Boosting Geoadditive Regression Models

Thomas Kneib

Department of Statistics
Ludwig-Maximilians-University Munich

joint work with

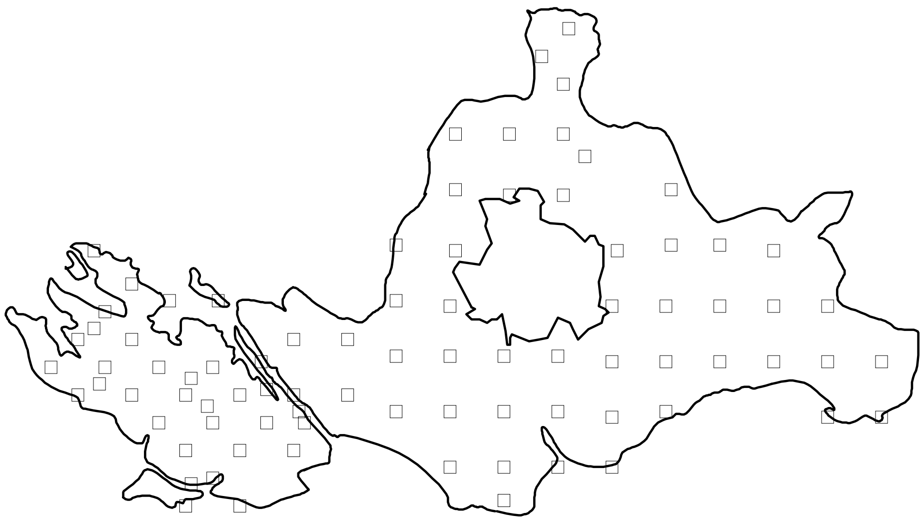
Torsten Hothorn   Gerhard Tutz

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

17.7.2008

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
	base saturation in 4 ordered categories
Binary	type of stand
	application of fertilisation

- Spatio-temporal data requires a model that 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.
- Two major difficulties:
  - How to estimate geoadditive regression models (inference)?
  - How to obtain a sensible model specification (model choice)?

$\Rightarrow$ Componentwise boosting.

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 (e.g. the negative log-likelihood).
- Simplicity: Estimation reduces to iterative fitting of base-learners to residuals (e.g. regression trees).

- Boosting a regression model with predictor η:

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 $m = m_{\text{stop}}$.

- The reduction factor ν turns the base-learner into a weak learning procedure (avoids to large steps in the boosting algorithm).

- 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 and model choice.

- In geoadditive models: Each additive component is assigned a separate base-learner.

- Crucial point: Determine optimal stopping iteration m_{stop}.

- Boosting implicitly implements model choice and variable selection (early stopping).

Base-Learners For Geoadditive Regression Models

- All base-learners in geoadditive regression models will be given by penalised least squares (PLS) fits
$$\hat{u} = X(X'X + \lambda K)^{-1}X'u.$$

- Example: Penalised splines.
  - Approximate a smooth function $f(x)$ using a moderate number of B-spline basis functions, i.e.
  $$f(x) = \sum_j \beta_j B_j(x).$$
  - Define a smoothness penalty
  $$\text{pen}(f) = \lambda \beta' K \beta$$
  based on an approximation to the second derivative $f''(x)$.

- 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 $x\beta$.

Complexity Adjustment & Decompositions

- To avoid biased selection towards more flexible effects, all base-learners should be assigned comparable degrees of freedom

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

- In many cases, a reparameterisation is required to achieve suitable values for the degrees of freedom.

- Example: A linear effect remains unpenalised with penalised spline and second derivative penalty

$$\Rightarrow \quad \mathrm{df}(\lambda) \geq 2.$$

- Decompose $f(x)$ into a linear component and the deviation from the linear component.

- Additional advantage: Allows to decide whether a non-linear effect is required.

Forest Health Data: Results

- Specification of a spatio-temporal logit model:

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

  where η_{it} is a suitable predictor.

