Boosting Geoadditive Regression Models

Thomas Kneib

Department of Statistics
Ludwig-Maximilians-University Munich

joint work with

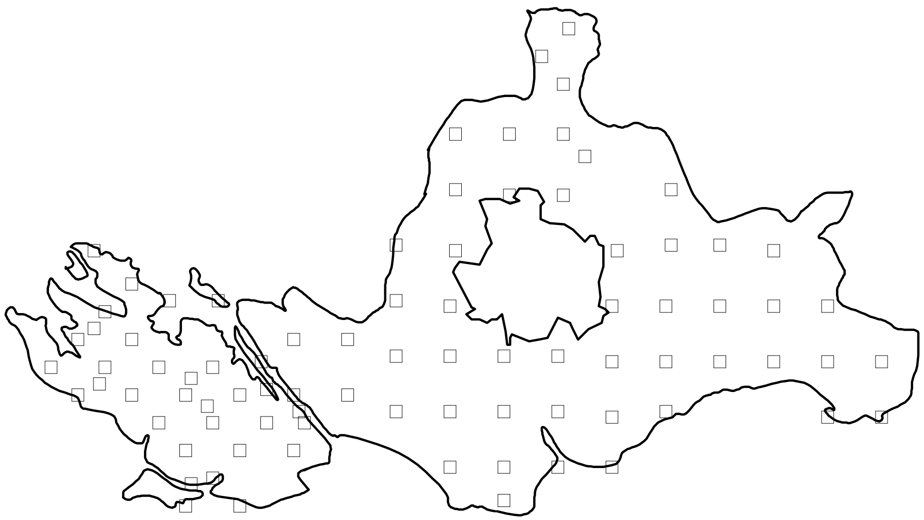
Torsten Hothorn Gerhard Tutz

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

8.5.2008

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**:

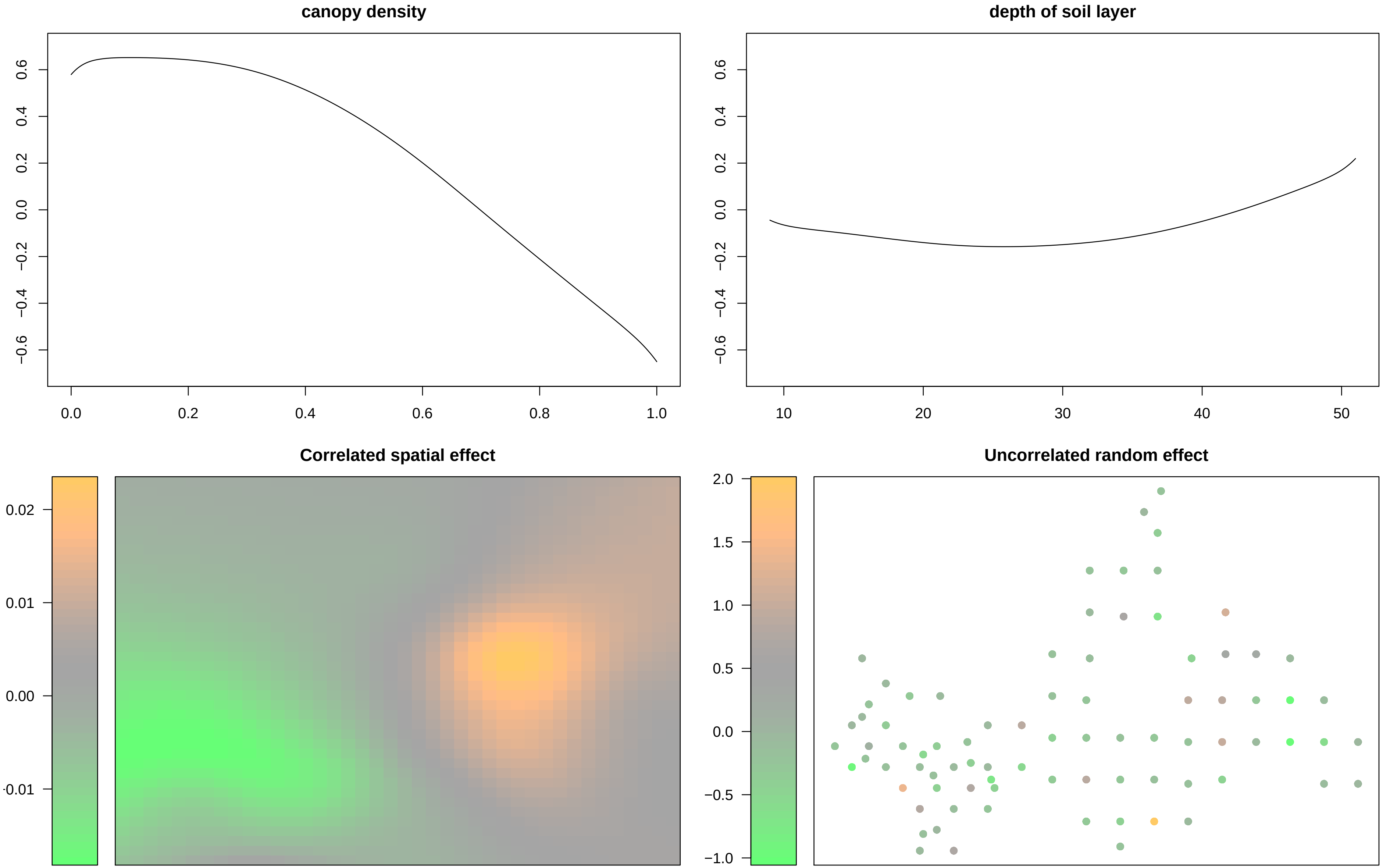
Continuous:	average age of trees at the observation plot
	elevation above sea level in meters
	inclination of slope in percent
	depth of soil layer in centimeters
	pH-value in 0 – 2cm depth
	density of forest canopy in percent
Categorical	thickness of humus layer in 5 ordered categories
	level of soil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

