Web-based Supplementary Materials for "Variable Selection and Model Choice in Geoadditive Regression Models" by Thomas Kneib, Torsten Hothorn and Gerhard Tutz.

Web Appendix A: Base-Learners for Random Effects

A further modelling component that can be included in geoadditive regresion models are cluster-specific random effects $f(\boldsymbol{z}) = b_c$ or $f(\boldsymbol{z}) = x_1 b_c$, where c is a cluster index that relates an observation to the corresponding cluster the observation pertains to. For each group, a separate effect b_c is specified which, under appropriate distributional assumptions, defines either a random intercept or a random slope of covariate x_1. This leads to a predictor of the form

$$\eta(\boldsymbol{z}) = \ldots + b_{c_i,0} + x_{i1} b_{c_i,1} + \ldots$$

where $c_i \in \{1, \ldots, C\}$ denotes the cluster observation i pertains to. For simplicity, we assume that the clusters are ordered consecutively from 1 to C. In case of longitudinal data, the clusters are defined by individuals whereas the repeated measurements forming the single observations are indexed by i. We utilize the standard assumption of Gaussian random effects, i.e., $b_{c_i,0} \sim \mathcal{N}(0, \tau_0^2)$ is a group-specific random intercept and $b_{c_i,1} \sim \mathcal{N}(0, \tau_1^2)$ is a group-specific random slope.

The corresponding base-learner can then be cast into a similar framework as penalized splines and spatial effects. More specifically, the vector of random intercept evaluations for the observations $i = 1, \ldots, n$ can be expressed as matrix-vector product $\boldsymbol{X}_0 \boldsymbol{b}_0$ where $\boldsymbol{b}_0 = (b_{1,0}, \ldots, b_{C,0})'$ is a vector collecting all random intercepts and $\boldsymbol{X}_0$ is a zero-one incidence matrix that links

each observation with the corresponding random intercept. Random slopes can also be considered as varying coefficient terms with a random intercept as effect modifier. For the vector of effects $x_{i1}b_{c_i,1}$ one obtains the expression $\text{diag}(x_{11}, \ldots, x_{n1})\boldsymbol{X}_0\boldsymbol{b}_1 = \boldsymbol{X}_1\boldsymbol{b}_1$. A random effects base-learner is then given by $\boldsymbol{S}_\lambda = \boldsymbol{X}_k(\boldsymbol{X}_k'\boldsymbol{X}_k + \lambda_k\boldsymbol{I}_C)\boldsymbol{X}_k'$, $k = 0, 1$, where λ_k is a smoothing parameter which is inversely proportional to the corresponding random effects variance.

Web Appendix B: Additional Results on Habitat Suitability

Table 2 presents relative selection frequencies of covariates in a non-spatial GLM, a spatial GLM with high degrees of freedom for the spatial component, and a spatial GLM with one degree of freedom for the spatial component. The relative selection frequency of a covariate is the number of times the corresponding effect has been selected for inclusion in the boosting algorithm, divided by the total number of iterations m_{stop}. The selection frequencies reveal a similar impact of spatial correlation as the estimated regression coefficients: For a high df spatial effect, the inclusion frequencies for several effects are reduced. However, it is more difficult to obtain a reliable picture since the estimated coefficients do not only depend on the selection frequency but also on the order of selection. Typically, covariates selected first in the boosting algorithm, will contribute more to the overall fit and will therefore have a larger effect. In contrast, covariates selected later may well be selected several times but still have comparably small impact on the response. It is therefore advisable to compare estimated effects instead of only selection frequencies.

Table 3 summarises some model fit characteristics for a non-spatial GLM,

Table 1

Environmental variables: Abbreviation, description, range, source and inventory area.

