DIPLOMA THESIS

Statistical methods in niche modelling

for the spatial prediction
of forest tree species

Ludwig–Maximilians–Universität München, Institut für Statistik

submitted by: *Veronika Fensterer*
supervision: *Prof. Dr. Helmut Küchenhoff*
date of submission: *June 1, 2010*

Abstract

A collective and aggregated overview of information is required for effective forest land management and a versed judgement on habitat suitability, especially against the background of climate change.

Based on the combination of varying data sources for the description of forest characteristics in the Bavarian Alps, modelling techniques, including generalised additive models, random forests and boosting algorithms, are compared in their capability of modelling the habitat of tree species. The properties of merely data–directed models are contrasted to hypotheses–directed approaches based on expert knowledge.

Sparse generalised additive models including a spatial trend provide a comparably good discriminatory power and suitable prediction maps. However, the spatial trend does not compensate the small–range effect of a disregarded predictor. Investigations concerning the properties on extrapolated data reveal unbiased and often more precise estimations than using data–directed models. To some extent, the models profit from weighting, the inclusion of interactive factors and a spatially balanced design.

The data–directed approaches, which are mainly designed for prediction, render valuable information for the selection of relevant predictors and for the comprehension of the relationships within the data by implicit variable selection, importance measures and partial effects. Therefore, the varying approaches afford diversified insights, especially into the interactive relations.

Preface

This diploma thesis was prepared at the Department of Statistics at the Ludwig-Maximilians-University Munich in cooperation with the environment consulting bureau AGWA.

First of all, I want to thank Prof. Dr. Helmut Küchenhoff from the Department of Statistics for taking over the supervision of the thesis and for supporting me in the planning of the proceedings and in the statistical implementation.

Further, my gratitude goes back to Karl Mellert from AGWA for the excellent co–operation during the entire project, for the selection of the expert models and for his help to grasp ecological issues.

Due to the methodological and computational advice, I want to thank Dr. Carolin Strobl, Prof. Dr. Torsten Hothorn, Manuel Eugster and Stefanie Kalus from the Department of Statistics.

Contents

1. Introduction

1.1. The role of statistical modelling in ecology

Ecology, as a sub-discipline of biology, deals according to a definition of Krebs (1985) with "the scientific study of the interactions that determine the distribution and abundance of organisms". For that purpose, statistical methods can support the exploration and the comprehension of the underlying structures.

The spatial or temporal dispersion of different kinds of plant or animal species, the relationships among them and also the correlation with environmental factors are analysed in order to link theoretical considerations with directly observed information (Ludwig and Reynolds, 1988), because the exploration of patterns in biotic communities is the central topic in statistical ecology.

Two different approaches are commonly used to examine ecological structures: data can be collected either experimentally or observationally. In an experimental study design, one or more parameters, which potentially influence the outcome, are varied systematically by fixing the remaining variables and the resulting effects can be attributed to the varied parameters. The favoured strategy of statistical ecology consists of the observation of ecological patterns and the corresponding parameters over time or within a determined region, with the result, that the analysis is focused on the existing situation, rather than on a systematic manipulation of the basic conditions.

However, some theoreticians are not convinced of the benefits of empirical research in ecology:

"The use of mathematical models in theoretical ecology is seen by some ecologists today as a symptom of a malignancy infecting the entire discipline of ecology." (Caswell, 1988)

Is the construction of statistical models really a waste of time, which even deteriorates ecological research? Caswell (1988) presumes the cause of this attitude in a misunderstanding of the relation between statistical modelling and ecological theory. Furthermore, he makes clear that an appropriate utilisation of statistical models can ameliorate the understanding of ecological theories.

Therefore, according to Toft (1990), it is essential,

- that the examined question is clearly formulated in order to choose the most appropriate method,
- that the data is as familiar as possible,

- that the biases, which can emerge from violations of the assumptions of the applied method, are well–known and
- that the practitioner is aware of the difference between statistical and biological significance.

Especially the third point causes sometimes difficulties in ecological problems. Besides the assumptions concerning the distributions and the asymptotics, etc., particularly the independence of the observations, which relates closely to the experimental design, is often not given. Hurlbert (1984) calls this lack of true replications, regardless of whether it is a consequence of spatial or of temporal dependency, "pseudoreplication".

Though, taking into account the weaknesses and limitations, the application of statistical modelling techniques are an effective tool to reveal and analyse ecological relationships.

1.2. Subjects and aims of the project and the thesis

This work is written in the context of the joint Bavarian–Austrian research project "Forest Information System for the Northern Alps" (WINALP). The project aims at a reliable, area–wide information system for the natural capacity of montane forests in order to support the decision making process for site–specific forest management. The examined area extends over the Northern Alps of Bavaria, Tyrol and parts of Salzburg.

Particularly against the background of climate change, which will effect the growth conditions, a forest information system provides information about the actual situation and the current species distribution. Hence, accurate predictions for the future distribution of ecological forest types can be gained.

The schedule of this project is divided into eight steps (Hochschule Weihenstephan–Triesdorf, 2010):

1. Development of a geographic information system with a high–resolution description of geological conditions.
2. Determining the requirements of the target user group.
3. Deduction of ecologically relevant parameters.
4. Modelling of forest types on a scale of (1:25,000), verification and recalibration.
5. Modelling of special maps requested from the users.
6. Construction of maps with ecological factors and forest types under the influence of climate change.
7. Development of a handbook for the forest–type specific cultivation, maintenance and restoration of montane forests.
8. Introduction of the system into practical application and user instruction.

This diploma thesis develops the statistical framework for point four. On the basis of the results from the previous steps, which link detailed point data and the information from digital maps, different modelling techniques for the dichotomously measured occurrence of various trees species in dependence of the corresponding environmental conditions, i.e. the ecological niches of the species, are examined. Ecologists define the ecological niche as "the conjunction of environmental conditions within which a species can maintain populations without immigration" (MacArthur, 1972).

Predominantly, two objectives are pursued with ecological modelling, which are often contrarious. The predictive capability of a statistical model is often considered more important than gaining new insights into the relationship between species and their environment. But, according to Austin (2002), the disregard of ecological plausibility in favour of more precise predictions is a limiting factor for the robustness and the explanatory power of a model. For that purpose, two strategies are embarked on:

On the one hand, a preferably parsimonious, hypotheses–directed model is developed with statistical modelling techniques as well as with expert knowledge. Thus, besides the goodness of fit of a model, also the plausibility of the results are incorporated in the model selection procedure for a more truthful image of the data generating process.

On the other hand, data–directed models are only established with the criterion of the best predictive performance as possible. However, models with a good predictive performance are usually black box methods, which do not contribute to the comprehension of a mechanism, whereas plausible models often achieve poorer predictive accuracy.

Moreover, a simulation study is conducted for the further examination of the applied modelling techniques. Different scenarios are supposed to provide information on the properties under various study designs, the role of the data generating mechanism and the effects of variable selection and extrapolation.

Guisan et al. (2007) showed, that the predictive performance of different modelling techniques strongly varies between different tree species. So it is reasonable to include several species in the model comparison, which differ in prevalence and spatial distribution. Besides the spruce as a frequent tree species, the ash, which is a pioneer tree species and the rare Swiss pine are considered.

1.3. Statistical realisation of modelling species–habitat relationships

Initially, a descriptive overview of the data is given in the second chapter. The data collection procedure of the previous steps of the WINALP project and the resulting data structure is briefly illustrated. Also the occurrence of the exemplary tree species and the appendant habitat properties, i.e. the properties of the area the species are living in, are described. Furthermore, a first insight into the dependence of the presence or absence of a species from environmental factors is given by univariate response shapes along ecological gradients.

The applied modelling techniques focus on three statistical methods, namely generalised additive models, random forest and boosting. An overview of theoretical concepts is given in the chapters 3 to 5.

Assuming an underlying functional association structure between the response and the predictor variables, the hypotheses–directed approach, which is described in chapter 3, should provide an ecologically reasonable and parsimonious model. Basically, this is realised through a generalised additive model. In addition to that, the most important interaction effects on the residuals of the final model, which are identified with a decision tree, are added to the model as a factor variable along the lines of Maggini et al. (2006). Further, the profit of imitating a balanced sample along ecological gradients by weighting is investigated, since this is the preferred sampling strategy for habitat suitability modelling (Hirzel and Guisan, 2002). Sparse models are selected with the generalised cross–validation criterion as well as with the verification of plausibility by an expert.

Instead, the data–directed models learn the patterns from the observed data in order to make predictions considering the data generating mechanism as unknown. This approach is realised with random forests described in chapter 4 and with two different boosting models, which are illustrated and explained in chapter 5.

The models are examined from two different points of view: In chapter 6 the three approaches are compared with different simulated data sets in terms of variations in the data generating process, the sampling design and the analysis methods.

The second perspective investigates the modelling techniques by analysing the WINALP data, which are divided into a training and a test data set. The former is used for fitting the models, whereas the test data set is taken for the quantification of predictive accuracy and for validation. For each modelling technique also the importance of the covariates for the species distribution and the partial effects is determined, because the individual species require particular demands on water, energy, nutrients and geomorphodynamics. Furthermore, the differences between the models in respect of the spatial illustration of the predictions is demonstrated. The results are depicted and evaluated in chapter 7, before a final summary and an outlook is given in chapter 8.

The simulation study and the evaluation of the data are accomplished with the statistical software R (R Development Core Team, 2009).

2. Data description

2.1. Data collection and data structure

The examined sites with varying grid intensity are located in the growing area "Northern Alps" in Bavaria as depicted in the chart below and comprise observation points within circa 4,600km^2. This area is well–suited for the analysis of the species–habitat relations, because the ecosystem of the forest is very diverse and hence, the ecological niches of the tree species are likely to be described more accurately.

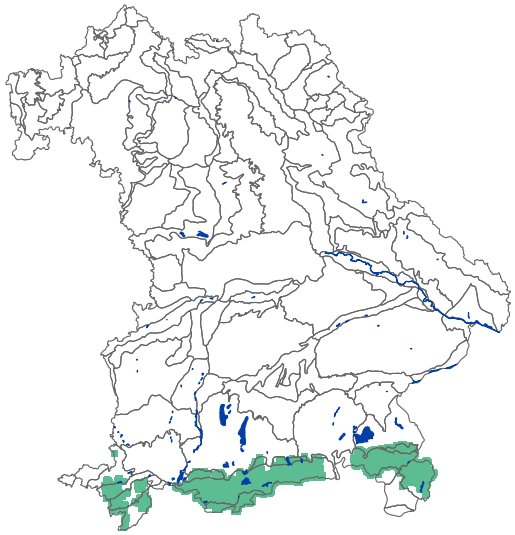


FIGURE 2.1.: *Examined observation sites (green) in Bavaria (separated in growing areas).*

The first task of the WINALP project was the collection of physiographical and vegetation–ecological raw data for a geographic information system. Data on soil climate and topography, which were available in the form of digital maps, were linked to detailed point data on soil profile, tree occurrence and phytosociological information, which was collected between 1978 and 1998, at 55,357 sites. The plot sizes vary mainly between 50m^2 and 300m^2.

Based on this raw data, relief, soil, climate and vegetation parameters, like temperature or precipitation, were deviated and index values, which e.g. describe the nutrient or the water balance of a site, were calculated.