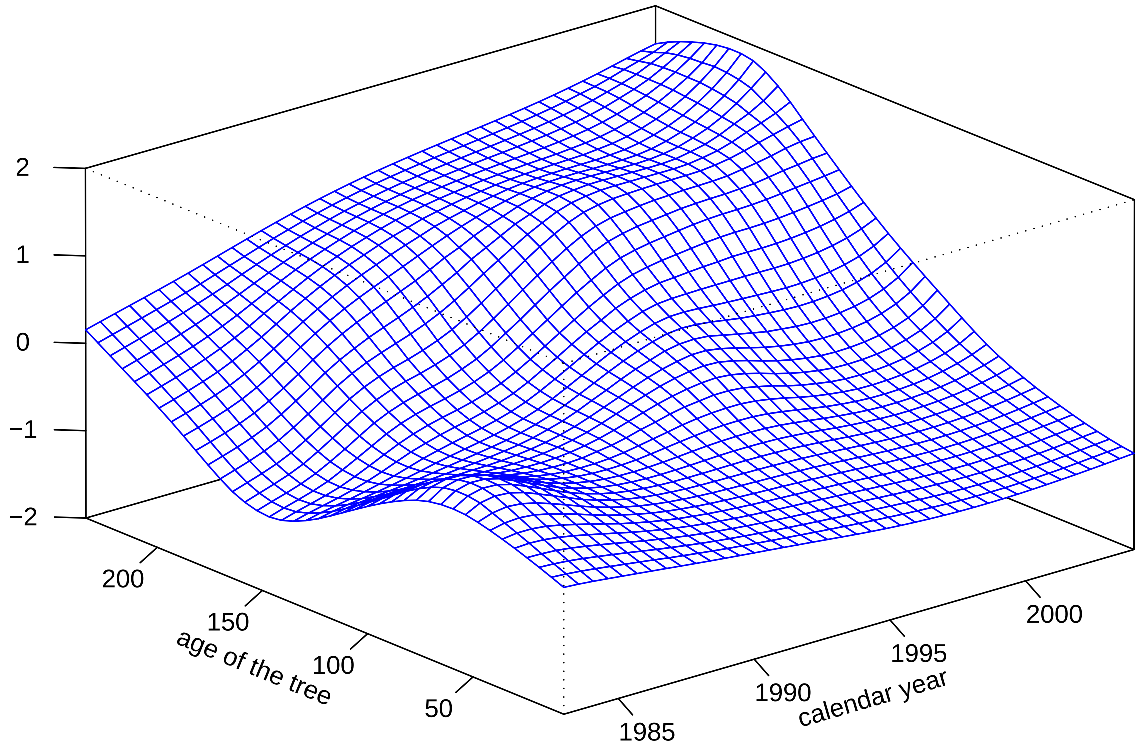
- Possible model:

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

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.
		$f_{spat}(s_{ix}, s_{iy})+$	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
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0.0
0.2
0.4
0.6
0.8
1.0
depth of soil layer
10
20
30
40
50
Correlated spatial effect
0.02
0.01
0.00
-0.01
Uncorrelated random effect
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.

Boosting in a Nutshell

- Boosting is a simple but versatile iterative stepwise gradient descent algorithm.
- Versatility: Estimation problems are described in terms of a loss function ρ.
- Simplicity: Estimation reduces to iterative fitting of base-learners to residuals.

1. Initialize $\hat{\eta}^{[0]} \equiv$ offset; set $m = 0$.
2. Increase m by 1. Compute the negative gradients ('residuals')