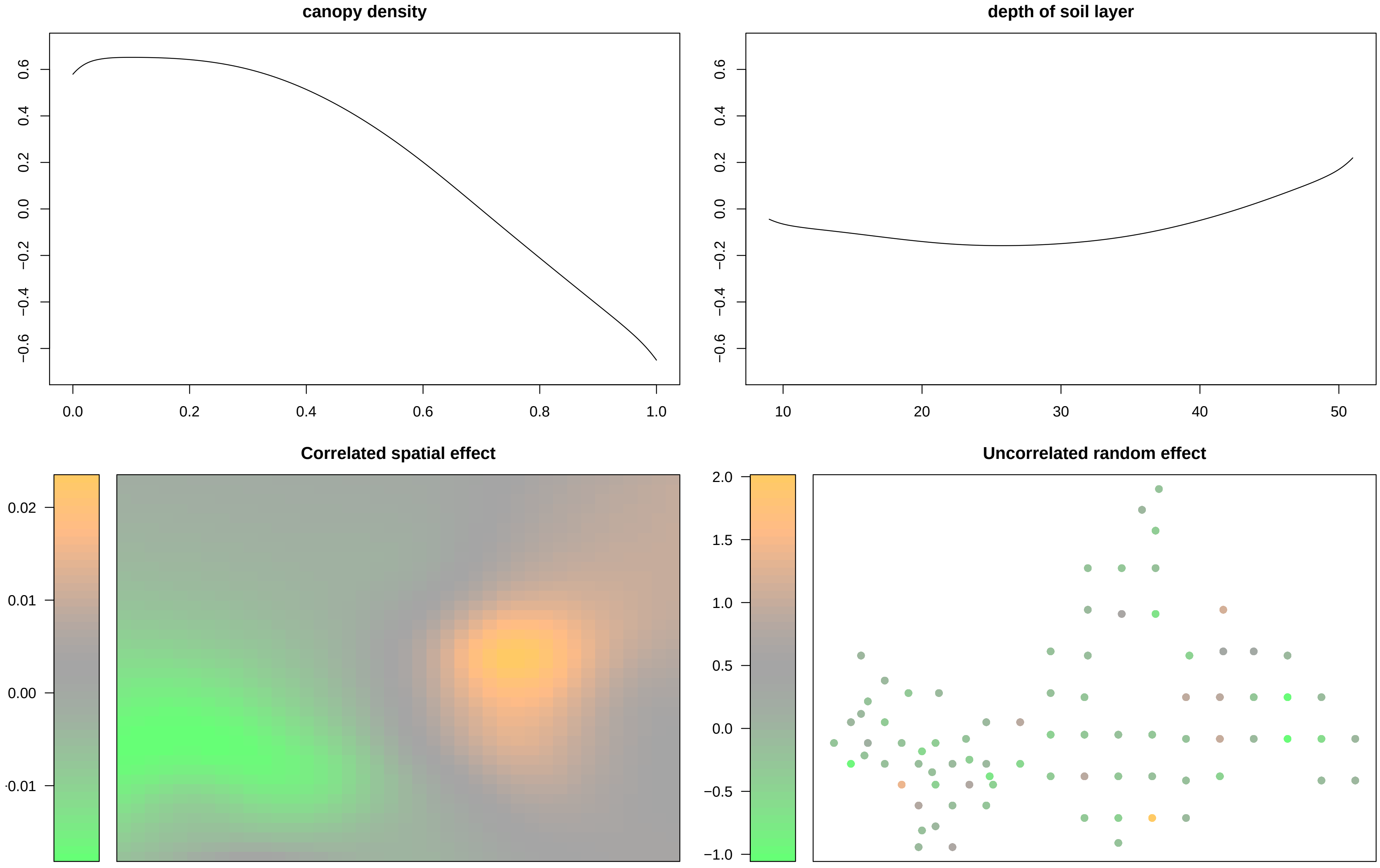
- Boosting relies on the specification of a candidate model with maximum complexity.