Besides the localisations of the observation points in terms of Gauss–Krüger coordinates, the resulting data set consists of dichotomous information about the occurrence of 14 tree species as well as of several ecological parameters:

ID	longitude	latitude	tree species 1	tree species 2	...	ecological parameters		
1			1	0				
2			0	0				
⋮			⋮	⋮				

TABLE 2.1.: *Data structure.*

The environment of the habitats is recorded through different ecological parameters, which can be divided into four groups: water (H), energy (E), nutrient (N) and geomorphodynamic (G) parameters. Table 2.2 gives an overview of the ecological variables of the WINALP data set.

category	**abbreviation**	**variable description**	**spruce/ Swiss pine**	**ash**
H	AWC1M	available water capacity (1m depth)	×	×
	HYD_UNIT	water logging level (categorial)		
	P_JJA	precipitation summer	×	×
	STAUTXT	water logging index (cat.)	×	×
	TWI10	topographic wetness index		
E	G05_20	degree value days		
	T01_20	temperature January	×	×
	T_JJA	temperature summer	×	×
	R_JJA	radiation summer	×	×
	SAFI	slope aspect favourability index		
	ASPECT10	exposition		
N	TGBS	depth gradient of base saturation (cat.)		×
	CLAY1M	proportion of clay (1m depth)		
G	MORDYN	morphodynamics (cat.)		×
	SLOPE10	slope angle		
	CURV10	curvature		

TABLE 2.2.: *Overview of the ecological variables and their usage in the hypotheses–directed models.*

For the hypotheses–directed models, not all available ecological variables, but the worthwhile ones are included into the model to achieve ecologically comprehensible

relationships. Based on expert knowledge, only these parameters were chosen, whose impact on the occurrence of the individual tree species is well–known. Therefore, tree specific variable sets were developed, which are marked with a cross in the last two columns of the table above.

All 16 predictors, which are listed in table 2.2 are used for the data–directed models, because as much information as possible should be incorporated into the models in order to improve the predictive performance.

Because the available sample of the project contains presence–absence measurements of the response, statistical methods for presence–only approaches will be disregarded (for the analysis of presence–only data refer to Elith and Graham (2009a)).

2.2. Descriptive analysis

In this chapter, a short, descriptive outline of the occurrence of the examined tree species and the environmental conditions at the examined sites is given. From the total of 14 collected tree species, the habitats of the ash, the spruce and the Swiss pine are exemplarily analysed.

At 9.85% of the observed locations, the pioneer tree species ash was present. Frequently, the ash occurs in combination with the sycamore, the spruce and the beech. Especially in the eastern part of the study region the ash is increasingly found, whereas only few species are located in the western Bavarian Alps:

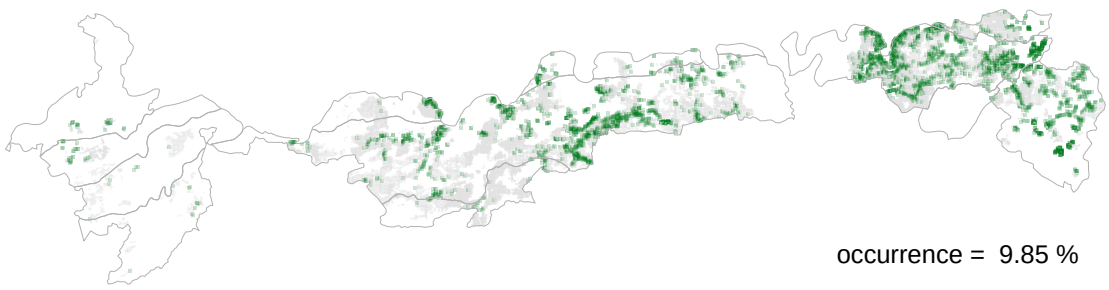


FIGURE 2.2.: *Spatial occurrence of the ash; marks for presences (green) are marked twice as large than absences (grey).*

The very prevalent spruce is situated in the most observation points (94.8%) with no spatial preference. This tree species grows in the environments of all investigated species, but at locations with a Swiss pine, spruces are monitored less often.

The Swiss pine is a very rare surveyed tree species in the Bavarian Alps. At only 85 of over 55,000 sites Swiss pines are found. Besides few, individually occurring species,

the Swiss pine exists in two small areas; one south of Garmisch–Partenkirchen and the other in the south–east of the study region.

As illustrated in chapter 2.1, a lot of possibly relevant predictors were collected. Nevertheless, the variation of the data could be explained with only a few topographical covariates, e.g. geographical position or altitude, because they can be seen as proxies for several physiological predictors. Since topographical information can be measured very precisely in comparison to physiological covariates, they tend to provide quite accurate models, even if they only indirectly influence vegetation. However, this is not the main objective in statistical ecology, because the transferability to other regions is not straightforward.

The temperature variables in the WINALP data set reveal spatial patterns through warm valleys and cold montane regions. Especially it is to note, that the eastern part of the study region shows slightly higher temperatures than the western regions:

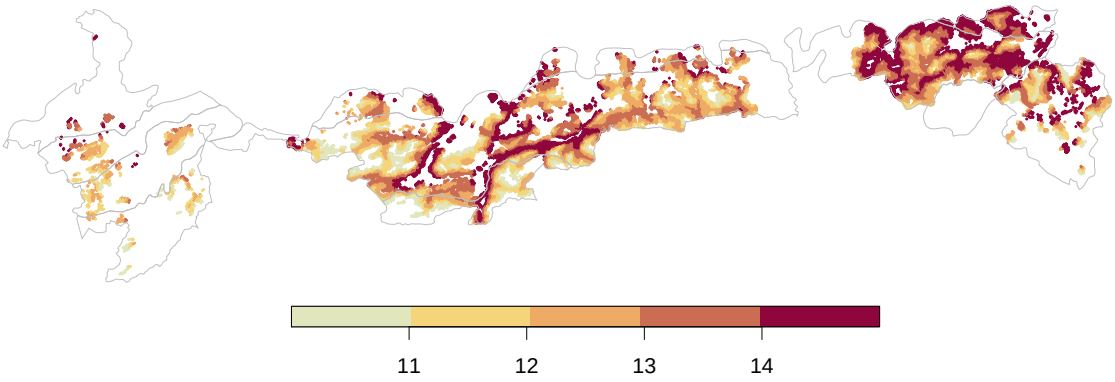


FIGURE 2.3.: *Spatial distribution of temperature (°C) in summer.*

The average temperature in summer is 13.09°C (sd=1.22) with a range from 8.63°C to 17.07°C. The temperature in January varies between -6.5°C and -0.59°C; the average is -2.79°C (sd=0.81).

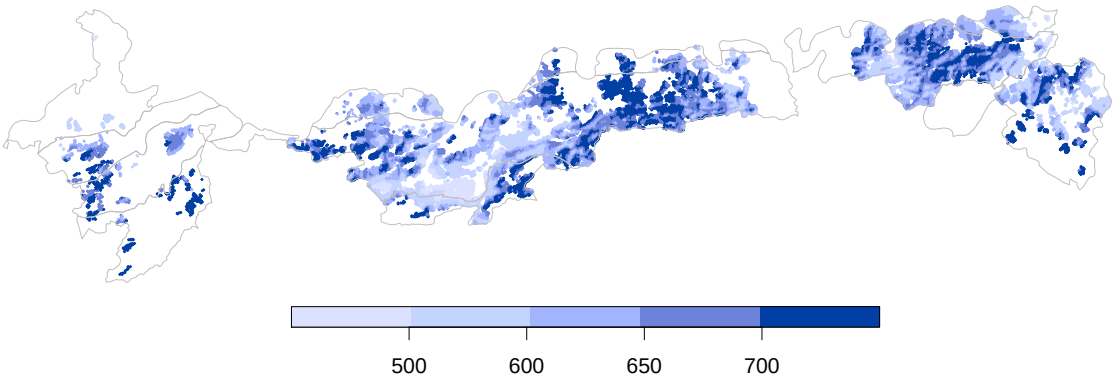


FIGURE 2.4.: *Spatial distribution of precipitation (mm) in summer.*

Furthermore, the precipitation exhibits a patchy pattern (cf. figure 2.4) and is, similar to most other predictors, a spatial acting variable. The average precipitation in the study region is 637.1mm (sd=74.7) and the measurements range from 409mm to 1046mm. The soil characteristics differ especially at the northern boundary, from the rest, because in that region, the soil has a lower base saturation and a higher stagnant moisture.

In figure 2.5 a descriptive overview of the – in experts' opinion – most important predictor variables for the individual tree species is given.

The spruce grows at the most of the observed locations and thus, it gets along with all environmental conditions of the Bavarian Alps. The ash differs from the spruce mainly in its energy request, because it prefers places with higher temperatures. Regarding water, nutrients and geomorphodynamics, spruces and ashes are growing under similar conditions.

The fact that the Swiss pine occurs especially in the upper altitudes of the Alps is also reflected in its environmental properties. This tree species grows at very low temperatures combined with a rather high radiation. Additionally, Swiss pines are found at locations with high base saturation and low water logging.

Among these predictors, the temperature in summer is correlated with the temperature in January ($r = 0.732$). A moderate correlation ($|r| < 0.6$) occurs between the predictors AWC1M, STAUTXT and TGBS, which describe the available water in the ground and the nutritional value.