	Description	Range	Source	Inventory
Variables at stand scale				
GST	Growing stock per grid	0-854m^3/ha	Forest inventory	0.05 ha
DBH	Mean diameter of the largest three trees	0-88cm	Forest inventory	0.05 ha
AOT	Age of oldest tree	27-350y	Forest inventory	0.5 ha circle
AFS	Age of forest stand	27-300y	Forest inventory	stand level
DWC	Amount of dead wood of conifers	0-127m^3/ha	Additional inventory	0.1 ha circle
LOG	Amount of logs per grid	0-293m^3/ha	Additional inventory	0.1 ha circle
SNA	Amount of snags and attached dead wood at living trees per grid	0-292m^3/ha	Additional inventory	0.1 ha circle
COO	Canopy over overstorey	5-100%	Estimation in field	1 ha grid
COM	Canopy over middlestorey	0-60%	Estimation in field	1 ha grid
CRS	Percentage of cover of regeneration and shrubs	0-95%	Estimation in field	1 ha grid
HRS	Mean height of regeneration and shrubs	0-10m	Estimation in field	1 ha grid
OAK	Percentage of oak trees	0-40%	Estimation in field	1 ha grid
COT	Percentage of coniferous trees	0-80%	Aerial photo	1 ha grid
PIO	Percentage of pioneer trees (Salix, Betula, Populus)	0-75%	Estimation in field	1 ha grid
ALA	Percentage of alder and ash trees	0-60%	Estimation in field	1 ha grid
MAT	Percentage of cover of mature trees	0-100%	Aerial photo	1 ha grid
GAP	Percentage of gaps per grid	0-19%	Aerial photo	1 ha grid
AGR	Percentage of agricultural land per grid	0-21%	Aerial photo	1 ha grid
ROA	Percentage of roads per grid	0-13%	Aerial photo	1 ha grid
LCA	Number of large cavities per grid	0-15n/ha	Additional inventory	0.5 ha circle
SCA	Number of small cavities per grid	0-33n/ha	Additional inventory	0.5 ha circle
HOT	Hollow trees per grid	0-10n/ha	Additional inventory	0.5 ha circle
CTR	Number of cavity trees per ha	0-14n/ha	Additional inventory	0.5 ha circle
Variables at landscape scale				
RLL	Length of roads at the landscape level	992-12647m	Aerial photo	78.5 ha circle
BOL	Length of patch borderlines	780-7800	Aerial photo	78.5 ha circle
MSP	Mean size of habitat patch	39268-261786	Aerial photo	78.5 ha circle
MDT	Percentage of mature deciduous trees at the landscape level	19-97%	Aerial photo	78.5 ha circle
MAD	Percentage of medium aged deciduous trees at the landscape level	0-69%	Aerial photo	78.5 ha circle
COL	Percentage of coniferous trees at the landscape level	0-77%	Aerial photo	78.5 ha circle
AGL	Percentage of agricultural land at the landscape level	0-41%	Aerial photo	78.5 ha circle
SUL	Percentage of succession at the landscape level	0-24%	Aerial photo	78.5 ha circle

Table 2

Structural guild 4: Relative selection frequencies of covariates in a non-spatial GLM, a spatial GLM with high degrees of freedom for the spatial component, and a spatial GLM with one degree of freedom for the spatial component. The relative selection frequency of a covariate is the number of times the corresponding effect has been selected for inclusion in the boosting algorithm, divided by the total number of iterations m_{stop}.

	GST	DBH	AOT	AFS	DWC	LOG	SNA	COO
non-spatial GLM	0	0.02	0	0	0	0	0	0.18
spatial with 5 df	0	0	0	0	0	0	0	0
spatial with 1 df	0	0	0	0	0	0	0	0.02
	COM	CRS	HRS	OAK	COT	PIO	ALA	MAT
non-spatial GLM	0.01	0.09	0.16	0.14	0	0.08	0	0
spatial with 5 df	0	0	0.01	0.01	0.01	0.01	0.03	0.02
spatial with 1 df	0.01	0	0.02	0.01	0	0.01	0.02	0
	GAP	AGR	ROA	LCA	SCA	HOT	CTR	RLL
non-spatial GLM	0.03	0.05	0.01	0	0.03	0.02	0	0
spatial with 5 df	0	0	0	0	0.01	0	0	0
spatial with 1 df	0	0.01	0	0	0.01	0	0	0
	BOL	MSP	MDT	MAD	COL	AGL	SUL	spatial
non-spatial GLM	0.05	0	0.06	0	0	0.05	0	0
spatial with 5 df	0	0.01	0	0	0.01	0.02	0.01	0.83
spatial with 1 df	0	0.01	0.01	0	0	0.02	0	0.84

Table 3

Structural guild 4: Model fit characteristics for a non-spatial GLM, a spatial GLM with high degrees of freedom for the spatial component, and a spatial GLM with one degree of freedom for the spatial component.

