Model Choice and Variable Selection in Geoadditive Regression Models

Thomas Kneib

Department of Statistics
Ludwig-Maximilians-University Munich

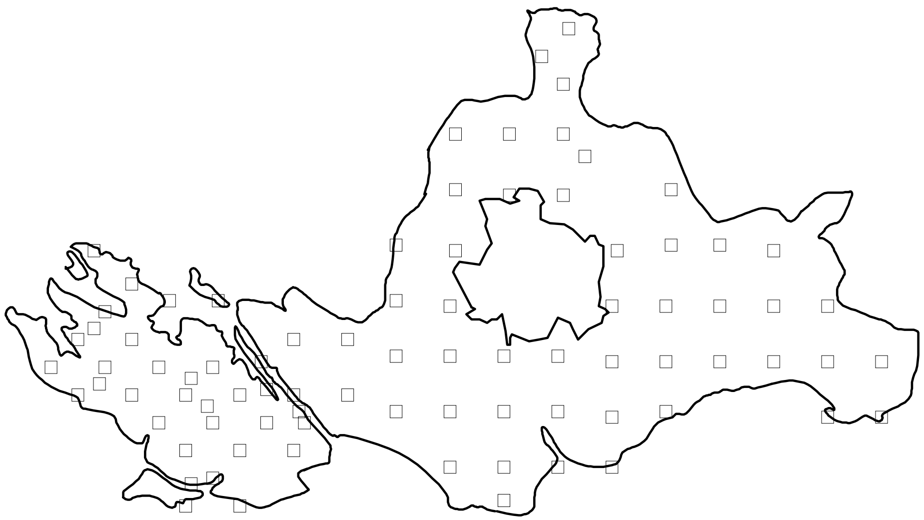
joint work with

Torsten Hothorn Gerhard Tutz

12.3.2008

Geoadditive Regression: Forest Health Example

- 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: binary defoliation indicator y_{it} of plot i in year t ($1 =$ defoliation higher than 25%).
- Spatially structured longitudinal data.



- Covariates:

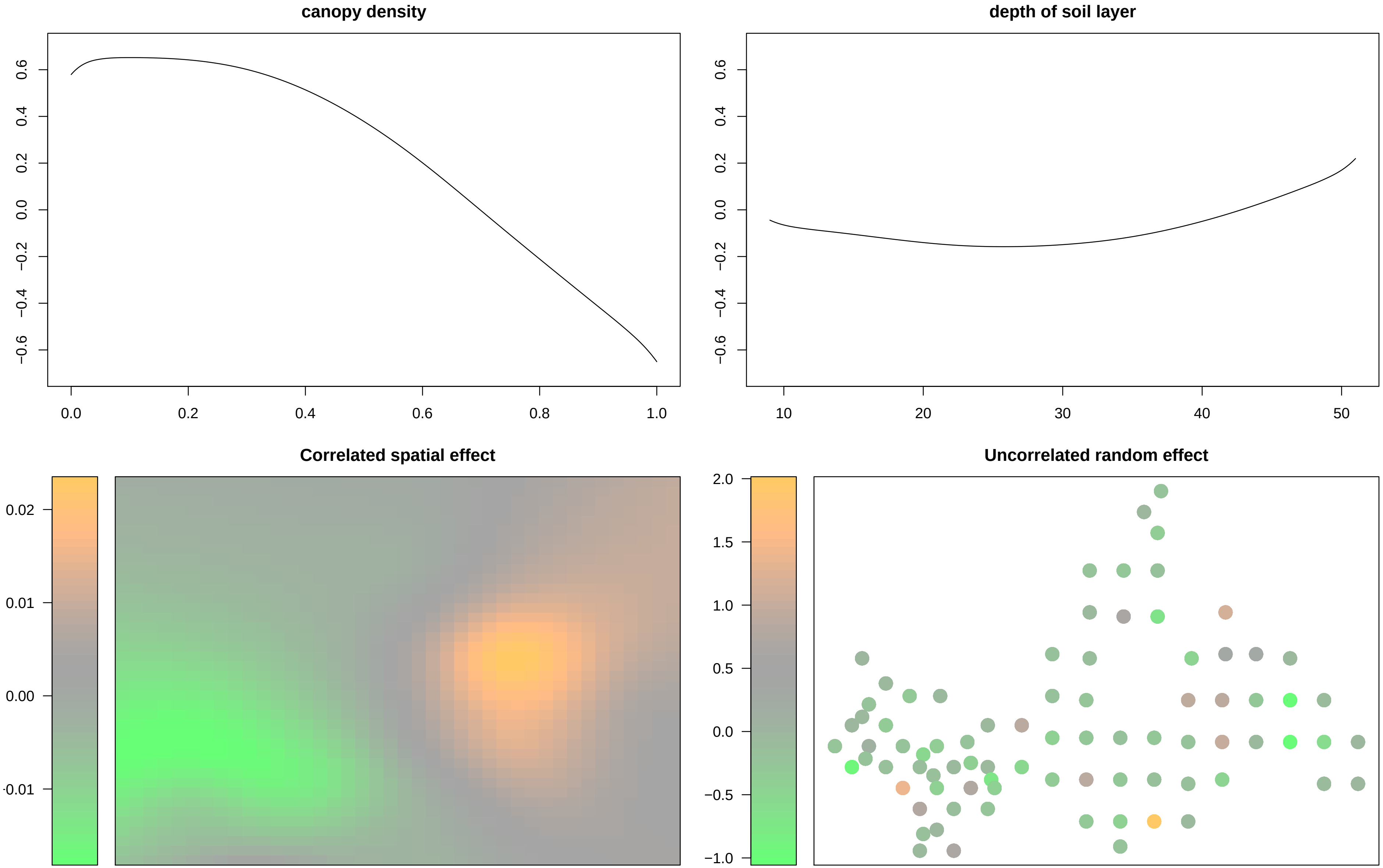
Continuous:	average age of trees at the observation plot
	elevation above sea level in meters
	inclination of slope in percent
	depth of soil layer in centimeters
	pH-value in 0 – 2cm depth
	density of forest canopy in percent
Categorical	thickness of humus layer in 5 ordered categories
	level of soil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