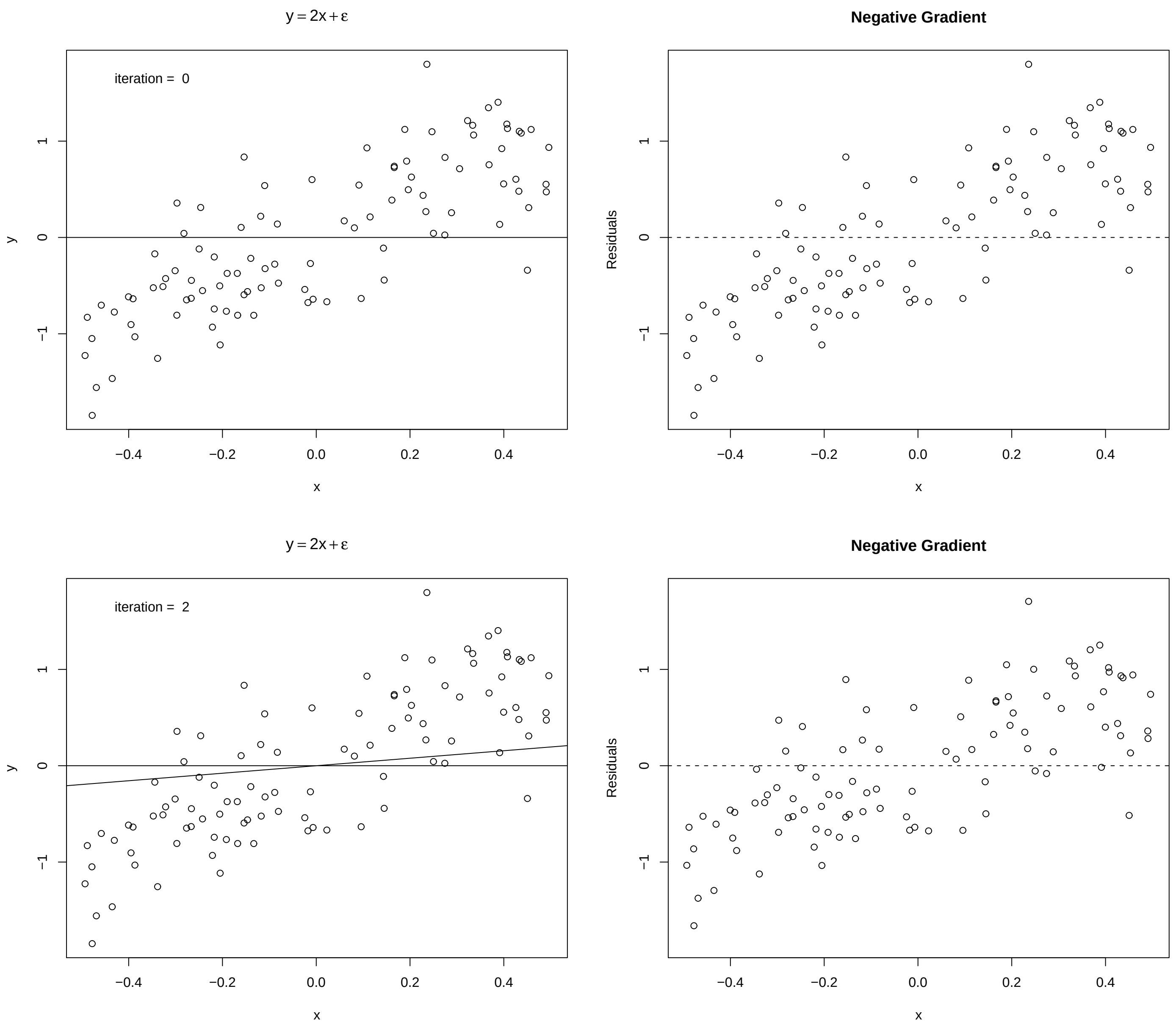
$$u_i = -\frac{\partial}{\partial \eta} \rho(y_i, \eta)\big|_{\eta = \hat{\eta}^{[m-1]}(x_i)}, \; i = 1, \ldots, n.$$

3. Fit the base-learner g to the negative gradient vector $u_1, \ldots, u_n$, yielding $\hat{g}^{[m]}(\cdot)$.
4. Up-date $\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]}(\cdot) + \nu \cdot \hat{g}^{[m]}(\cdot)$
5. Iterate steps 2.-4. until $m = m_{\text{stop}}$.

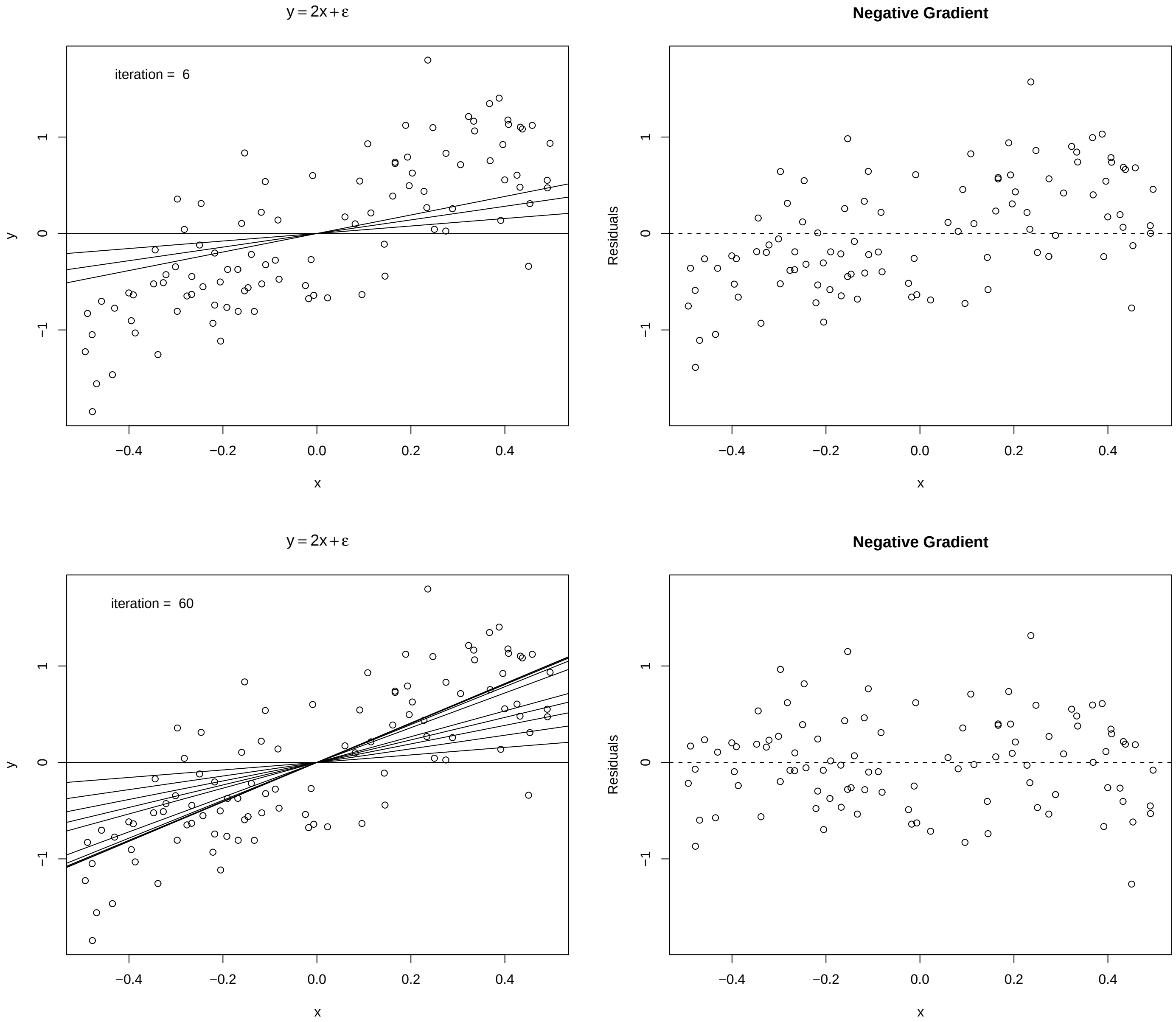
- Example: Linear model with quadratic loss function $\rho(y, \eta) = |y - \eta|^2$.
  - The gradient of the loss function yields the least squares residuals.
  - Base-learner: Least-squares fit $\hat{g}$.
  - In each iteration, update η via