	Deviance	Pearson χ^2	df	AIC
non-spatial GLM	121.066	215.829	8.986	212.623
spatial with 5 df	52.248	72.043	37.438	200.708
spatial with 1 df	77.808	133.071	26.786	204.966

a spatial GLM with high degrees of freedom for the spatial component, and a spatial GLM with one degree of freedom for the spatial component. The inclusion of a spatial effect leads to a better model fit not only in terms of the deviance and Pearson‘s χ^2 (which solely reflect the increased model flexibility) but also in terms of AIC. Both spatial models are roughly comparable in terms of AIC, although the increased flexibility of the high df spatial model seems to provide a somewhat better fit. However, when comparing Anscombe residuals for the three models (Figure 1), the high df model actually introduces extreme residuals. In contrast, the low df spatial model shows preferrable residual variation, which is also comparable to the residual variation of the purely parametric model. Combining the small difference between the spatial models in terms of AIC with the findings from the residual plots, the low df spatial model seems to be the most appropriate choice from the three models under consideration.

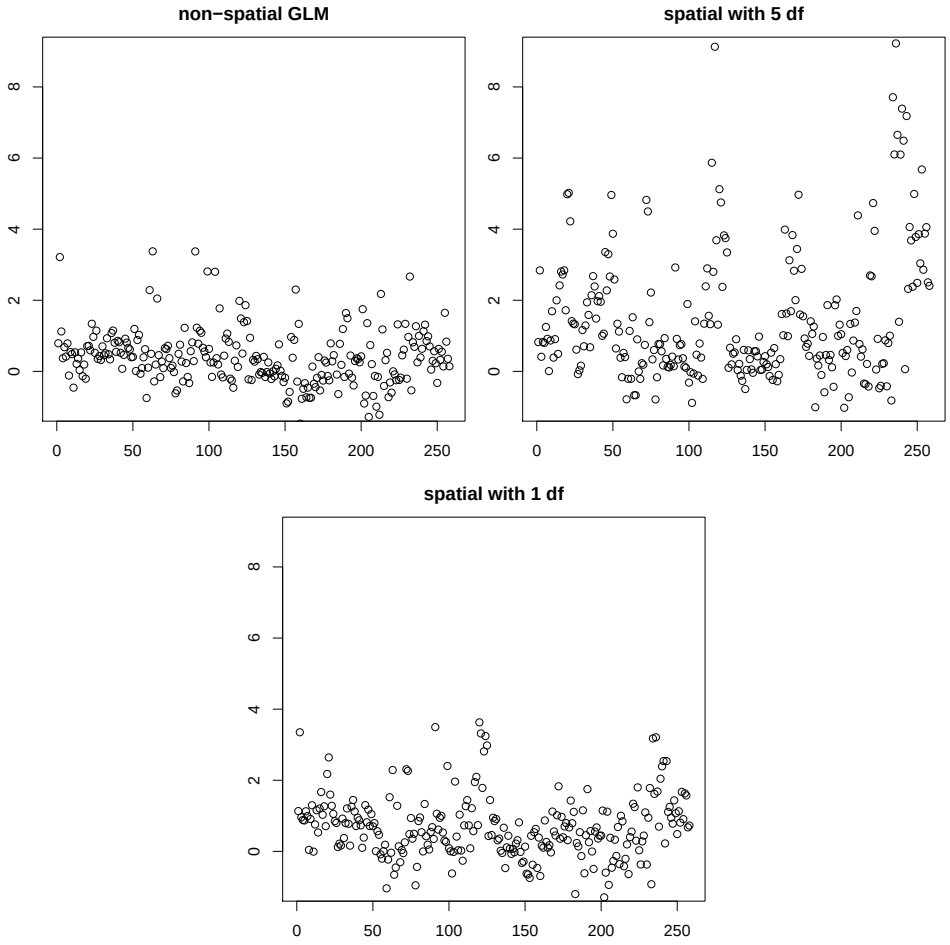

Figure 1. *Structural guild 4: Anscombe residuals for a non-spatial GLM, a spatial GLM with high degrees of freedom for the spatial component, and a spatial GLM with one degree of freedom for the spatial component.*