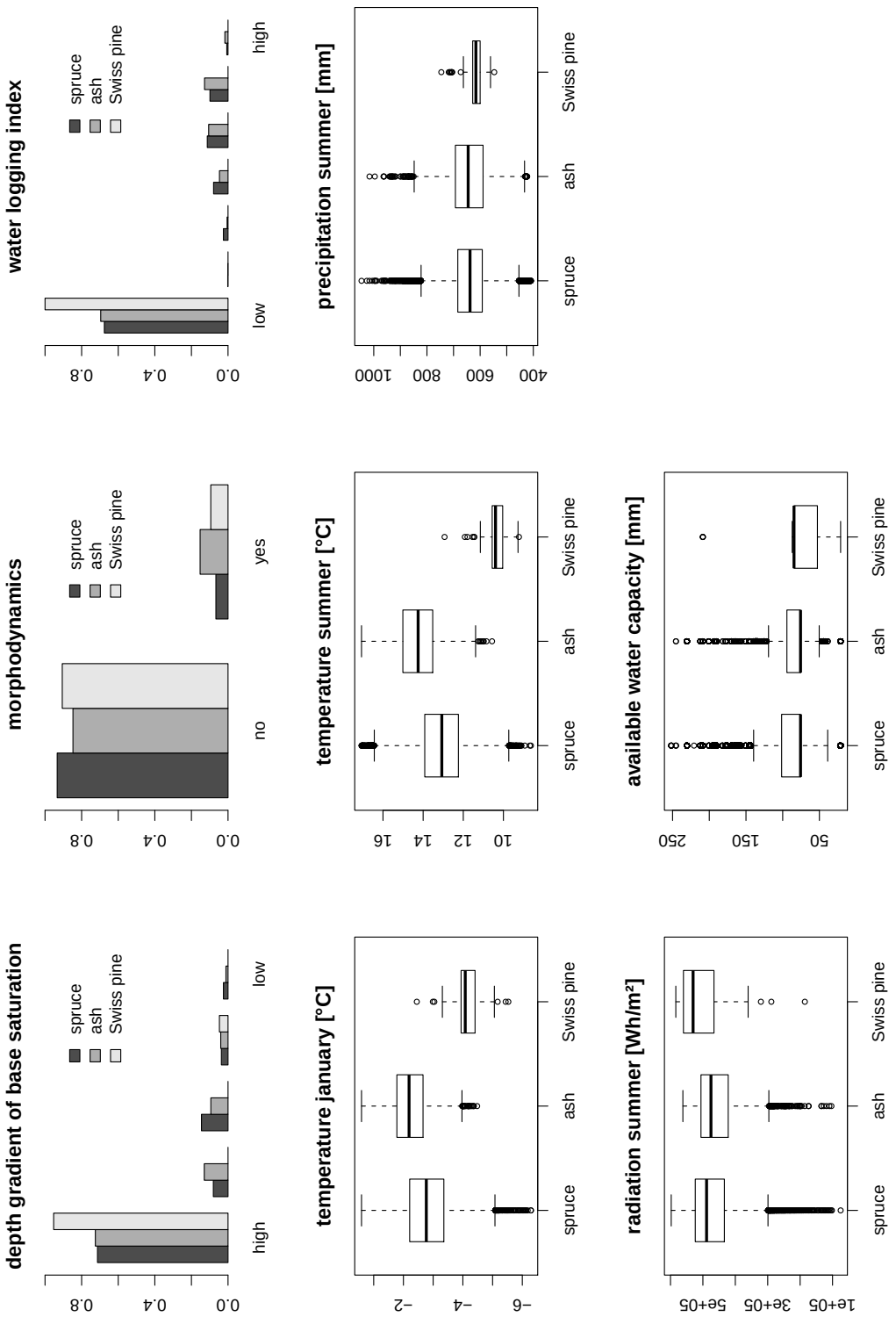


FIGURE 2.5.: *Descriptive predictor analysis for the different tree species.*

2.3. Analysis of the species response shapes along ecological gradients

In niche theory a distinction is drawn between a fundamental and a realised niche. Whereas the fundamental niche of a species describes the conditions, under which growing is possible, the realised niche denotes the actual conditions, under which the species occurs.

For a simple description of the complex ecological niche, in particular the realised niche of a tree species, the observed response patterns are analysed along individual ecological gradients, i.e. environmental predictors.

Contrary to former assumptions of unimodal, symmetric and bell–shaped curves, e.g. Gauch Jr. and Whittaker (1972), newer approaches, e.g. Austin (1999, 1987), suggest, that the shape of a curve is influenced by superior competitors of the evaluated species and by environmental stress. This results in many different possible shapes, e.g. skewed, plateau or bimodal curves.

Analysis strategies for the univariate description of the ecological niche include HOF–models (Huisman, Olff and Fresco, 1993), generalised linear models (GLMs) and generalised additive models (GAMs). The parametric HOF–models provide a hierarchic test procedure for the shape of the response curves, but Oksanen and Minchin (2002) showed the similarity between the results and GAMs. The advantage of GAMs is their flexibility in contrast to the restricted shapes of other approaches. The true, but unknown response curve, lies most likely within the calculated confidence region.

Figure 2.6 depicts univariate response curves of GAMs along a temperature and a precipitation gradient. The curves of the spruce along temperature and of the Swiss pine along precipitation show, that the ecological niche concerning these gradients is completely covered within the study region. The slightly bimodal shape of the Swiss pine has to be attributed to the few presences rather than to competition effects.

For some species the ecological niche is not totally explored and so it is impossible to model the full range of their distribution. For instance, the temperature gradient of the study region comprises the clear lower growing boundary of the ash, whereas the upper boundary is not included. Instead of that, the Swiss pine prefers the lower part of the gradient.

The truncated shape of the precipitation response curve of the ash indicates, that physiological limitations in terms of precipitation are not reached within the observed locations. Thus, only an extract of the real distribution is given.

These examples show, that even the diverse ecosystem of the Bavarian Alps does not encompass the full environmental scope, in which the tree species are able to live. Therefore it is to notice, that response curves based on restricted environmental gradients exhibit unrealistic shapes, especially at the edges of the gradient, and a reduced predictive suitability (Thuiller et al., 2004).

The analysis of species occurrence along environmental gradients provides a valuable tool for data description and an exploratory insight, but actually, the underlying processes, which govern the dispersion of a species, are more complex and relate to several variables or interactions with other predictors.

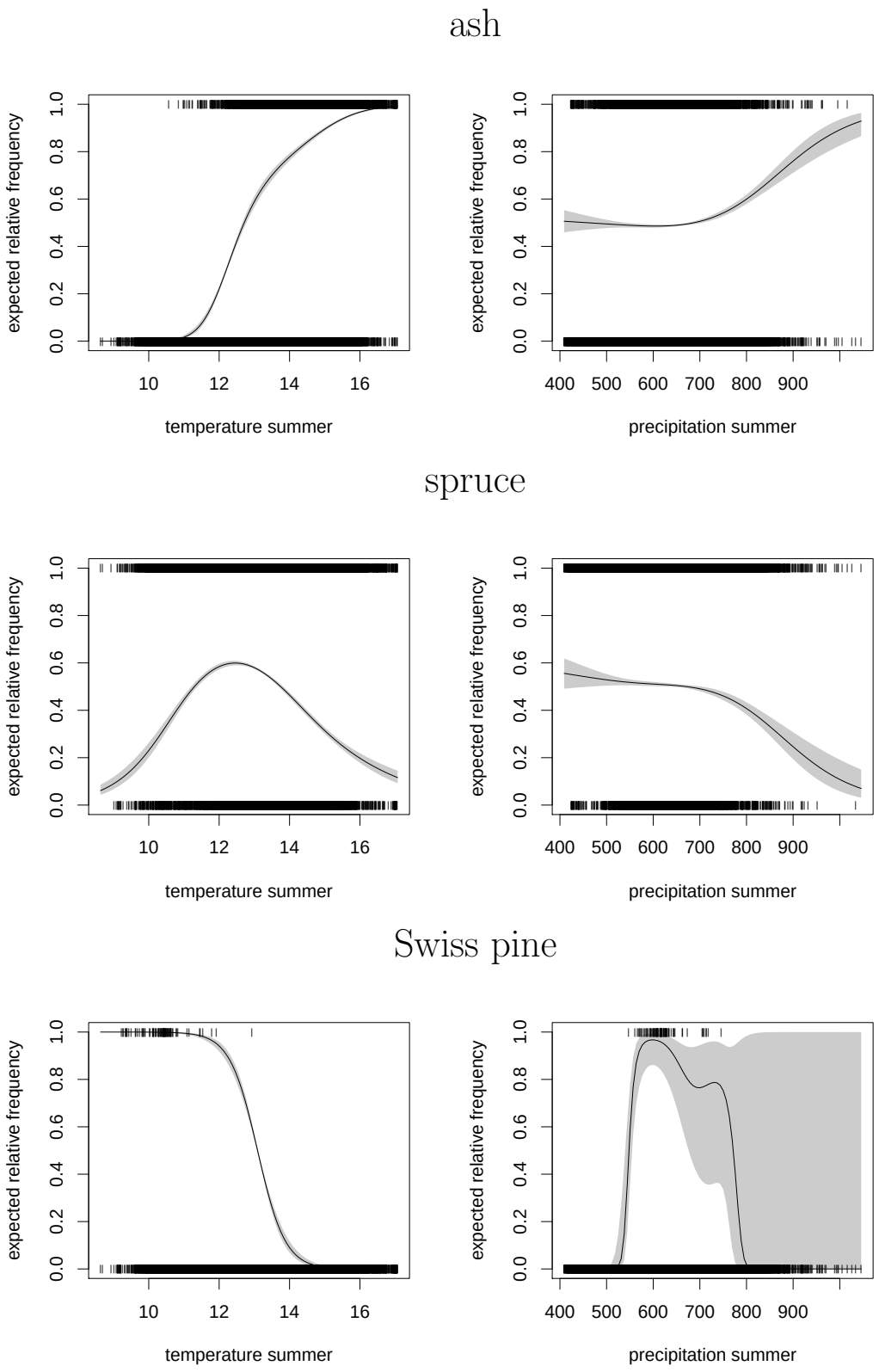


FIGURE 2.6.: *Response curves for ash, spruce and Swiss pine along the ecological gradients "temperature in summer" and "precipitation in summer".*

3. Hypotheses–directed approach using a generalised additive model

3.1. Model description

Generalised additive models (Hastie and Tibshirani, 1990) allow a flexible modelling of the relations between dependent and independent variables by extending the simple linear approach.

Since in niche modelling bell–shaped or linear response curves are unrealistic assumptions (e.g Austin and Smith (1989) and Austin (2002)), especially regarding the mechanisms of competition, flexible response shapes describe the ecological niche of a species more accurately. Even in comparison to generalised linear models with polynomial terms or with transformed variables, GAMs provide a better approximation of the true response surface, because they are not limited by the shape available from the predetermined model equation.

The first application of GAMs in an ecological context is found in the work of Yee and Mitchell (1991). The improvements compared to general linear models regarding the flexible modelling of species distributions are illustrated and the clear graphical interpretability as well as the predictive capacity is emphasised.

In recent years, GAMs have become an important and widely used tool for modelling species distributions (Guisan, Edwards and Hastie, 2002) and are often used as a reference procedure for the application of new and unexperienced methods, e.g. Moisen and Frescino (2002), Thuiller, Araújo and Lavorel (2003) or Leathwick, Elith and Hastie (2006).

A lot of effort is put in the improvement of generalised additive models. For instance, variable selection procedures, e.g. Wood and Augustin (2002), incorporation of interactions (Maggini et al., 2006), simultaneous models for several species (Araújo and Luoto, 2007) and the appropriate account for spatial autocorrelation, e.g.Maggini et al. (2006) can increase the model performance.

Some of these ideas are adopted by the analysis of the WINALP data with GAMs and their impact on model performance is examined in order to calibrate the models. Furthermore, a weighted GAM approach for the imitation of a balanced sampling design is investigated.

3.2. Estimation

At first, a short overview of the estimation procedure of generalised additive models is given focusing mainly on binomial response variables. GAMs are based on generalised linear models (McCullagh and Nelder, 1989), which are determined by the following three components.

1. Random component: Conditioned on the covariates $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})$ the density of the response variables y_i belongs to the mono–parametric exponential family $(i = 1, \ldots, n)$.
2. Systematic component: The structure of the linear predictor η_i is a linear combination of the covariates: $\eta_i = \sum_{k=1}^{p} \beta_k x_{ik}$.
3. The link function g defines the connection between the expected value $\mathbb{E}(Y_i) = \mu_i$ and the linear predictor: $g(\mu_i) = \eta_i$.