$$\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]} + 0.1\hat{g}$$

    i.e. multiply the current fit with a reduction factor.



y = 2x + ε
iteration = 0
y
x
Negative Gradient
Residuals
x
y = 2x + ε
iteration = 2
y
x
Negative Gradient
Residuals
x


$y = 2x + \varepsilon$
iteration = 6
Negative Gradient
$y = 2x + \varepsilon$
iteration = 60
Negative Gradient
y
Residuals
x
−0.4 −0.2 0.0 0.2 0.4
−1 0 1

- Scales to more complex models:
  - Define a loss function (e.g. the negative log-likelihood).
  - Define a simple base-learning procedure (e.g. a regression tree).
- The reduction factor ν turns the base-learner into a weak learning procedure (avoids to large steps in the boosting algorithm).
- Crucial point: Determine optimal stopping iteration m_{stop}.
- Componentwise boosting: Replace the single base-learning procedure by a sequence of base-learners. Only the best-fitting one is updated in each iteration

  $\Rightarrow$ Structured model fit.
- In geoadditive models: Each additive component is assigned a separate base-learner.
- Boosting implicitly implements variable selection (early stopping).

Base-Learners For Geoadditive Regression Models

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

$$u = g(x) + \varepsilon \qquad \text{smooth nonparametric effect}$$
$$u = g(x_1, x_2) + \varepsilon \qquad \text{smooth surface / spatial effect}$$
$$u = x_1 g(x_2) + \varepsilon \qquad \text{varying coefficients}$$

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

- All base-learners in geoadditive regression models will be given by penalised least squares (PLS) fits

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

  characterised by the hat matrix

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

- Univariate spline smoothing: Approximate the function $g(x)$ by a linear combination of B-spline basis functions, i.e.