Web Appendix C: Forest Health

In this application, the health status of beeches at 83 observation plots located in a northern Bavarian forest district has been assessed in visual forest health inventories carried out between 1983 and 2004. Originally, the health status is classified on an ordinal scale, where the nine possible categories denote different degrees of defoliation. The domain is divided in 12.5% steps, ranging from healthy trees (0% defoliation) to trees with 100% defoliation. Since data become relatively sparse already for a medium amount of defoliation, we model the dichotomized response variable defoliation with categories 1 (defoliation above 25%) and 0 (defoliation less or equal to 25%). Table 4 contains a brief description of the covariates in the data set.

Obviously, the collected data have both a temporal and a spatial component that has to be considered in the analysis. Moreover, due to the longitudinal structure of the data, we are interested in estimating plot-specific random effects. Previous studies described in Kneib and Fahrmeir (2006) and Kneib and Fahrmeir (2008) also suggest the presence of interaction effects and non-linear influences of some continuous covariates. Based on these results we consider a logit model with candidate predictor

$$\begin{aligned}\eta(\boldsymbol{z}) &= \boldsymbol{x}'\boldsymbol{\beta} + f_1(\text{ph}) + f_2(\text{canopy}) + f_3(\text{soil}) + f_4(\text{inclination}) \\ &\quad + f_5(\text{elevation}) + f_6(\text{time}) + f_7(\text{age}) + f_8(\text{time}, \text{age}) \\ &\quad + f_9(s_1, s_2) + b_{\text{plot}},\end{aligned}$$

where $\boldsymbol{x}$ contains the parametric effects of the categorical covariates and the base-learners for the smooth effects $f_1, \ldots, f_7$ are specified as univariate cubic penalized splines with 20 inner knots and second order difference penalty.

For both the interaction effect f_8 and the spatial effect f_9 we assumed bivariate cubic penalized splines with first order difference penalties and 12 inner knots for each of the directions. Finally, the plot-specific random effect b_{plot} is assumed to be Gaussian with random effects variance fixed such that the base-learner has one degree of freedom. Similarly, all univariate and bivariate nonparametric effects are decomposed into parametric parts and nonparametric parts with one degree of freedom each. Since the number of observations is too large for AIC-based choice of the stopping rule, m_{stop} was determined by a bootstrapping procedure.

After applying the stopping rule, no effect was found for the ph-value, inclination of slope and elevation above sea level. The univariate effects for age and calendar time where strictly parametric linear but the interaction effect turned out to be very influential. The sum of both linear main effects and nonparametric interaction is shown in Figure 2. The spatial effect was selected only in a relatively small number of iterations whereas the random effect was the component selected most frequently. We can therefore conclude that spatial variation in the data set seems to be present mostly very locally, which is also confirmed by the results found in Kneib and Fahrmeir (2008). For canopy density and soil depth nonlinear effects where identified as visualized in Figure 2. In summary, our results resemble those found in previous analyses but have the advantage that model choice and variable selection can be addressed simultaneously with model fitting.

Table 4

Forest health data: Description of covariates.

Covariate	Description
age	age of the tree in years (continuous, 7≤age≤234)
time	calendar time (continuous, 1983≤time≤2004)
elevation	elevation above sea level in meters (continuous, 250≤elevation≤480)
inclination	inclination of slope in percent (continuous, 0≤inclination≤46)
soil	depth of soil layer in centimeters (continuous, 9≤soil≤51)
ph	ph-value in 0-2cm depth (continuous, 3.28≤ph≤6.05)
canopy	density of forest canopy in percent (continuous, 0≤canopy≤1)
stand	type of stand (categorical, 1=deciduous forest, -1=mixed forest).
fertilisation	fertilisation (categorical, 1=yes, -1=no).
humus	thickness of humus layer in 5 categories (ordinal, higher categories represent higher proportions).
moisture	level of soil moisture (categorical, 1=moderately dry, 2=moderately moist, 3=moist or temporary wet).
saturation	base saturation (ordinal, higher categories indicate higher base saturation).