The linear predictor in generalised additive models contains not only linear, parametric terms, but also smooth, non-parametric and unspecified functions of the covariates, which are additively combined:

$$\eta_i = \sum_{k=1}^{p} \beta_k x_{ik} + f_1(x_{i1}) + f_2(x_{2i}) + f_3(x_{i3}, x_{i4}) + \ldots + \varepsilon_i$$

or respectively

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f}_1 + \boldsymbol{f}_2 + \boldsymbol{f}_3 + \ldots + \boldsymbol{\varepsilon} \quad .$$

The residuals ε_i are assumed to be independent and identically distributed with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Besides one–dimensional smooth functions of covariates, the linear predictor can include higher–dimensional, functional relationships, like f_3, which represent an interaction between two metric covariates. Imposing constraints for the identifiability of the smooth functions, for example $\sum_{i=1}^{n} f_j(x_{ij}) = 0, j = 1, \ldots, q$, determines the level of the functions.

To represent the smooth functions constructively, the use of a linear combination of basis functions is popular:

$$f_j(x_j) = \sum_{l=1}^{d_j} \gamma_{jl} B_l(x_j) \quad , \quad j = 1, \ldots, q \quad .$$

With different types of basis functions $B_l, l = 1, \ldots, d_j$, various types of splines, e.g. regression splines, in particular truncated power splines and B–splines, or smoothing splines, can be displayed. Natural cubic splines and their multi–dimensional analogon, the thin–plate–splines, are exemplarily described in detail below, because they are used for the analysis in chapter 6 and 7. For more information on other representations of the smooth functions refer to Fahrmeir, Kneib and Lang (2007).

The basis functions of natural cubic splines are chosen in a way, in which their linear combination $f(x)$ makes up a cubic polynomial spline. The knots are represented by the observations expanded through two boundary knots: $a \leq x_1 < \ldots < x_n \leq b$. Furthermore, $f''(a) = f''(b) = 0$ is presumed.

Natural cubic splines minimise the penalised least–squares criterion

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx \quad ,$$

with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q)$. Thus, they represent that function in the space of all possible smoothing functions with the smallest curvature. Also the penalty term can be demonstrated with the corresponding basis functions:

$$\int \left(f''(x)\right)^2 dx = \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_i \gamma_j \int B_i''(x) B_j''(x) dx = \boldsymbol{\gamma}^\top \boldsymbol{K} \boldsymbol{\gamma} \quad .$$

Generalising this optimisation concept into higher dimensions, leads to the so called thin–plate splines. In the two–dimensional case, they are assessed by minimising

$$\sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \int \int \left[\left(\frac{\partial^2}{\partial^2 x_1} + 2 \frac{\partial^2}{\partial x_1 \partial x_2} + \frac{\partial^2}{\partial^2 x_2} \right) f(x_1, x_2) \right]^2 dx_1 dx_2 \quad ,$$

with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_q})$.

The thin–plate regression splines can also be displayed through a basis function approach, in which radial basis functions are used. Due to the isotropy of these basis functions, they are especially appropriate for modelling two–dimensional smooth effects of geographical coordinates (Wood, 2006).

The consistent representation of smooth functions as a linear combination of basis functions results in a unified optimisation strategy for GAMs. The estimation is conducted either with the penalised least–squares criterion for normal distributed response

$$PKQ = (\boldsymbol{y} - \boldsymbol{B}_1 \boldsymbol{\gamma}_1 - \ldots - \boldsymbol{B}_q \boldsymbol{\gamma}_q - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{B}_1 \boldsymbol{\gamma}_1 - \ldots - \boldsymbol{B}_q \boldsymbol{\gamma}_q - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=1}^{q} \lambda_j \boldsymbol{\gamma}_j^\top \boldsymbol{K}_j \boldsymbol{\gamma}_j$$

or with the penalised log–likelihood

$$l_{\text{pen}} = l(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots \boldsymbol{\gamma}_q) - \frac{1}{2} \sum_{j=1}^{q} \lambda \boldsymbol{\gamma}_j^\top \boldsymbol{K}_j \boldsymbol{\gamma}_j$$

for the generalised case. $\boldsymbol{B}_j$ denotes a matrix, whose entries are the basis functions for covariate j evaluated at the observations:

$$\boldsymbol{B}_j = \begin{pmatrix} B_{j1}^l(x_1) & \cdots & B_{jd}^l(x_1) \\ \vdots & & \vdots \\ B_{j1}^l(x_n) & \cdots & B_{j1}^l(x_n) \end{pmatrix} \quad .$$

Simple GAMs are estimated either with the penalised least–squares estimator or with the Fisher Scoring algorithm, respectively (Fahrmeir, Kneib and Lang, 2007). Multiple generalised additive models require more complex methods.

Besides the common backfitting algorithm (Hastie and Tibshirani, 1990), in which the degree of smoothness is difficult to estimate (Wood and Augustin, 2002), Marx and Eilers (1998) suggest a penalised iteratively re–weighted least squares approach for direct and simultaneous modelling of the smooth components of a GAM. Therefore, the penalised log–likelihood is approximated by

$$\left\| \sqrt{\boldsymbol{W}^{(k)}} \left(\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{\eta} \right) \right\|^2 + \lambda_1 \boldsymbol{\gamma}_1^\top \boldsymbol{B}_1 \boldsymbol{\gamma}_1 + \lambda_2 \boldsymbol{\gamma}_2^\top \boldsymbol{B}_2 \boldsymbol{\gamma}_2 + \dots \quad , \tag{3.1}$$

near by the actual parameter estimate.

$\boldsymbol{W}^{(k)}$ denotes the diagonal matrix of the IRLS–weights in iteration k. The elements are calculated with

$$w_i^{(k)} = \frac{1}{V\left(\mu_i^{(k)}\right) g'\left(\mu_i^{(k)}\right)^2} \quad ,$$

in which $V(\cdot)$ determines the variance function of the corresponding exponential family and $g(\cdot)$ denotes the link function. $\tilde{\boldsymbol{y}}$ is a vector of pseudo–data which is given by

$$\tilde{y}_i = g'\left(\mu_i^{(k)}\right)\left(y_i - \mu_i^{(k)}\right) + \hat{\eta}^{(k)} \quad .$$

The model is estimated iteratively by repeating the following steps:

Algorithm 1: Penalised iteratively re–weighted least squares algorithm (P–IRLS)

1. Update the weights $\boldsymbol{W}$ and pseudo–observations $\tilde{\boldsymbol{y}}$ with the actual value of the parameter estimate.
2. Solve the minimisation problem 3.1 with $\boldsymbol{W}$ and $\tilde{\boldsymbol{y}}$ from step 1 to achieve an updated estimation for the parameter vector $(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q, \boldsymbol{\beta})$.

3.3. Model selection and variable importance

The parameters $\lambda_j, j = 1, \dots, q$, control the degree of smoothness of the functional components. In order to prevent overfitting as well as estimating a too general model (bias–variance tradeoff), the smoothing parameter $\boldsymbol{\lambda}$ has to be chosen appropriately. In the case of a known scale parameter σ^2, the unbiased risk estimator (UBRE)

$$\text{UBRE} = \frac{1}{n} \left\| \sqrt{\boldsymbol{W}}(\tilde{\boldsymbol{y}} - \boldsymbol{\eta}) \right\|^2 - \sigma^2 + \frac{2}{n} \text{tr}(\boldsymbol{H}) \sigma^2$$

is proposed as an optimisation criterion. If the scale parameter is unknown, the generalised cross validation criterion (GCV)

$$\mathrm{GCV} = \frac{n \left\| \sqrt{\boldsymbol{W}}(\tilde{\boldsymbol{y}} - \boldsymbol{\eta}) \right\|^2}{[n - tr(\boldsymbol{H})]^2}$$

will be used (Wood, 2006). Thereby, the hat matrix is denoted by $\boldsymbol{H}$ and depends on the smoothing parameter $\boldsymbol{\lambda}$. GCV is based on the leave–one–out cross–validation criterion, which is calculated with the sum of the squared prediction errors, when each observation is predicted by a model assessed by the rest of the data.

The smoothing parameter vector can be updated either in each iteration step or it is estimated with an additional outer loop. The first alternative is indeed more effective, but, besides other disadvantages, it cannot be guaranteed in particular that the overall best smoothing parameter is chosen (Wood, 2006).

Model selection is often closely related to the appropriate selection of covariates for the explanation of the response variable in a regression context. This important issue is now gaining support also in the modelling of ecological issues (Johnson and Omland, 2004; Reineking and Schröder, 2006).

Wood (2009a) proposes to compare GAMs with different sets of covariates, but within the same distribution family, concerning the GCV or UBRE scores as "the most logically consistent method", which is also recommended by Guisan, Edwards and Hastie (2002) in an ecological context.

The UBRE–criterion provides the same results as the AIC; furthermore, a variable selection with GCV differs marginally from a selection with UBRE/AIC, but in the case of binomial models the results should be compared (Wood, 2006). Covariates for removal are determined according to their p–value.

To evaluate the relevance of a single term, the p–value is used as an indicator for the distance to the null effect. For smooth components, the test statistic to the hypothesis

$$\mathbb{E}\left(\hat{\boldsymbol{\gamma}}_j\right) = \mathbf{0} \quad ,$$

which is calculated with the squared Mahalanobis distance of $\hat{\boldsymbol{\gamma}}$ from zero, determines the p–value. Thus, under the null hypothesis the p–value shows how strong the coefficients of the smooth term deviate from null.

An alternative possibility of model selection represents the usage of a shrinkage component, which is imposed on the smoothness penalty term.

A further possibility for the model selection procedure in generalised additive models with GCV or UBRE is the adaptive backfitting approach of Hastie and Tibshirani (1990), BRUTO, which can also be extended to binomial response values (Leathwick, Elith and Hastie, 2006). Other criteria, like AIC, BIC and Cross Selection are considered in Araújo and Luoto (2007).

The variable selection in regression models is well–suited for introducing ecological expert knowledge in order to heighten plausibility. This proceeding is confirmed through the findings of Meynard and Quinn (2007), who propose expert–based variable selection, because automated procedures lack selecting the relevant variables out of several correlated predictors.

3.4. Approach of a pseudo–balanced design by weighting observations

Since ordinary GAMs are often used in ecological modelling, some ideas for their amelioration are taken into account.

Comprising the maximum variability of the main predictors in the habitat of the examined species is crucial for the success of species distribution modelling. This requirement relates closely to the appropriate choice of the utilised sampling strategy.

Hirzel and Guisan (2002) showed with a simulation study, that equal and random sampling along stratified environmental gradients tends to provide models with more accurate predictions than a random sampling design. In particular, the extremes of the gradients are represented more strongly with this approach, and thus, according to Mohler (1983), the species distribution models can be improved.

Since the WINALP data set is constructed out of two databases, the design is a random pattern. To imitate a balanced design along environmental gradients, an extended GAM approach with weighted observations for a subsequent stratification is regarded. In comparison to a stratified subsample, this proceeding has the advantage of using the entire information of the data set.

The weights for the individual observations along a single gradient $\boldsymbol{x}_j, j = 1, \ldots, q$, are determined through the quartiles:

$$w_{ij} = \begin{cases} \frac{N}{n_j^{0.25}} , & \text{for } x_{ij} \text{ within the first quartile of } \boldsymbol{x}_j \\ \frac{N}{n_j^{0.25-0.75}} , & \text{for } x_{ij} \text{ within the second or the third quartile of } \boldsymbol{x}_j \\ \frac{N}{n_j^{0.75}} , & \text{for } x_{ij} \text{ within the forth quartile of } \boldsymbol{x}_j \end{cases}$$

$n_j^{0.25}$ denotes the number of observations within the first quartile of gradient $\boldsymbol{x}_j$, $n_j^{0.25-0.75}$ the number within the second and the third quartile and $n_j^{0.75}$ the number within the fourth quartile, respectively. Thus, the weights of observations within the middle quartiles of the gradient are half of those of the external quartiles.