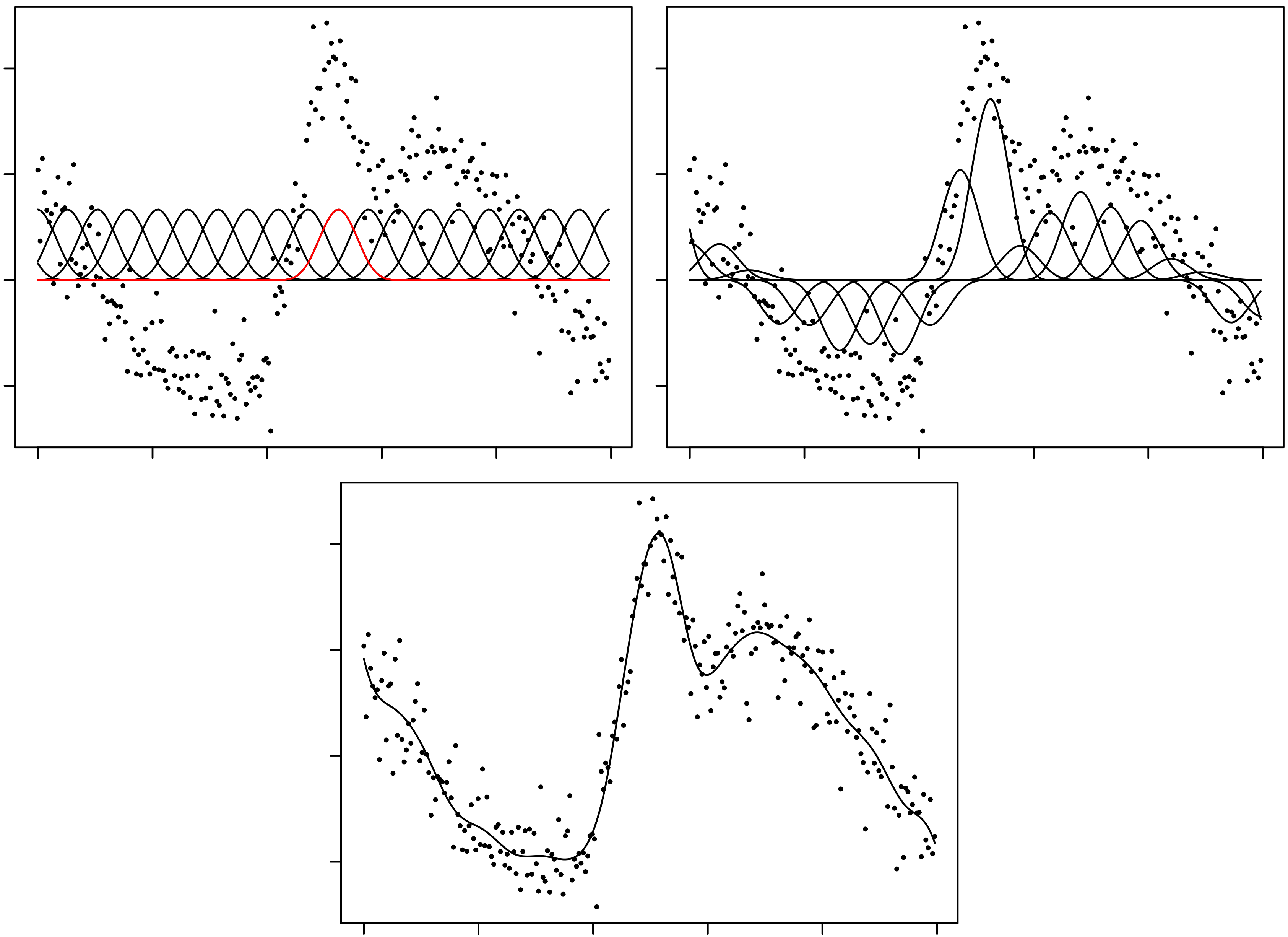
$$g(x) = \sum_j \beta_j B_j(x)$$

- In matrix notation:

$$u = X\beta + \varepsilon.$$

- Least squares estimate for β and predicted values:

$$\hat{\beta} = (X'X)^{-1}X'u \qquad \hat{y} = X(X'X)^{-1}X'u$$



- B-spline fit depends on the number and location of basis functions

  $\Rightarrow$ Difficult to obtain a suitable compromise between smoothness and fidelity to the data.

- Add a roughness penalty term to the least squares criterion.

- Simple approximation to squared derivative penalties: Difference penalties

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

- Can be written as quadratic forms

$$\lambda \beta' D' D \beta = \lambda \beta' K \beta$$

  based on difference matrices D.

- Replace the least-squares estimate and fit with penalised least squares (PLS) variants:

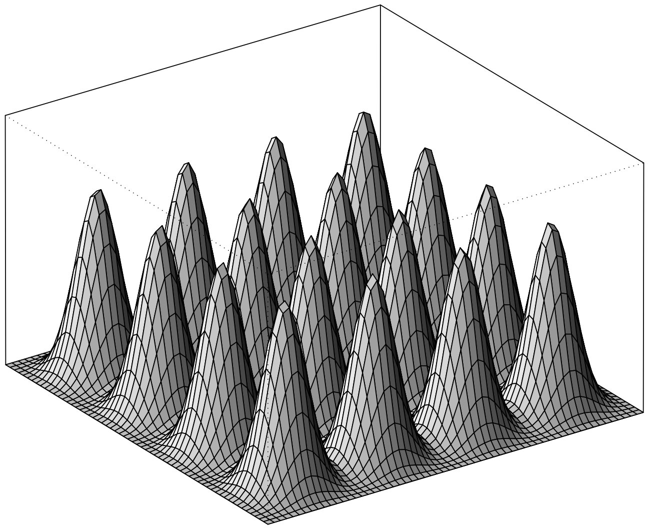
$$\hat{\beta} = (X'X + \lambda K)^{-1} X'u \qquad \hat{u} = X(X'X + \lambda K)^{-1} X'u$$

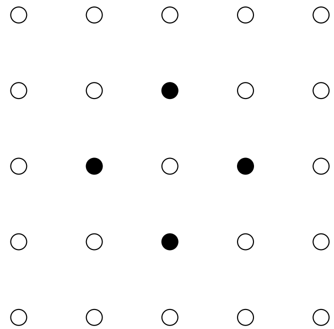
- The base-learner is characterised by the hat matrix

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

- PLS base-learners can also be derived for
  - Interaction surfaces $f(x_1, x_2)$ and spatial effects $f(s_x, s_y)$,
  - Varying coefficient terms $x_1 f(x_2)$ or $x_1 f(s_x, s_y)$,
  - Random intercepts b_i and random slopes xb_i, and
  - Fixed effects $x\beta$.

- 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 rowwise and columnwise penalties as

$$\begin{aligned}\text{pen}_{\text{row}}(\beta) &= \lambda\beta'\left(I \otimes K_1\right)\beta \\ \text{pen}_{\text{col}}(\beta) &= \lambda\beta'\left(K_2 \otimes I\right)\beta.\end{aligned}$$

- The overall penalty is then given by

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

- Varying coefficient terms $x_1 f(x_2)$ or $x_1 f(s_x, s_y)$:

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

  where X^* is the design matrix representing $f(x_2)$ or $f(s_x, s_y)$.

- Cluster-specific random intercepts: The design matrix is a zero/one incidence matrix linking observations to clusters and the penalty matrix is a diagonal matrix.

- Fixed effects: Set the smoothing parameter to zero (unpenalised least squares fit).

- All base-learners can be described in terms of a penalised hat matrix

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

  with suitably chosen design matrix X and penalty matrix K.

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.

- The complexity of a linear model can be assessed by the trace of the hat matrix, since

$$\text{trace}(X(X'X)^{-1}X') = \text{ncol}(X).$$

- In analogy, the effective degrees of freedom of a penalised least-squares base-learner are given by

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

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

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

- Difficulty: For most PLS base-learners, the penalty matrix K has a non-trivial null space, i.e.

$$\dim(\mathcal{N}(K)) \geq 1.$$

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

  $\Rightarrow \mathrm{df}(\lambda) = 1$ can not be achieved.

- 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.

- This will also allow to choose between linear and nonlinear effects within the boosting algorithm.