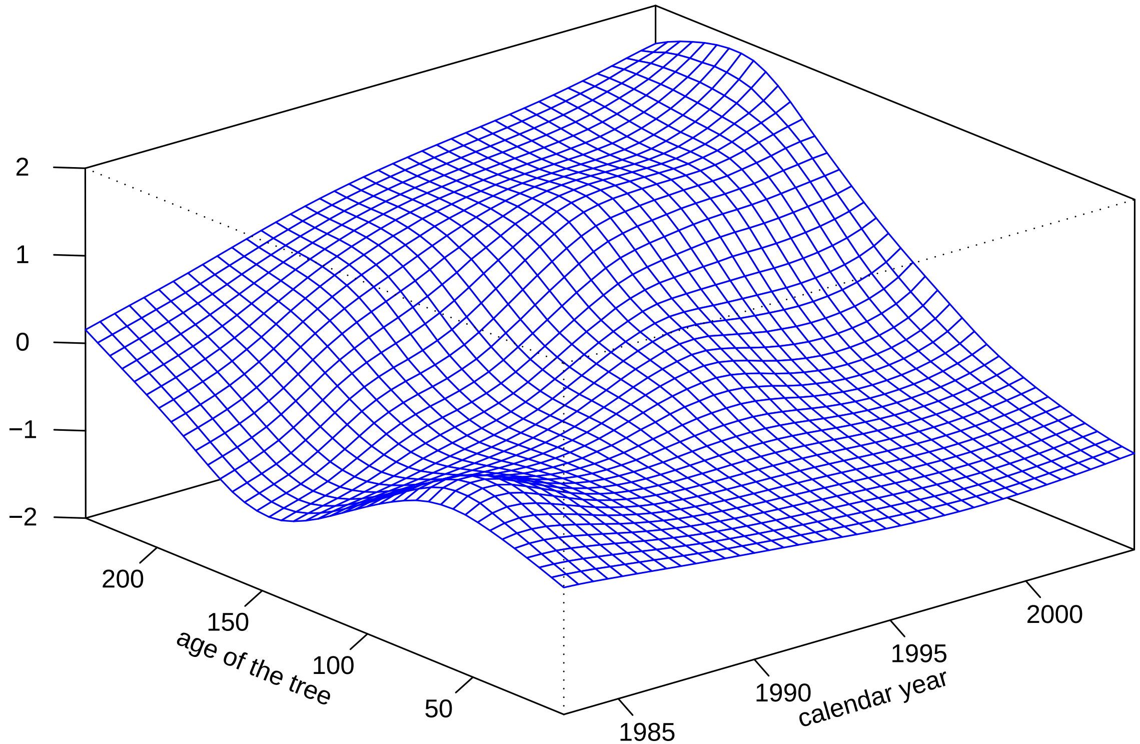
- Possible model:

$$P(y_{it} = 1) = \frac{\exp(\eta_{it})}{1 + \exp(\eta_{it})}$$

where η_{it} is a geoadditive predictor of the form

η_{it}	$=$	$f_1(\text{age}_{it}, t)+$	interaction between age and calendar time.
		$f_2(\text{canopy}_{it})+$	smooth effects of the canopy density and
		$f_3(\text{soil}_{it})+$	the depth of the soil layer.
		$f_{\text{spat}}(s_{ix}, s_{iy})+$	structured and
		b_i+	unstructured spatial random effects.
		$x'_{it}\beta$	parametric effects of type of stand, fertilisation, thickness of humus layer, level of soil moisture and base saturation.


canopy density
depth of soil layer
Correlated spatial effect
Uncorrelated random effect
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 40 50
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6
0.02 0.01 0.00 -0.01
2.0 1.5 1.0 0.5 0.0 −0.5 −1.0



- Questions:
  - How do we estimate the model? ⇒ Inference.
  - How do we come up with the model specification? ⇒ Model choice and variable selection.

⇒ Componentwise boosting for geoadditive regression models.

Base-Learners For Geoadditive Regression Models

- Componentwise base-learning procedures for geoadditive regression models can be derived from univariate Gaussian smoothing approaches such as

$$y = g(x) + \varepsilon \qquad \text{smooth nonparametric effect}$$

$$y = g(x_1, x_2) + \varepsilon \qquad \text{smooth surface / spatial effect}$$

$$y = x_1 g(x_2) + \varepsilon \qquad \text{varying coefficients}$$

where $\varepsilon \sim N(0, \sigma^2 I)$.

- All base-learners will be given by penalised least squares (PLS) fits

$$\hat{y} = X(X'X + \lambda K)^{-1} X'y$$

characterised by the hat matrix

$$S_\lambda = X(X'X + \lambda K)^{-1} X'.$$

- Recall univariate penalised spline smoothing: Approximate $g(x)$ by a linear combination of B-spline basis functions, i.e.

$$g(x) = \sum_j \beta_j B_j(x)$$

and define a difference penalty

$$\text{pen}(\beta) = \lambda \sum_j (\beta_j - \beta_{j-1})^2 \quad \text{or} \quad \text{pen}(\beta) = \lambda \sum_j (\beta_j - 2\beta_{j-1} + \beta_{j-2})^2.$$

to ensure smoothness.

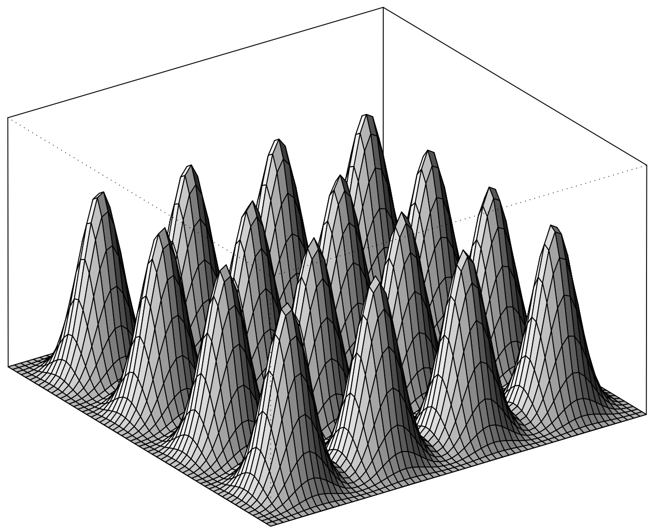
- Model and penalty in matrix notation:

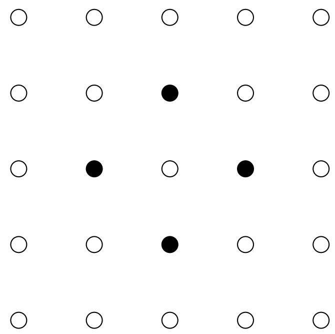
$$y = X\beta + \varepsilon \qquad \text{and} \qquad \text{pen}(\beta) = \lambda \beta' K \beta.$$

- Penalised least squares estimate and fit:

$$\hat{\beta} = (X'X + \lambda K)^{-1} X'y \qquad \hat{y} = X(X'X + \lambda K)^{-1} X'y.$$