The weight for each observation $i, i = 1, \ldots, n$, along several gradients is calculated as the product of the individual weights:

$$w_i = \prod_{j=1}^{q} w_{ij} \quad .$$

Since difficulties occur with the interpretation of confidence regions of weighted models, the weights are so scaled that they sum up to n.

The described weighting procedure increases the influence of observations, which are located at the boundary of the predictor space. With the WINALP data it will be investigated to what extent this strategy contributes to the improvement of the model.

3.5. Integration of an interaction factor

According to Austin (2002), the negligence of interactions between predictors restricts ecological modelling. Therefore, multi–dimensional effects of environmental covariates should also be incorporated into the model equation. However, this proceeding significantly enhances the complexity of the model and the identification of meaningful interaction terms from all possible ones is highly elaborate. Furthermore, the level of interaction has to be determined.

A possible solution to the integration of interaction terms are multivariate adaptive regression splines (MARS), which were introduced by Friedman (1991). This approach is a generalisation of regression trees, described in section 4.2. The fit in each node is a continuous function. The algorithm contains an implicit data–directed choice of the number of basis functions, the degree of interaction terms and the knot locations. Moisen and Frescino (2002) used this approach to the modelling of forest characteristics.

Maggini et al. (2006) applied a very simple, but efficient solution to the simultaneous integration of the most important predictor interactions within the commonly used linear or additive regression models. A qualitative interaction factor is included as covariate. The categories of this factor variable are defined by the paths of a regression tree, in which the residuals of the initial regression model are fitted with the ecological variables. The complexity of the regression tree is limited by pruning.

This appealing method, which adaptively accounts for relevant interactions, can be useful for the improvement of the predictive performance of a baseline model.

3.6. Spatial autocorrelation

Autocorrelation occurs when dependencies among different observations are persistent. Spatial and also temporal autocorrelated data is very common in ecology, e.g. (Legendre, 1993), because often, observations are collected across geographic space or along time series, respectively. Since the WINALP data set does not contain a temporal structure, only spatial autocorrelation is regarded.

Various reasons for spatial autocorrelation can be considered, which can affect the observations on a large scale as well as on a smaller scale. Besides spatially acting, ecological variables like water, energy, nutrients and geomorphodynamics, also biological processes, e.g. speciation, extinction, dispersal or species interactions (Dormann et al., 2007), and cultivation proceedings induce spatial structure in the data.

Disregarding a contributing predictor Z results in the violation of the independence assumption regarding the error term $\boldsymbol{\varepsilon}$ and leads to biased estimations. Since no information about the structure $s(Z)$ of the unobserved predictor is available except the position of the observations, a standard method to allow for Z is the inclusion of a smooth effect of the position, e.g. a smooth interaction between longitude and latitude:

$$\eta_i = \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{j=1}^{q} f_j(x_{ij}) + s(\text{longitude}_i, \text{latitude}_i) + \tilde{\varepsilon}_i \quad .$$

Therefore, it is assumed, that $\mathbb{E}(\varepsilon_i) = s(z_i)$ and $\text{Var}(\varepsilon_i) = \sigma_i^2$ so that $\mathbb{E}(\tilde{\varepsilon}_i) = 0$ and $\text{Var}(\tilde{\varepsilon}_i) = \sigma^2$ holds $(i = 1, \ldots, n)$.

Not only unobserved confounders, but also the resolution of the measuring scale may be a cause for misspecification, because in the case of a too high resolution, long–range, continental effects are not displayed in the data and otherwise a too small resolution lacks describing small–scaled patterns. Kühn (2007) describes this aspect and proposes the allowance for spatial autocorrelation as a way of alleviating misspecification.

3.7. Implementation

For the analysis with GAM, a reduced predictor set (cf. table 2.2), which was predetermined by ecology experts and which varies for the different tree species, is used, mainly in order to prevent multicollinearity. Merely the main effects of these environmental predictors are included to obtain sparse and comprehensible models.

The binary presence–absence response variables are modelled by using the binomial family. The smooth components are constructed with thin–plate regression splines as basis functions. The degrees of freedom are restricted to at most 4 in order to get reasonable and smooth response shapes. Categorial predictors are included either as factors or with a linear trend. Isotropic spatial autocorrelation is accounted for with a bivariate smooth surface, modelled with an interaction between longitude and latitude.

The `gam` function from the package `mgcv` (Wood, 2009b) is used for the model calculations, which implements the penalised likelihood approach. To compare several models in terms of their GCV–score adequately, the smoothing parameter selection is accomplished with an outer loop in addition to the ordinary P–IRLS algorithm, which provides the best GCV as possible.

The impact of allowing for the spatial structure, weighting observations and integrating an interaction factor is examined with the full, unselected models by means of the

GCV criterion. Product–weighting (cf. section 3.4) is accomplished with all covariates besides T01_20 and MORDYN, because of correlation with T_JJA and respectively, because MORDYN is binary. To determine the interaction factor, a regression tree is calculated on the residuals of the model until an assessable depth with conditional inference trees (cf. section 4.3) implemented in the function `ctree` of the package `party` (Hothorn et al., 2009).

A stepwise backward variable selection procedure is adopted, in which the covariate with the highest p–value is removed at each stage until the GCV–score strongly increases. The order of exclusion serves as a rough guide for the variable importance in the GAM approach. To identify the best model the statistical relevance of a predictor in the variable selection procedure as well as expert knowledge on the plausibility of the resulting response curves and predictions are taken into account. For more reasonable models, the degrees of freedom are adjusted in some cases.

3.8. Profits and limitations of the GAM approach

Generalised additive models are a powerful and efficient tool for the flexible modelling of smooth effects of several covariates simultaneously. They can detect several shapes including e.g. bimodal or skewed ones. Especially the quantification of uncertainty of the resulting estimation, e.g. with confidence regions, is appealing and allows statistical inference.

However, assuming additive covering of the single effects and a smooth response surface restricts the functional form of the modelled relationship. Moreover, the independence assumption of the model is often not given and techniques to alleviate this violation are necessary.

Especially in the context of ecology, GAMs produce unrealistic estimations at the boundary of a predictor, if the gradient is truncated, as mentioned in section 2.3.

Another weakness of regression models relates to the integration of interaction effects and correlated predictors. Containing all main and interaction terms of interest the model is often affected by identification and multicollinearity problems and consequently, the estimation fails or is instable.

4. Data–directed approach using random forests

4.1. Model description

Although stochastic data models, like e.g. regression models, sometimes lack fitting the data well, they are, due to simplicity and interpretability, commonly used. Whereas data models try to imitate the underlying mechanism of the data, algorithmic approaches, like neural nets, random forests or support vector machines, directly specify the functional relation between the predictors and the response (Breiman, 2001b).

The flexibility of machine learning methods allows the modelling of complex and also nonlinear relations and patterns as they occur in ecology. In comparison to conventional, parametric modelling techniques, these approaches have no need for restrictive assumptions on the relationships among response and predictor variables and they often outperform standard models. Within this work random forests will be described and applied in detail, as one possible data–directed strategy to model species distributions.

Random forests (Breiman, 2001a) base on classification and regression trees (CART, Breiman et al., 1984), which have been frequently used in recent years even in ecology. An introduction to the usage of CARTs for ecological questions is given by Olden, Lawler and Poff (2008). De'ath and Fabricius (2000) emphasise the simplicity and the convenient interpretation of the powerful CART algorithm in contrast to traditional linear regression, mainly with regard to variable selection and nonlinear modelling of multiple interactions in complex ecological data sets.

The random forest algorithm is, like boosting, which is described in detail in section 5, an ensemble method. Therefore, several models are combined in order to improve the predictive performance. In the case of random forests, an ensemble of decision trees is created by bootstrapping or subsampling.

Whereas random forests are well established in other subjects, they are barely applied in ecology. Prasad, Iverson and Liaw (2006) and Lawler, White and Blaustein (2006) detected a superior predictive capacity of random forests, applied to the modelling of tree and mammal species, in comparison to CARTs, bagging trees, generalised regression models, multivariate adaptive regression splines (MARS) and artificial neural networks.

Analysing different ecological presence–absence data sets, Cutler et al. (2007) explored the classification accuracy and the variable importance of random forests in contrast to classification trees, additive logistic regression and linear discriminant analysis. The superiority of random forests in the case of strong interactions among variables is pointed out, while the advantage is only moderate when additive structures are modelled.

The simulation study of Elith and Graham (2009a) discovered the superior performance of boosted regression trees and random forests in contrast to generalised linear models in most instances. Furthermore, advice for the appropriate application of the examined modelling techniques for specific use is given.

4.2. Classification and regression trees

Classification and regression trees, introduced by Breiman et al. (1984), are a non–parametric approach, because the relationship between response and predictor variables has not to be prespecified.

With the CART method special decision trees for categorial and continuous response variables, which only allow binary splits into mutually exclusive, preferably homogeneous groups, are created. This algorithm is not limited, because every multiway split can be represented by a series of binary splits. The other, very popular decision tree algorithm C4.5 accomplishes additionally multiway splits. Detailed information can be found in Quinlan (1993).

The resulting tree can be depicted graphically and therefore, the CART–model is not only suitable for prediction, but also for data description.

The CART procedure can be divided into three basic steps (Olden, Lawler and Poff, 2008):

1. tree building
2. stopping the tree building process
3. tree pruning and optimal tree selection.

For tree building CARTs subdivide the feature space recursively and parallel to the coordinate axes. In each resulting section $R_m, m = 1, \ldots, M$, the response variable is modelled with the averaged response value c_m in the regression situation:

$$\hat{f}(\boldsymbol{x}) = \sum_{m=1}^{M} \hat{c}_m \mathrm{I}(\boldsymbol{x} \in R_m) \quad .$$

I represents the indicator function. In the context of classification the fit is determined by majority decision.

With the resulting stepwise constant function, arbitrary association rules, e.g. linear or polynomial ones, can be approximated. Besides accounting for relevant main effects, CARTs are able to select important interactions implicitly, even though they are possibly high dimensional.

The splitting variable and the position of the cut is achieved by reducing an information measure of node impurity locally optimal, e.g.

- the sum of squares: $\frac{1}{|R_m|}\sum_{x_i \in R_m}(y_i - \hat{c}_m)^2$

in the regression case and

- the misclassification error: $1 - \max_k \hat{p}_{mk}$,
- the Gini Index: $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$ or
- the Shannon Entropy: $\sum_{k=1}^{K} \hat{p}_{mk}\log(\hat{p}_{mk})$

within classification problems. Therefore, $p_{mk} = \frac{1}{|R_m|}\sum_{x_i \in R_m} I(y_i = k)$ holds; $\{1, \ldots, K\}$ are the classes. The cut–point, which maximises the homogeneity of the resulting groups regarding the response, is selected. This procedure results in a decision tree, whose branches represent the splitting rules.