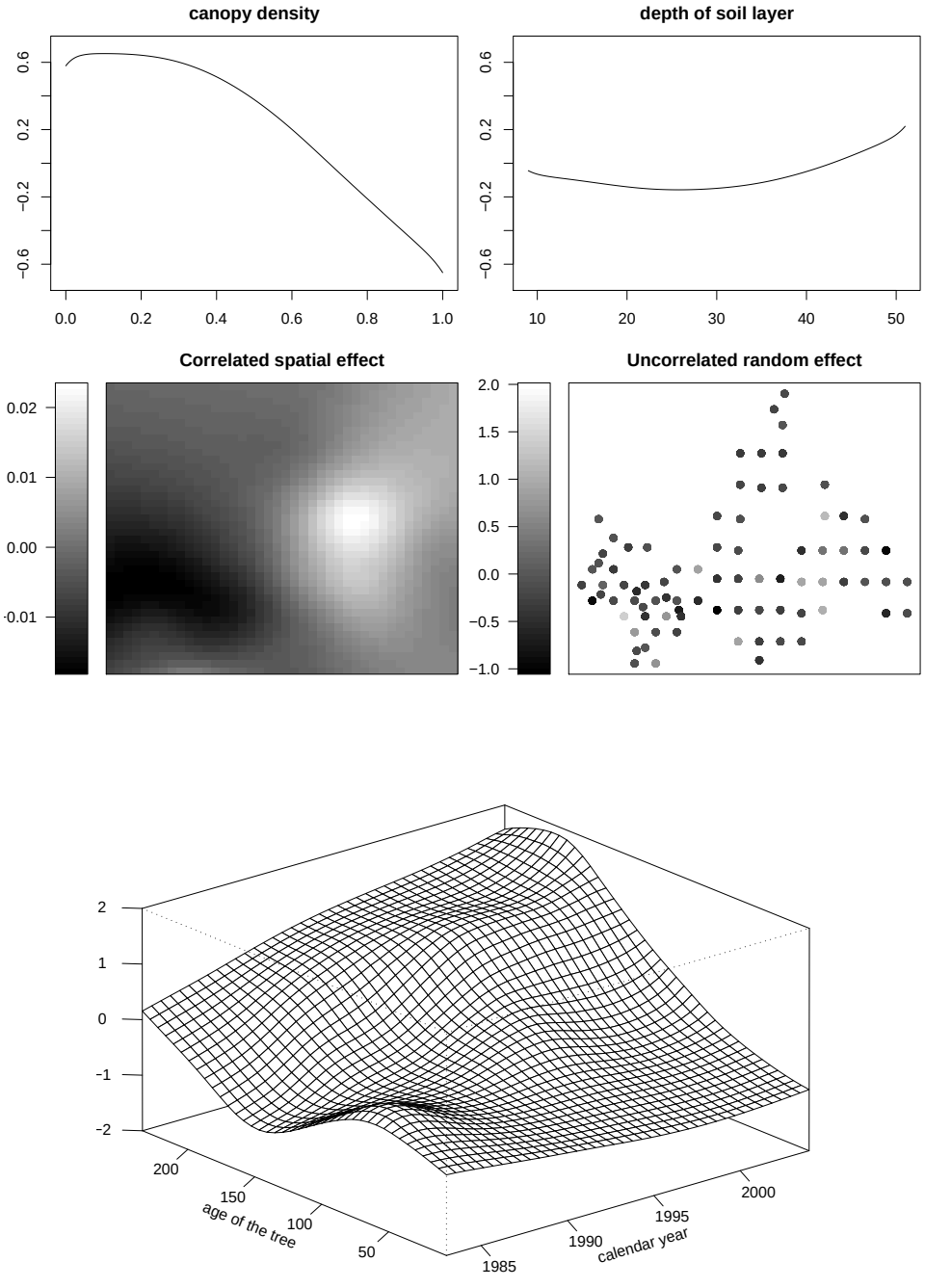

Figure 2. *Forest health: Estimation results.*

References

Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics* **62**, 109–118.

Kneib, T. and Fahrmeir, L. (2008). A space-time study on forest health. In R. Chandler and M. Scott, editors, *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*. New York: John Wiley & Sons.