- PLS base-learner for interaction surfaces and spatial effects $f(x_1, x_2)$:





- Define bivariate Tensor product basis functions

$$B_{jk}(x_1, x_2) = B_j(x_1)B_k(x_2).$$

- Based on penalty matrices K_1 and K_2 for univariate fits define an overall penalty as

$$\text{pen}(\beta) = \lambda\beta' \underbrace{(I \otimes K_1 + K_2 \otimes I)}_{=K} \beta.$$

- PLS base-learner for varying coefficient terms

$$y = x_1 g(x_2) + \varepsilon$$

Representing $g(x_2)$ as a penalised spline yields $y = X\beta + \varepsilon$, where

$$X = \text{diag}(x_{11}, \ldots, x_{n1})X^*$$

and X^* is the design matrix corresponding to $g(x_2)$.

- PLS base-learners can also be defined for
  - Random intercepts and random slopes,
  - Space-varying effects.

Complexity Adjustment

- The flexibility of penalised least squares base-learners depends on the choice of the smoothing parameter.

- Typical strategy: fix the smoothing parameter at a large pre-specified value.

- Difficult when comparing fixed effects, nonparametric effects and spatial effects.

  $\Rightarrow$ More flexible base-learners will be preferred in the boosting iterations leading to potential selection and estimation bias.

- We need an intuitive measure of complexity.

- Effective degrees of freedom of a penalised least-squares base-learner:

$$\mathrm{df}(\lambda) = \mathrm{trace}(X(X'X + \lambda K)^{-1}X').$$

- Choose the smoothing parameters for the base-learners such that

$$\mathrm{df}(\lambda) = 1.$$

- Can not be achieved for most base-learners since

$$\lim_{\lambda \to \infty} \mathrm{df}(\lambda) \geq 1.$$

- For example, a polynomial of order $k-1$ remains unpenalised for penalised splines with k-th order difference penalty.

- A reparameterisation has to be applied, leading for example to

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_{k-1} x^{k-1} + f_{\text{centered}}(x).$$

- Assign separate base-learners to the parametric components and a one degree of freedom PLS base-learner to the centered effect.

Boosting Geoadditive Regression Models

- Generic representation of geoadditive models:

$$\eta(\cdot) = \beta_0 + \sum_{j=1}^{r} f_j(\cdot)$$

  where the functions $f_j(\cdot)$ represent the candidate functions of the predictor.

- Each candidate function is associated with a PLS base-learner.
- Early stopping of the boosting algorithm implements variables selection.
- Defining concurring base-learners implements model choice (for example linear vs. nonlinear modelling).
- The number of boosting iterations can be determined based on AIC reduction or cross-validation.