Node splitting will be recursively repeated until

- a minimum number of observations in the terminal node is reached,
- impurity reduction is too small,
- the terminal node is pure or if
- a different stopping criterion is fulfilled.

The hierarchical structure of the decision tree implicates the modelling of interactions between the predictor variables.

In order to avoid overfitting, the optimal tree size, which describes the complexity of the model, has to be determined. Hastie, Tibshirani and Friedman (2001) suggest a strategy of cost–complexity pruning for that reason. Thereby, the goodness of fit of a tree T, which is measured by its loss $\mathrm{L}(T)$, e.g. with the Gini Index, is penalised by the complexity of the tree, measured with the number of terminal nodes $|T|$:

$$\mathrm{L}_\alpha(T) = \mathrm{L}(T) + \alpha|T| \quad .$$

Minimising this cost–complexity criterion for several α–values leads to a sequence of smallest subtrees of the full tree $T_0 \supseteq T_1 \supseteq T_2 \supseteq \ldots$. The best value of the complexity parameter α is determined by cross–validation.

The tree building procedure can be modified and extended to some special options:

- lower limitation of the minimum node size
- definition of a priori class weights
- handling of missing predictor values with surrogate variables.

For that purpose, Hastie, Tibshirani and Friedman (2001) provide further information.

In particular, missing values often appear in ecological data. Whereas many modelling techniques cannot include missing values of predictor variables in the estimation, CARTs provide so called surrogate variables. A surrogate variable is a predictor, which produces a split similar to the split of the primary splitting variable.

However, the classification and regression trees produce not necessarily the best possible tree. Converging to a local optimum instead of the global optimum, the CART algorithm, like other greedy algorithms, can be suboptimal and moreover instable. A strategy to overcome this shortcoming is presented in section 4.4.

4.3. Conditional inference trees

Despite the instability of decision trees, the suboptimality of the final decision tree and the rawness of the predicted surface (Hastie, Tibshirani and Friedman, 2001), another limitation of the CART algorithm is the variable selection bias (Strobl, Boulesteix and Augustin, 2007). Variables with many possible splits or many missing values will be favoured in the selection of a splitting variable, if an ordinary entropy based measure, i.e. the Gini Index, is applied.

Conditional inference trees, which were proposed by Hothorn, Hornik and Zeileis (2006), are more appropriate for predictors measured at arbitrary scales, because they avoid the variable selection bias by means of accomplishing variable selection separated from the splitting procedure:

1. Variable selection: Test the global null hypothesis of independence between all covariates and the response. If this hypothesis is rejected, select the covariate with the strongest association to the response, e.g. the covariate with the lowest p–value resulting from an independence test. Otherwise, the recursion will be stopped.
2. Splitting procedure: Choose the best split of the selected variable by optimising an appropriate splitting criterion.

At the variable selection step, the global null hypothesis consists of m partial hypotheses:

$$H_0 = \bigcap_{j=1}^{m} H_0^j \quad \text{with} \quad H_0^j : D(Y|X_j) = D(Y) \quad , \quad j = 1, \ldots, m \quad .$$

Thus, it will be tested, if the conditional distribution $D(Y|X_j)$ of the response Y given the j–th covariate X_j depends on X_j. For each partial hypothesis a linear test statistic T_j is constructed, which measures the strength of the association between Y and X_j. For detailed statistical notation, refer to Hothorn, Hornik and Zeileis (2006).

Because the family of permutation tests provides unified tests for independence with regard to variables of arbitrary scales, this non–parametric approach is used for statistical inference. The general idea of permutation tests is the imitation of the distribution of the test statistic under the null hypothesis with random permutation of the considered response variable among the observations.

In the case of the independence tests, which are applied in the context of conditional inference trees, the situation under the j–th null hypothesis is generated by all possible permutations of the response Y, whereas X_j is fixed. Thus, the p–value of the realised

test statistic t_j can be determined and some method for the adjustment of multiple testing, e.g. Bonferroni correction, can be accomplished.

Despite the different scales of the predictors, an unbiased comparison of the m p–values is possible. If the minimum of the adjusted p–values does not exceed a prespecified level α, the global null hypothesis will be rejected and the covariate with the minimal p–value will be the splitting variable. Otherwise, further node splitting will be stopped. The parameter α can be seen either as significance level of the test of the global null hypothesis or as a hyper–parameter determining the tree size.

The permutation test framework also provides an approach to the splitting procedure, i.e. finding the optimal binary split: For each split, a linear test statistic is constructed, which measures the discrepancy between the splitted sets. The split, which causes the maximal, standardised test statistic, is chosen.

Thus, the association between the response and the potential splitting variable is measured with the test statistic of a formal test, in contrast to the classification and regression trees, where some measure of impurity reduction is applied.

Furthermore, an implicit solution to the overfitting problem is provided by the incorporation of the distribution of the test statistics, which avoids an additional pruning step.

4.4. Random forests

Random forests extend the decision tree approach and solve two essential limitations:

- Instability: Small changes in the data set can yield quite different trees, which induce a high variability, whereas the bias is low.
- Suboptimality: Locally optimal splits are not obliged to result in the global optimal tree.

The basic idea of ensemble methods, in particular random forests and boosting algorithms (cf. chapter 5) derives from the bias–variance decomposition for the prediction error (James and Hastie, 1997):

$$\mathrm{PE} = \mathrm{Var}(Y) + \mathrm{bias}^2(\hat{Y}, \mathbb{E}(Y)) + \underbrace{\mathrm{Var}(\hat{Y})}_{\sigma^2} \quad .$$

The prediction error PE of a model can be partitioned into an irreducible error $\mathrm{Var}(Y)$ and a reducible error, which consists of the squared bias $\mathrm{bias}^2(\hat{Y}, \mathbb{E}(Y))$ and the variance of the prediction $\mathrm{Var}(\hat{Y})$.

Ensemble methods in general aim at the reduction of the variance of the prediction σ^2 through model aggregation. If several i.i.d. models are averaged, e.g. B models, $\mathrm{Var}(\hat{Y})$ decreases according to the Law of Large Numbers to $\frac{\sigma^2}{B}$.

Since Breiman (2001a) showed, that the predictive accuracy improves with a decreasing correlation between the ensemble members, the construction of the ensemble should aim at a low correlation. Different ensemble methods vary in the manner how they generate several models.

Within the context of random forests, the optimal variance reduction in the case of an ensemble with B i.i.d. models is approximated by the following steps (Hastie, Tibshirani and Friedman, 2001):

Algorithm 2: Random forest

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample of size n from the training data.
   (b) Repeat the following steps recursively to build a tree on the bootstrapped data until the minimum node size is reached:
      i. Select m variables at random from the p covariates.
      ii. Pick the best variable/split–point among the m variables.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

Hence, variability is inserted into the ensemble with a random subset of the data, with a random selection of the predictor variables, which are available for each split and also with the omission of pruning. For regression, predictions are obtained by averaging the predictions over all trees; in a classification analysis, majority votes deliver the predictive values.

The tuning parameter m controls the diversity between the single trees and the bias–variance trade–off (Hastie, Tibshirani and Friedman, 2001). Increasing m leads to more correlated trees and thus to a higher variance, whereas the squared bias decreases. If m equals the number of predictors the procedure is called "bagging", which was introduced by Breiman (1996a, 1998).

Therefore, random forests improve the predictive accuracy, because they reduce the instability and suboptimality of a single tree, but otherwise the interpretability of a single decision tree gets lost.

4.5. Model selection and variable importance

A very appealing feature regarding the tuning parameter selection, the accuracy measures and the error rates of random forest models is, that model fitting and model validation can be accomplished all at once. This proceeding is possible, because only a subsample of the data is used for the fitting of a single tree, the remaining observations are used for out–of–bag predictions. Thus, an estimate for the out–of–bag error can be achieved, which approximates the generalisation error (Breiman, 1996b).

Especially in applications and for interpretability, not only the prediction, but also the mechanism behind a procedure, is concerning. For that reason a measure for the

variable importance is the matter of interest (Oppel, Strobl and Huettmann, 2009), as also applied in Prasad, Iverson and Liaw (2006) and Cutler et al. (2007). Four approaches are considered (Strobl et al., 2007, 2008):

- Selection frequency: number of selections of each variable by all individual trees
- Gini importance: (weighted) mean of decrease in Gini Index at every split produced by the variable
- Permutation importance: averaged difference in prediction accuracy before and after randomly permuting the covariate
- Conditional permutation importance: permutation importance with additional adjustment for previous splitting variables.

For detailed formula, refer to the denoted literature, because the support for explaining is not too valuable.

(Strobl et al., 2007) investigated that the variable importance measure for random forests based on the Gini Index is affected by the number of categories and the scale of measurement. This is to some extent due to the bias of the individual classification and regression trees in favour of variables with more categories as depicted in section 4.3. By the use of a forest based on conditional inference trees in combination with the permutation importance, this kind of bias decreases.

The idea behind the permutation importance is to simulate the situation of independence between response and predictor, which delivers a higher prediction error, if the predictor is important. However, the variable selection bias decreases, but does not vanish, because the bootstrap step in the algorithm also affects the variable selection preferring multi–categorial predictors (Strobl et al., 2007). The background of this is, that for statistical inference, in the individual conditional inference trees the distribution in the bootstrap sample is used instead of the distribution of the test statistic under the null hypothesis .

Since subsampling avoids this effect, it is more appropriate for random forests with predictors varying in their number of categories and in the scale of measurement. For further, more detailed reading refer to Strobl et al. (2007) and Bickel and Ren (2001).

Selection frequency and also the ordinary permutation importance were detected to overestimate the importance of predictors, which actually have just a minor influence, but correlate to a contributing covariate (Strobl et al., 2008). The reason is, that these measures depend on the marginal correlation, i.e. the correlation of the considered predictor with the response as well as with the previous splitting variables.

Strobl et al. (2008) suggest an importance measure based on conditional permutation, which adjusts for the unintended impact of the correlation with previous splitting variables more accurately than the permutation importance. For further information on the technique of conditional permutation, refer to the denoted literature.

Besides the univariate effect of a predictor variable, variable importance measures for random forests also account for the impact of high–dimensional interaction effects.

4.6. Implementation

In order to implement the data–directed approach with random forests the 16 environmental predictor variables, which are described in chapter 2.1, are used for modelling.

The random forest approach is conducted with the `cforest` procedure of the package `party` (Hothorn et al., 2009), in which a random forest algorithm with conditional inference trees is implemented.

Besides the construction of unbiased decision trees, subsamples of the size $0.632 \cdot n$ instead of bootstrapping are used for drawing random samples of the data set, in order to prevent the variable selection from bias, which is explained in the previous chapter.

The stability of the resulting model is based on the number of trees `ntree` and the number of predictors, which are available for each split, `mtry`. Though, the selection of the tuning parameters is not as crucial as in other modelling techniques.

For stable results as many trees as possible are included in the forest and thus, 300 trees are computed for each model.

Although the conditional variable importance measure would be preferable because of the correlated predictors, the calculation thereof is not possible due to the high number of observations in the WINALP data set. However, according to Strobl et al. (2008), similar values to the conditional importance will be obtained with the ordinary permutation importance, if `mtry` is increased. Therefore, `mtry`= 7 is chosen.