A Generic Boosting Algorithm

- 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.

- Componentwise boosting procedure based on the loss function $\rho(\cdot)$:

1. Initialize the model components as $\hat{f}_j^{[0]}(\cdot) \equiv 0$, $j = 1, \ldots, r$. Set the iteration index to $m = 0$.

2. Increase m by 1. Compute the current negative gradient

$$u_i = -\frac{\partial}{\partial \eta} \rho(y_i, \eta) \Big|_{\eta = \hat{\eta}^{[m-1]}(\cdot)}, \quad i = 1, \ldots, n.$$

3. Choose the base-learner g_{j^*} that minimizes the L_2-loss, i.e. the best-fitting function according to

$$j^* = \underset{1 \le j \le r}{\operatorname{argmin}} \sum_{i=1}^{n} (u_i - \hat{g}_j(\cdot))^2$$

where $\hat{g}_j = S_j u$.

4. Update the corresponding function estimate to

$$\hat{f}_{j^*}^{[m]}(\cdot) = \hat{f}_{j^*}^{[m-1]}(\cdot) + \nu S_{j^*} u,$$

where $\nu \in (0, 1]$ is a step size. For all remaining functions set

$$\hat{f}_{j}^{[m]}(\cdot) = \hat{f}_{j}^{[m-1]}(\cdot), \quad j \neq j^*.$$

5. Iterate steps 2 to 4 until $m = m_{\text{stop}}$.

- Determine m_{stop} based on AIC reduction or cross-validation.

- Boosting implements both variable selection and model choice:
  - Variable selection: Stop the boosting procedure after an appropriate number of iterations (for example based on AIC reduction).
  - Model choice: Consider concurring base-learning procedures for the same covariate, e.g. linear vs. nonlinear modeling.

Habitat Suitability Analyses

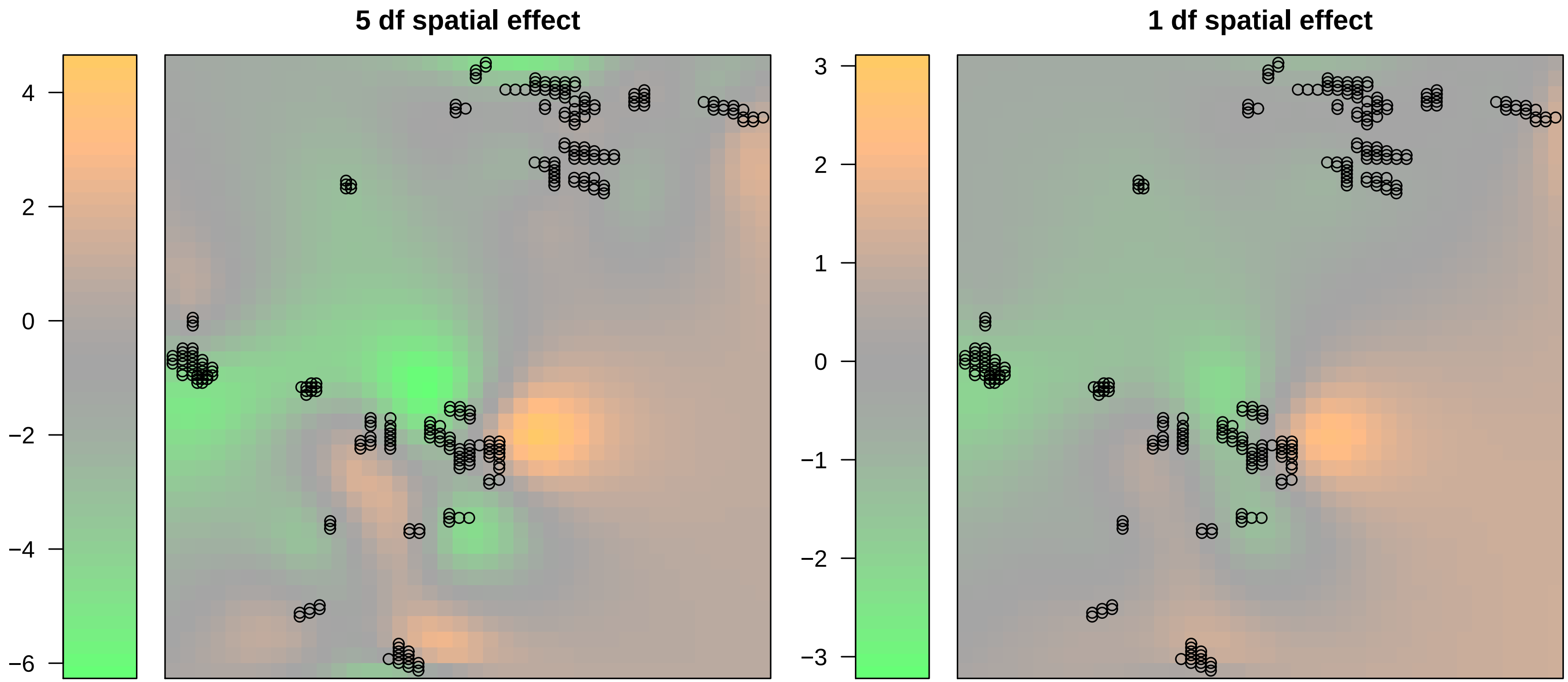
- Identify factors influencing habitat suitability for breeding bird communities collected in seven structural guilds (SG).
- Variable of interest: Counts of subjects from a specific structural guild collected at 258 observation plots in a Northern Bavarian forest district.
- Research questions:
  a) 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.

