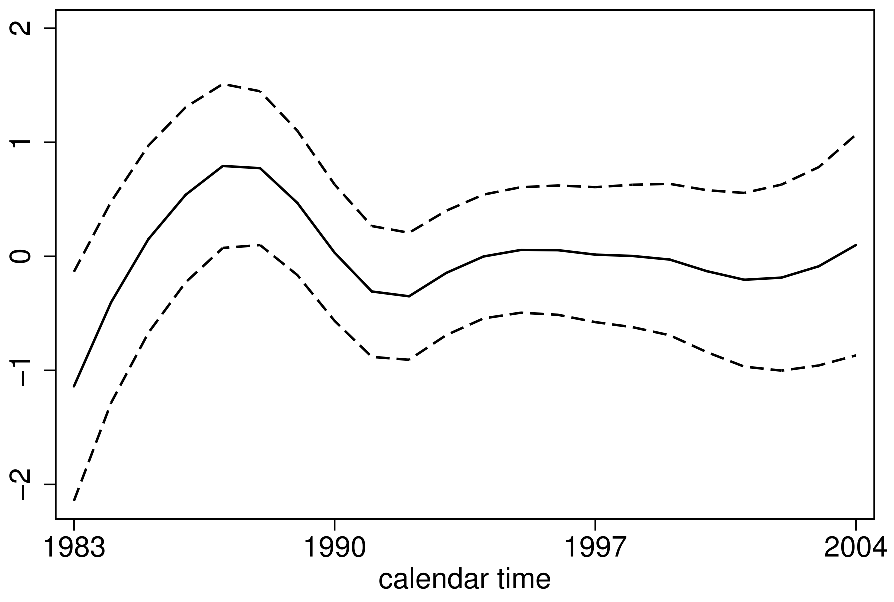

geoadditive model
−2
−1
0
1
2
0
9
18
27
36
45
inclination of slope in percent

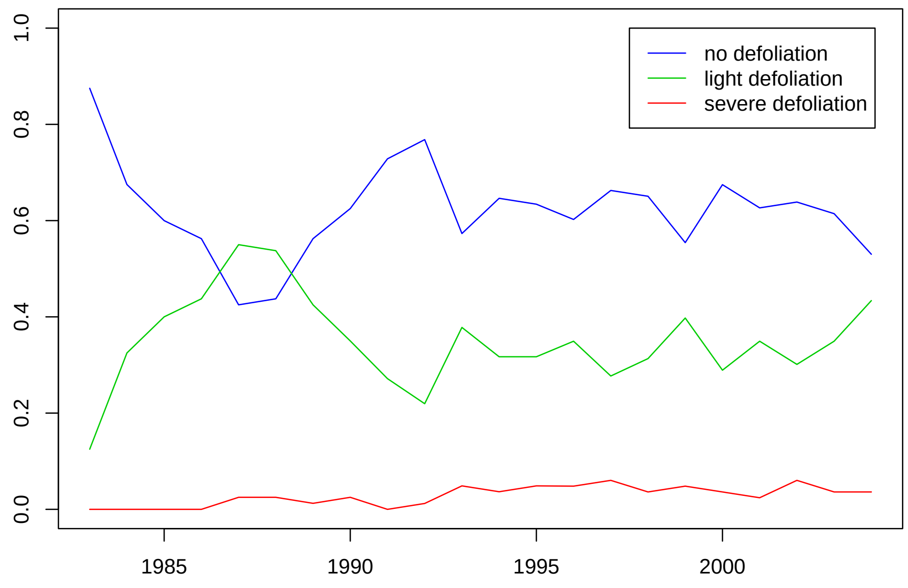

no spatial effect
−2
−1
0
1
2
0
9
18
27
36
45
inclination of slope in percent



variable	$\hat{\beta}_j$	std. dev.	p-value
ph	-0.037	0.212	0.860
humus 0-1cm	-0.261	0.108	0.015
humus 1-2cm	-0.135		
humus 2-3cm	0.139	0.086	0.105
humus 3-4cm	0.135	0.102	0.185
humus >4cm	0.122	0.142	0.391
moderately dry	-0.597	0.320	0.061
moderately moist	0.185		
moist or temporary wet	0.412	0.229	0.071