Lin and Jeon (2006) showed, that the predictive capacity of random forests depend, particularly in the case of many observations coexistent with few predictors, on the terminal node size. Thus, for the limitation of overfitting, the optimal depth of the tree is chosen with regard to the out–of–bag AUC criterion based on random forests with 100 trees for the analysis of the data and with 50 trees in the simulation study.

The results are described by the variable importance measure as well as by partial dependence plots of the predictors, which are not available by default in the `party` package. A routine, based on the marginal average of the effect, was implemented as described by Hastie, Tibshirani and Friedman (2001) and applied in an ecological issue by Cutler et al. (2007). For the partial dependence of variable X_j, predictions are accomplished for varying values of variable X_j by fixing the other covariates X_{-j} to their empirical values:

$$f_{\text{partial},j}(X_j) = \frac{1}{n} \sum_{i=1}^{n} f(X_j, x_{i,-j}) \quad .$$

4.7. Profits and limitations of the random forest approach

The random forest approach is a powerful tool for accurate predictions. They are characterised by a high classification accuracy and a flexible handling of different types of response. Averaging prevents from overfitting due to the Law of Large Numbers. Also the robust algorithm can handle correlated predictors with complex and nonlinear interactions, even if the number of predictors exceeds the number of observations.

In comparison to decision trees, random forests are more appropriate for modelling functional relations and can approximate arbitrary decision boundaries. An appealing side effect of averaging is the fact, that the raw, piecewise constant decision boundaries of an individual tree are smoothed. Also the limited assortment of predictor variables at each split accounts for additional interaction effects, because locally suboptimal splits are generated.

Another feature of random forests is their treatment of missing values with surrogate variables. Thus, all observations are included in the modelling procedure.

Additionally, the calculation of variable importance is provided, which can be used for the identification of relevant predictor variables . The random forest algorithm offers an unbiased variable selection and implicit pruning of each tree based on statistical inference by utilising conditional inference trees in combination with subsampling.

Since the random forest algorithm renders the calculation of the out–of–bag prediction error, the estimation of predictive performance measures comes along with model fitting. This error is often used as a lower bound for the prediction error of other modelling strategies.

Because random forest is a non–parametric modelling technique, it is not affected by limiting distribution assumptions, e.g. it does not require uncorrelated predictors. However, a limitation of random forests is, that this approach is not based on a probabilistic model. Consequently, a probability statement on the accuracy of the prediction, e.g. a confidence interval, is not provided by the estimation procedure and has to be calculated with additional techniques, like e.g. bootstrapping.

Furthermore, the algorithm is a computationally intensive routine concerning computing time as well as computer resources, especially if partial dependence curves are calculated. The examination of the individual trees is often not meaningful and hence, the method is called a black box approach. The straightforward interpretability of the single regression tree disappears by applying a random forest.

5. Data–directed approach using boosting

5.1. Model description

Boosting, which connects statistical modelling to machine learning, is an ensemble method for improving the predictive performance of a single weak learning algorithm by an adaptive combination of the ensemble. In contrast to other ensemble methods, like i.e. random forests, boosting algorithms are aimed at the stepwise improvement of the performance. This is achieved through a strengthened or literally "boosted" emphasis on poorly fitted observations.

Originally, Freund and Schapire (1995) introduced boosting as a machine learning algorithm for binary classification problems and called the procedure "AdaBoost". Hastie, Tibshirani and Friedman (2001) extended this approach to a more general, statistical framework, in which many different baselearners, e.g. trees or component–wise linear procedures, can be used. Thus, also the analysis of regression problems and further statistical issues, e.g. survival analysis, is possible with boosting.

For modelling presence–absence data with boosting, two approaches are considered:

- BRT: boosted regression trees
- GAMBoost: component–wise smoothing splines used as base procedure

Whereas BRT allows for interactions between the predictor variables more appropriately, GAMBoost has the advantage, that the modelled effects on the response are smooth.

In ecology, boosting methods have just recently attracted interest and are rarely applied at present. Elith, Leathwick and Hastie (2008) provide a broad guidance for the application of boosted regression trees in ecology and illustrate the usage by means of the distribution analysis of a fish species in New Zealand.

De'ath (2007) gives an introduction to boosted trees in ecology and discovers their superiority in terms of predictive accuracy compared to bagged trees, random forest and generalised additive models. Also Guisan et al. (2007) showed, that boosted regression trees perform better than a series of other modelling techniques, e.g. GAM, BRUTO or MARS.

Examples for the application of boosted regression models, e.g. GLMBoost or GAMBoost are not available in literature. So the performance and the properties on species distribution modelling must still be investigated.

5.2. Estimation

5.2.1. Generic boosting algorithm

Boosting, as an ensemble method, aims at decreasing the prediction error by averaging (similar to random forests, which are described in chapter 4.4) over several models $g_m(x)$:

$$g(x) = \sum_{m=1}^{m_{\text{stop}}} \alpha_m g_m(x) \quad .$$

The individual models $g_m(x)$ result from the analysis of different re–weighted data versions with a proper statistical modelling technique, which is called base procedure or baselearner.

Apart from the fact, that the basic idea of boosting is not limited to decision trees as baselearners, the original conception of boosting, especially the AdaBoost algorithm (Freund and Schapire, 1995), predominantly differs from random forests in the generation of the ensemble in terms of two aspects:

The first difference is related to the construction of the base procedure. Instead of unbiased, highly overfitting models, a sequence of weak learners is produced to build up the ensemble. Kearns and Valiant (1994) proved that an ensemble of models, which are individually only slightly better than random guessing, achieves better predictions in conjunction.

Secondly, variation in the data sets, which are used for each single model $g_m(x)$, is not achieved by bootstrap sampling, but rather through varying observation weights. The modification of the weights for each ensemble member depends on the fit of the previous model. The weight of an observation is lowered in the case of a good previous fit, whereas a bad fit yields increased weights. This weighting procedure concentrates increasingly on observations, which are difficult to learn.

From a statistical point of view, the original AdaBoost algorithm, in which re–weighted versions of the data produce several models, can be seen as an optimisation problem for the expected loss L between the response variable and a function of the covariates (Friedman, 2001):

$$f^* = \underset{f}{\operatorname{argmin}} \mathbb{E}\left[L(Y, f(\boldsymbol{X}))\right] \quad .$$

Therefore, the function f is described as a linear combination of base procedures $b(\boldsymbol{x}; \boldsymbol{\gamma}_m)$ (Friedman, Hastie and Tibshirani, 2000)

$$f(\boldsymbol{x}; \{\nu_m, \boldsymbol{\gamma}_m\}_1^{m_{\text{stop}}}) = \sum_{m=1}^{m_{\text{stop}}} \nu_m b(\boldsymbol{x}; \boldsymbol{\gamma}_m) \quad .$$

The configuration of the parameter vector $\boldsymbol{\gamma}$, which characterises the basis functions, depends on the utilised base procedure. The minimisation problem is solved numerically

with the steepest-descent method (Friedman, 2001; Bühlmann and Hothorn, 2007):

Algorithm 3: Functional gradient descent algorithm

1. Initialise $\hat{f}_0(x)$, e.g. $\hat{f}_0(x) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, c)$.
2. For $m = 1, \ldots, m_{\text{stop}}$ iterate:
   a) Compute the negative gradient evaluated at the predictions of the previous iteration $\hat{f}_{m-1}(\boldsymbol{x})$:

$$\tilde{y}_i = - \left[\frac{\partial L(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)} \right]_{f(\boldsymbol{x})=\hat{f}_{m-1}(\boldsymbol{x})}, \quad i = 1, \ldots, n \quad .$$

   b) Choose the parameter vector $\boldsymbol{\gamma}_m$, with which the base procedure fits the negative gradient vector $\tilde{y}_1, \ldots, \tilde{y}_n$ at the best in terms of L_2–loss.
   c) Update the function estimate:

$$\hat{f}_m(\boldsymbol{x}) = \hat{f}_{m-1}(\boldsymbol{x}) + \nu b(\boldsymbol{x}; \boldsymbol{\gamma}_m)$$

   Parameter estimations of previous iterations remain steady.
3. Terminate, if $m = m_{\text{stop}}$.

The basic idea behind the algorithm is, that an approximate value for f^* is iteratively improved by a small modification through the addition of fitted baselearners towards the locally best solution. Also the original AdaBoost algorithm, in which the data is iteratively re–weighted, was shown to be a gradient descent algorithm (Breiman, 1999).

The demonstrated, stage–wise boosting algorithm constitutes a general framework for various applications. Depending on the choice of the base procedure and the loss function, structural assumptions for the data can be incorporated into the specific algorithm in order to ameliorate the predictive performance and the interpretation of the resulting model.

5.2.2. Loss functions

In step 2a) of the functional gradient descent algorithm, the negative gradient is calculated with the loss between the response variable and the corresponding fit of the previous iteration. The usage of appropriate loss functions, which results in $\tilde{y} \in \mathbb{R}$ for all types of responses, permits a uniform concept of the algorithm.

Whereas for regression problems, the loss is calculated with the residuals $y - f(x)$, the so–called "margin" $yf(x)$ is utilised in classification problems with responses $y \in \{-1, 1\}$. Because a positive margin value indicates a correctly classified observation and a negative value a misclassified one, an appropriate loss function should penalise the latter more strongly. Commonly used loss functions are for example (Bühlmann and Hothorn, 2007).

- a re–parameterised and scaled version of the negative binomial log–likelihood $L_{\text{log–lik}}(y, f) = \log_2(1 + \exp(-2yf))$,
- the exponential loss $L_{\text{exp}}(y, f) = \exp(-yf)$ or
- the hinge function (loss function for support vector machines) $L_{\text{SVM}}(y, f) = [1 - yf]_+$,

which are convex and differentiable approximations of the misclassification loss $L_{0\text{–}1}(y, f) = \text{I}_{\{yf \leq 0\}}$:

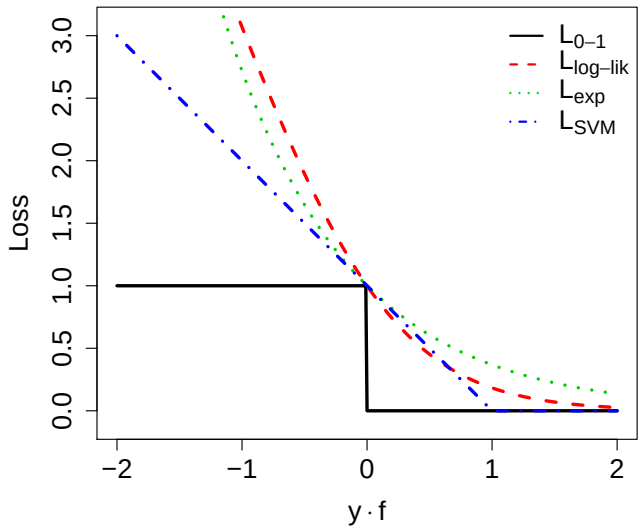


FIGURE 5.1.: *Loss functions for binary classification.*

The exponential loss function leads to the AdaBoost algorithm. Boosting with the negative binomial log–likelihood as a loss function is called BinomialBoosting. For both versions, the population minimiser is given by the half of the log–odds ratio of the class probabilities (Friedman, Hastie and Tibshirani, 2000):

$$f^*_{\text{log–lik}}(x) = f^*_{\text{exp}}(x) = \frac{1}{2}\log\left(\frac{p(x)}{1 - p(x)}\right) , \quad \text{with} \quad p(x) = \mathbb{P}(Y = 1|X = x) \quad .$$

On the other hand, the Bayes classifier minimises the expected misclassification loss and also the loss function based on the hinge function.