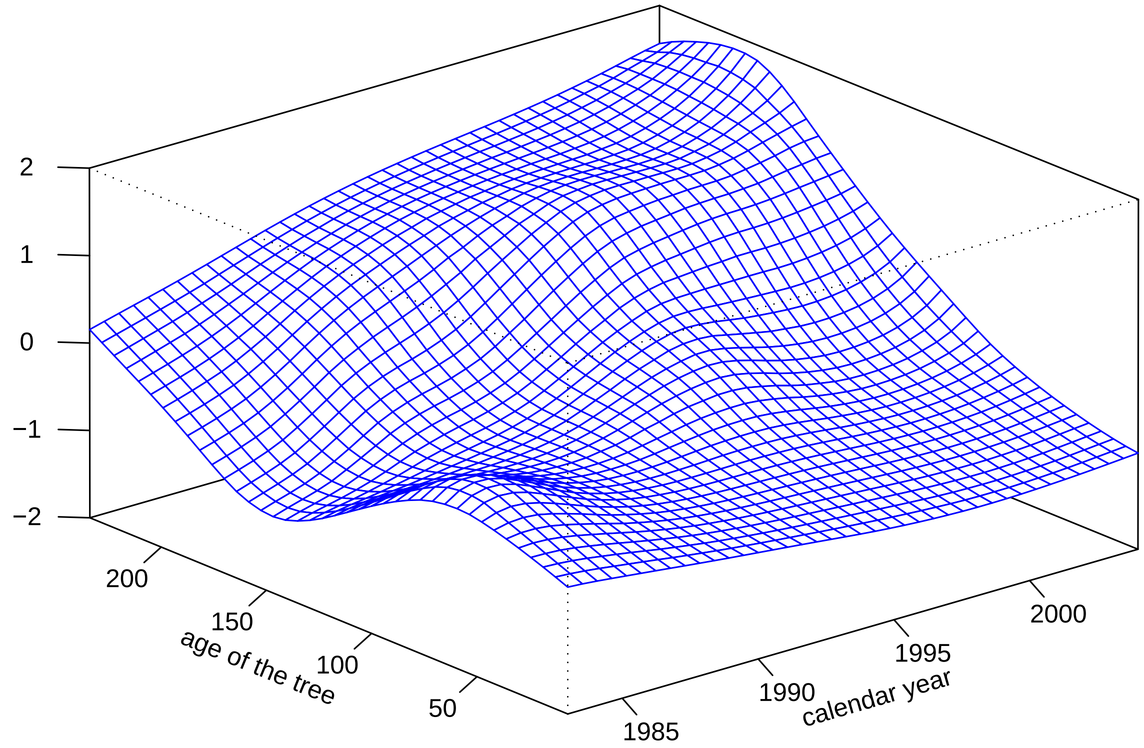
- We considered a candidate model where
  - All continuous covariates are included with penalised spline base-learners decomposed into a linear component and the orthogonal deviation, i.e.

$$g(x) = x\beta + g_{\text{centered}}(x).$$

  - An interaction effect between age and calendar time is included in addition (centered around the constant effect).
  - The spatial effect is included both as a plot-specific random intercept and a bivariate surface of the coordinates (centered around the constant effect).
  - Categorical and binary covariates are included as least-squares base-learners.

- Results:
  - No effects of ph-value, inclination of slope and elevation above sea level.
  - Parametric effects for type of stand, fertilisation, thickness of humus layer, and base saturation.
  - Nonparametric effects for canopy density and soil depth.
  - Both spatially structured (surface) and unstructured effect (random effect) with a clear domination of the latter.
  - Interaction effect between age and calendar time.


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


2
1
0
−1
−2
200
150
100
50
age of the tree
1985
1990
1995
2000
calendar year

Summary

- Boosting provides both a structured model fit and a possibility for model choice and variable selection in geoadditive regression.
- Simple approach based on iterative fitting of negative gradients.
- Implemented in the R package **mboost**.
- Current limitations:
  - Measures of uncertainty are difficult to derive.
  - Maximum complexity structure has to be imposed on the data at hand.

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

- Find out more:

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