Variable Selection in the presence of spatial correlation

- 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

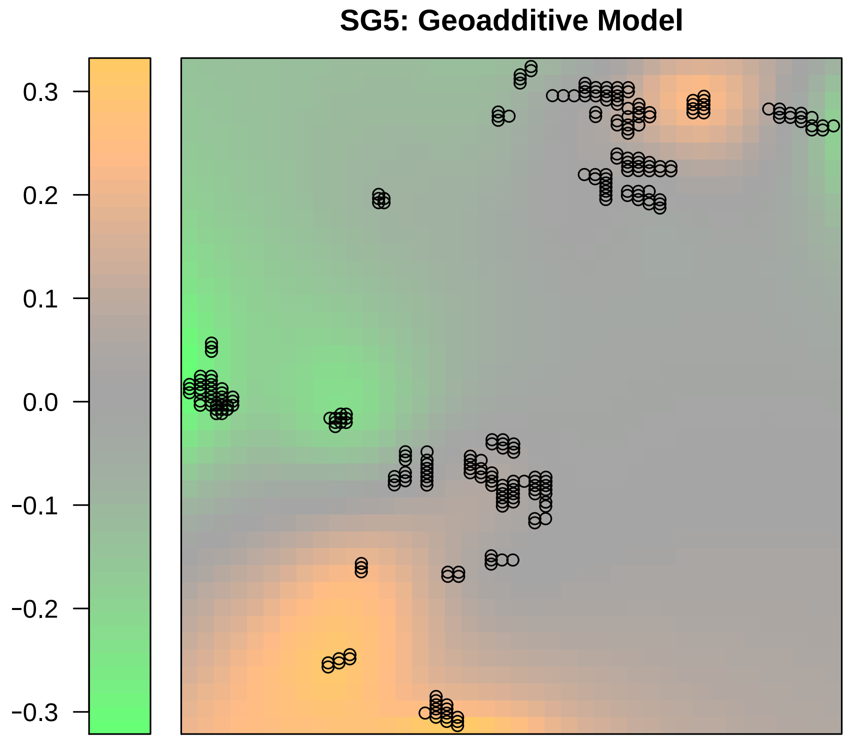
- Spatial effects for high and low degrees of freedom (SG4):

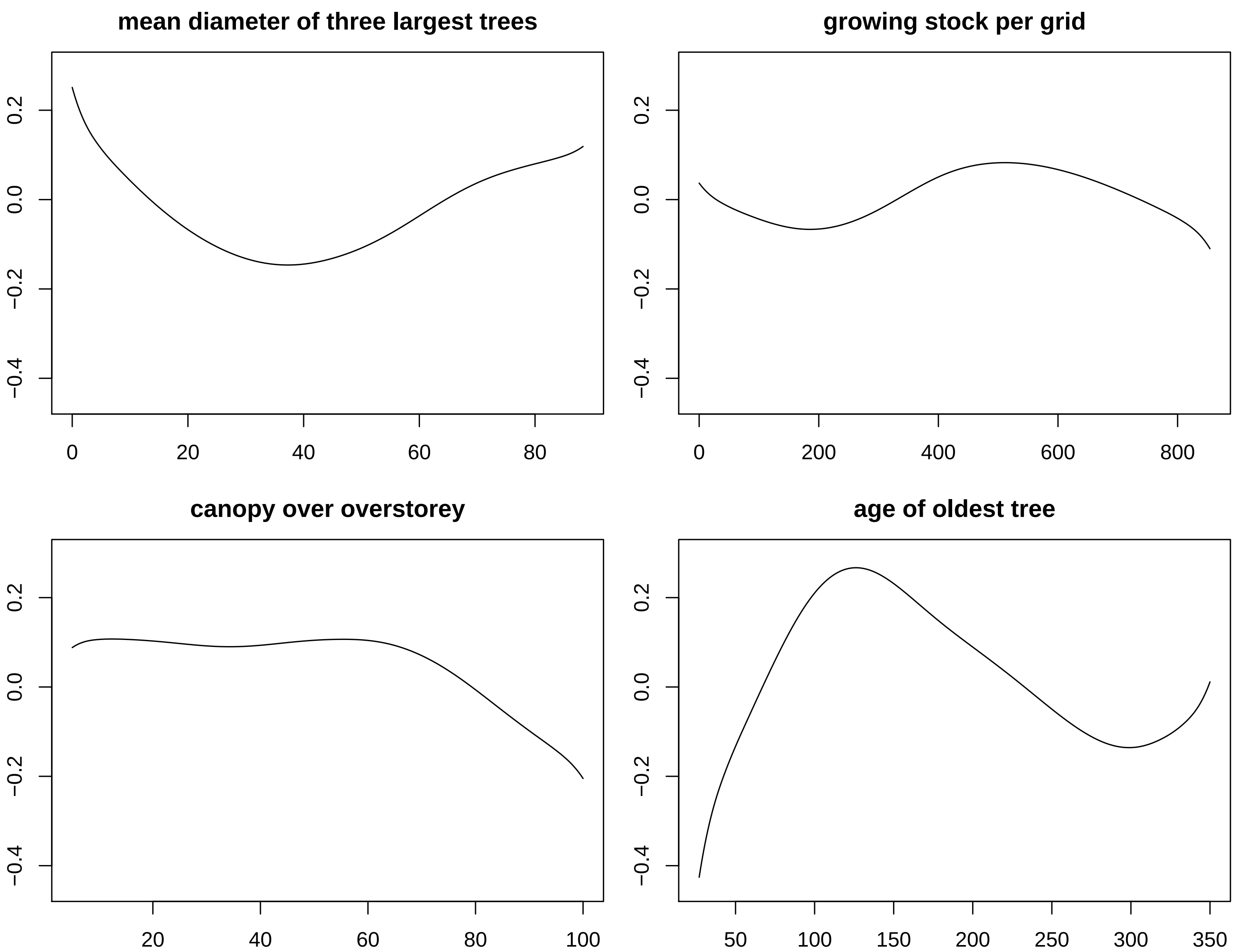


- Spatial correlation has non-negligible influence on variable selection.
- Making terms comparable in terms of complexity is essential to obtain valid results.