Habitat Suitability Analyses

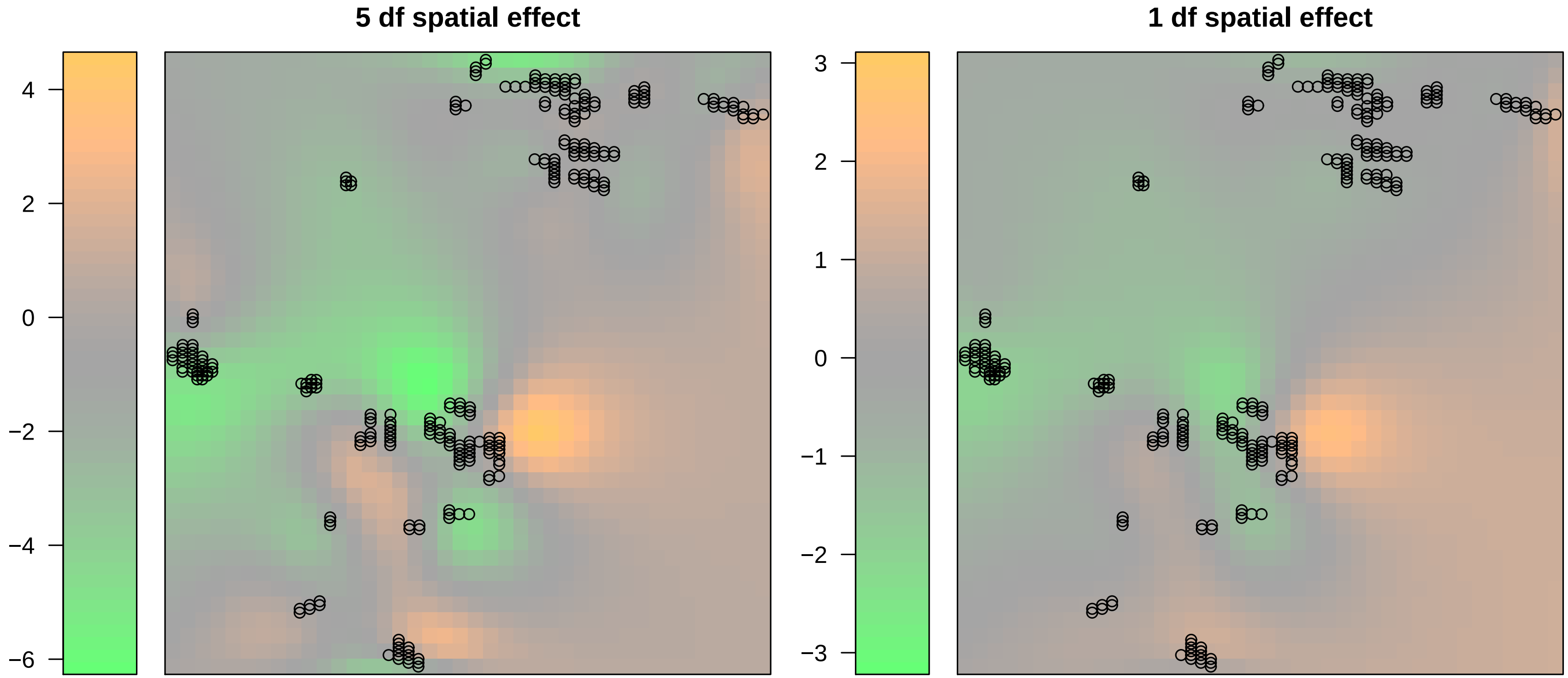
- Identify factors influencing habitat suitability for breeding bird communities collected in seven structural guilds (SG).
- Variable of interest: Counts of subjects from a specific structural guild collected at 258 observation plots in a Northern Bavarian forest district.
- Research questions:
  a) Which covariates influence habitat suitability (31 covariates in total)? Does spatial correlation have an impact on variable selection?
  b) Are there nonlinear effects of some of the covariates?
  c) Are effects varying spatially?
- All questions can be addressed with the boosting approach (but we focus on a)).

- Selection frequencies in a spatial Poisson-GLM:

	GST	DBH	AOT	AFS	DWC	LOG	SNA	COO
non-spatial GLM	0	0	0	0.06	0.3	0	0.01	0
spatial with 5 df	0	0.02	0	0.01	0.05	0	0.01	0
spatial with 1 df	0	0	0	0.06	0.15	0	0	0
	COM	CRS	HRS	OAK	COT	PIO	ALA	MAT
non-spatial GLM	0.03	0.04	0.03	0.05	0.06	0	0.04	0.06
spatial with 5 df	0	0.01	0	0	0	0	0.01	0.05
spatial with 1 df	0.03	0.02	0.02	0.04	0.05	0	0.03	0.04
	GAP	AGR	ROA	LCA	SCA	HOT	CTR	RLL
non-spatial GLM	0.03	0	0	0.1	0.07	0	0	0
spatial with 5 df	0.01	0	0.01	0.01	0.01	0	0	0
spatial with 1 df	0.03	0	0	0.07	0.06	0	0	0
	BOL	MSP	MDT	MAD	COL	AGL	SUL	spatial
non-spatial GLM	0	0.06	0	0	0.05	0	0	0
spatial with 5 df	0	0	0	0	0.03	0	0	0.76
spatial with 1 df	0	0.04	0	0	0.04	0	0	0.3

- A similar picture is obtained from considering the estimated regression coefficients.

- Spatial effects for high and low degrees of freedom:



- Spatial correlation has non-negligible influence on variable selection.

- Making terms comparable in terms of complexity is essential to obtain valid results.

Summary & Extensions

- Generic boosting algorithm for model choice and variable selection in geoadditive regression models.
- Avoid selection bias by careful parameterisation.
- Implemented in the R-package **mboost**.
- Future plans:
  - Derive base-learning procedures for other types of spatial effects (regional data, anisotropic spatial effects).
  - Construct spatio-temporal base-learners based on tensor product approaches.

- Reference: Kneib, T., Hothorn, T. and Tutz, G.: Model Choice and Variable Selection in Geoadditive Regression. Under revision for *Biometrics*.

- Find out more:

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