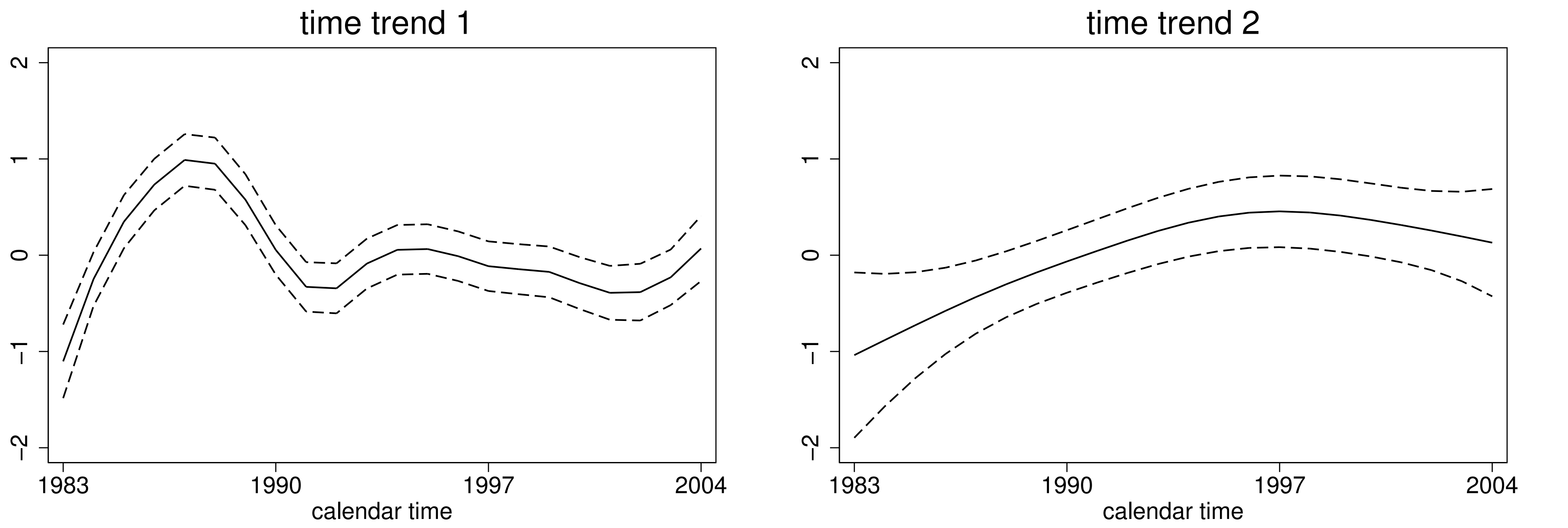
- Limitation of the model: All effects are globally defined





- Possible refinement: Category-specific trends

$$P(Y_{it} \leq r) = \Phi\left[\theta^{(r)} - \ldots - f_{time}^{(r)}(t) - \ldots\right]$$



- More complicated constraints:

$$\theta^{(1)} - f^{(1)}_{time}(t) < \theta^{(2)} - f^{(2)}_{time}(t) \qquad \text{for all } t.$$

Conclusions

- Inclusion of any kind of spatial effect leads to a dramatically improved model fit.
- The unstructured part dominates the structured spatial effect.
- Nonparametric effects allow for more realistic models.
- Category-specific effects give additional insight but may require a larger database.

References

- Kneib, T. & Fahrmeir, L. (2006): Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, to appear.

- Kneib, T. & Fahrmeir, L. (2007): 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`