Geoadditive Models

- Instead of linear modelling, allow for nonlinear effects of all 31 covariates.
- Decompose nonlinear effects into a linear part and a nonlinear part with one degree of freedom.
- Variable selection for SG5 results in 7 variables without any influence, 3 linear effects, and 21 nonlinear effects.


SG5: Geoadditive Model
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3


mean diameter of three largest trees
growing stock per grid
canopy over overstorey
age of oldest tree
0.2
0.0
−0.2
−0.4
0
20
40
60
80
200
400
600
800
100
50
150
250
300
350

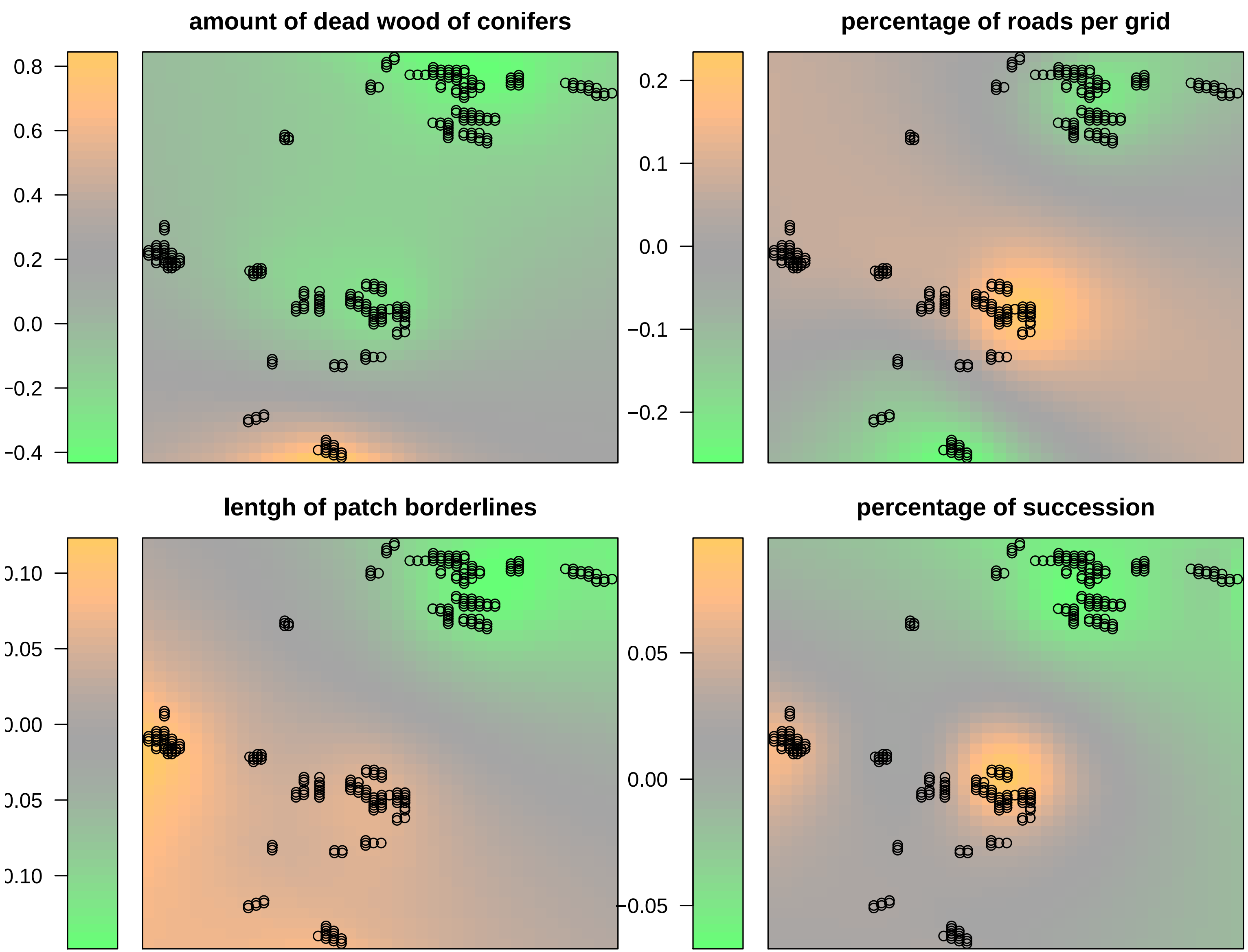
Space-varying effects

- Instead of allowing for nonlinear effects, consider space-varying effects $xg(s_x, s_y)$ for all covariates.

- Decompose space-varying effects into a linear part and a space-varying part with one degree of freedom.

- For SG3, 6 variables have no influence at all, 13 variables have linear effects, and 12 variables are associated with space-varying effects.

- The spatial effect is completely explained by the space-varying effects of the covariates.


amount of dead wood of conifers
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
percentage of roads per grid
0.2
0.1
0.0
−0.1
−0.2
lentgh of patch borderlines
0.10
0.05
0.00
0.05
0.10
percentage of succession
0.05
0.00
−0.05

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.
  - Extend methodology to model selection in continuous time survival models.

- 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`