Besides the mentioned, monotone loss functions for binary classification problems, it is also possible to use the non–monotone L_1– and L_2–loss (scaled):

$$L_1(y, f) = |y - f| \quad \text{and} \quad L_2(y, f) = \frac{1}{2}\,|y - f|^2 \quad .$$

The L_1– and L_2–loss functions are mainly common for regression problems, but they can also be applied for classification problems. Their population minimiser is the Bayes classifier and the conditional expectation, respectively. Using the L_2–loss function yields the popular L_2–Boosting algorithm, in which the negative gradient vector is equal to the residuals. Thus, L_2–Boosting improves the estimation in each step by iteratively fitting a new, weak model on the resulting residuals. A further, robust alternative for regression is the Huber–loss function.

Because binary classification is in the centre of interest in this work, refer to Bühlmann and Hothorn (2007) for a detailed description of the regression case.

5.2.3. Base procedures

The general framework of boosting as an additive combination of several models allows a flexible specification of the base procedure. The choice depends on the structural characteristics of the relationship of interest: Trees are more appropriate for the modelling of the effects of interacting predictor variables. If a linear or an additive structure is suggested, component–wise linear models (GLMBoost) or respectively component–wise smoothing procedures (GAMBoost) are suitable baselearners.

The base procedures differ in the method of parameter estimation (cf. table 5.1):

	base procedure	parameter estimation
GLM	$\hat{b}(x) = \hat{\beta}^{\text{best}} x^{\text{best}}$	$\hat{\beta}^{(j)} = ((X^{(j)})^\top X^{(j)})^{-1}(X^{(j)})^\top \tilde{Y}_i = \sum_{i=1}^{n} X_i^{(j)} \tilde{Y}_i / \sum_{i=1}^{n} \left(X_i^{(j)}\right)^2$
GAM	$\hat{b}(x) = \hat{f}^{\text{best}}(x^{\text{best}}) = \sum_{l=1}^{d^{\text{best}}} \hat{\beta}_l^{\text{best}} B_l(x^{\text{best}})$	$\hat{\boldsymbol{\beta}}^{(j)} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left[\sum_{i=1}^{n} \left(\tilde{Y}_i - f\left(X_i^{(j)}\right)\right)^2 + \lambda \int f''(x))^2 dx \right]$
Tree	$\hat{b}(x) = \sum_{k=1}^{K} \gamma_k^{\text{best}} \text{I}\left(x \in R_k^{\text{best}}\right)$ $\left(\boldsymbol{\beta}^{\text{best}} = (\gamma_1^{\text{best}}, \ldots, \gamma_k^{\text{best}}, R_1^{\text{best}}, \ldots, R_k^{\text{best}})\right)$	$\hat{\boldsymbol{\beta}}^{(j)} = \underset{\gamma_1^{(j)}, \ldots, \gamma_k^{(j)}}{\text{argmin}} \sum_{k=1}^{K} \sum_{x_i \in R_k^{(j)}} L_2\left(\tilde{Y}_i, \gamma_k^{(j)}\right)$

TABLE 5.1.: *Examples for base procedures and the corresponding optimisation problems for parameter estimation (Hastie, Tibshirani and Friedman, 2001; Bühlmann and Hothorn, 2007).*

In each boosting step, simple models with the negative gradient $\tilde{Y}_i$ as dependent variable are calculated for all predictor variables (cf. the second column of table 5.1).

For the different base procedures, the model parameters $\boldsymbol{\beta_j}$ have different meanings: for GLMBoost, $\boldsymbol{\beta_j}$ is the ordinary regression coefficient and for GAMBoost, $\boldsymbol{\beta_j}$ represents the coefficients for the basis functions. For decision trees, the model parameters represent a vector, which includes the vector of subregions $R_1, \ldots, R_k$ of that region, which is chosen for splitting in iteration m, and the constant estimations $\gamma_1, \ldots, \gamma_k$ for

these subregions.

The parameter estimators for the GLMBoost and the GAMBoost are assessed with the method of the least–squares and the penalised least–squares, respectively. If a decision tree is chosen for the base procedure, a subtree in each existing region of the predictor space will be required. The calculation of the parameter estimators for each subtree results from the minimisation of the L_2–loss.

For each predictor or – in the case of a decision tree as baselearner – for each combination of predictor variables, the base procedure is fitted to the negative gradient. The best fitting base procedure $\hat{b}(\boldsymbol{x})$ of the actual boosting iteration m is added (in a shrunken version) to the function estimate $\hat{f}_{m-1}(\boldsymbol{x})$.

5.3. Properties of the boosting algorithm

In this section some outstanding properties of the boosting algorithm will be overviewed.

Firstly, using boosting with component–wise least squares or component–wise smoothing splines, the boosted model converges to the ordinary regression model and the regression coefficients of the boosted model are shrunken versions of the coefficients of the ordinary model.

A very attractive feature of boosting is the implicit and efficient variable selection. In each boosting iteration few predictors, which improve the model at best, are chosen for the extension of the model. Thus, unimportant covariates do not or do only little contribute to the model and indications for the importance of the individual predictors can be deviated, which is explained in detail in chapter 5.5. However, the mathematical properties of variable selection with boosting are not clear.

As explained in chapter 5.2.1, boosting iteratively fits the negative gradient vector. Hence, the approximate boosting hat matrix in iteration m, $\mathcal{B}_m$, can be calculated with the hat matrix of the component–wise estimator for the negative gradient vector

$$\mathcal{H}_m : (\tilde{y}_{1,m}, \ldots, \tilde{y}_{n,m}) \mapsto \hat{\tilde{y}}_{1,m}, \ldots, \hat{\tilde{y}}_{n,m} \quad .$$

In the case of BinomialBoosting, e.g., the approximate boosting hat matrix can be depicted as follows Bühlmann and Hothorn (2007):

$$\mathcal{B}_m = \mathcal{B}_{m-1} + 4\nu W_{m-1}\mathcal{H}_m(\mathrm{I} - \mathcal{B}_{m-1}) \quad \text{with} \quad \mathcal{B}_1 = \rho 4 W_0 \mathcal{H}_1 \quad .$$

W denotes a diagonal matrix, which is calculated with the predictors. This alternative notation of the boosting step allows the calculation of the degrees of freedom as the trace of the approximate boosting hat matrix and thus, the calculation of information criteria, like AIC or BIC.

It is also known, that boosting can be susceptible to overfitting, but it overfits very slowly. Bartlett and Traskin (2007) showed, that the early stop of the algorithm results in its consistency.

5.4. Model selection

Because poor generalisability of a model can result from overfitting to the training data set, regularisation methods are applied to discover a good fitting degree. Therefore, the boosting algorithm provides two alternatives: constraints on the number of components m_{stop} as well as on the step–size ν in each boosting iteration are possible (Friedman, 2001).

The optimal number of components m_{stop} is determined by some selection criterion, e.g. AIC, cross–validation or bootstrapping. It is to remark, that regularising the parameter m_{stop} implicitly assumes that sparse models perform better in prediction.

Furthermore, the step–size ν controls the generalisability of the fitted model. The parameter ν operates like a shrinkage parameter (Friedman, 2001) and hence, by decreasing ν, the variance is reduced on account of the bias.

This leads to contrary objectives: Choosing a small step–size ν results only in a slight improvement of the model, many boosting iterations are required and vice versa, if the size of the boosting steps is large, fewer iterations will be needed. Consequently, a trade–off between ν and m_{stop} occurs.

For small step–sizes, e.g. $\nu = 0.1$, the effect of this parameter on the predictive accuracy is negligible small (Friedman, 2001; Bühlmann and Hothorn, 2007). Thus, it is sufficient to search the optimal number of boosting iterations m_{stop} for a constant and small value of the step–size ν.

However, overfitting can result not only from the tuning parameters m_{stop} and ν, but also from the degree of complexity of the base procedure. Bühlmann and Hothorn (2007) demonstrate, that base procedures with low bias and high variance, e.g. high–order smoothing splines, lead to a poorly generalising model. Additionally, Bühlmann and Yu (2003) showed for $L2$–boosting, that the flexibility of a model does not need to be governed by the degrees of freedoms of the base procedure, because a high–order degree of smoothness can also be achieved with sufficient boosting iterations by using weak baselearners.

5.5. Variable importance

Especially in high–dimensional problems, in which even the number of predictor variables exceeds the number of observations, boosting is an appropriate tool for evaluating the importance of the different predictors. The variable importance measures depend on the utilised base procedure.

For boosted regression trees, Hastie, Tibshirani and Friedman (2001) propose the following importance measure. The importance of the predictor X_l, $I_l^2(T)$, of a single decision tree T can be assessed by adding up the squared impurity reduction i_t^2 in that inner nodes t, $t = 1, \ldots, J-1$, which were produced by predictor X_l:

$$I_l^2(T) = \sum_{t=1}^{J-1} i_t^2 \mathrm{I}\left(v(t) = l\right) \quad ,$$

where I denotes the indicator function. The importance measure for a boosted regression tree can be calculated by averaging:

$$I_l^2 = \frac{1}{m_{\text{stop}}} \sum_{m=1}^{m_{\text{stop}}} I_l^2(T_m) \quad .$$

Since up to now, an importance measure for the GAM–Boost algorithm does not exist, a possibility thereof is described in this paragraph. The importance of the predictor X_l can be quantified with the deviation of the partial response $f(x_l)$ from the null, which can be measured with the area under the normalised response curve:

$$I_l = \int_0^1 \left| \hat{f}(\tilde{x}_l) \right| \mathrm{d}\tilde{\mathrm{x}}_l \; , \quad \tilde{x}_l \in [0,1] \quad .$$

For the comparability of the measures between different predictors, the domain of the predictors is standardised on the interval $[0, 1]$.

Because variable importance measures are relative, the proportion of the cumulative importance is specified and interpreted.

5.6. Implementation

As mentioned in section 5.1, two different baselearners are used for the analyses with boosting: regression trees and component–wise smoothing splines.

Analogously to the analysis with the random forests, the data–directed approach with the boosting algorithms is modelled with the same 16 predictor variables, which have been described in chapter 2.1. The analysis with GAMBoost is conducted with the inclusion of the coordinates.

The boosted regression trees are calculated by means of the library `gbm` (Ridgeway, 2007), which implements the gradient boosting method with classification and regression trees as base procedure and with the deviance of the bernoulli distribution as loss function. The individual trees are grown until the depth is two. This approach is denominated with "GBM" below.

The GAMBoost models are fitted with the `gamboost` function of the library `mboost` (Hothorn et al., 2009) with a loss based on the negative binomial log–likelihood, because, according to Bühlmann and Hothorn (2007), the advantages of the negative log–likelihood loss for classification problems are the estimation of probabilities, the monotonicity of the loss function and the robustness towards very poor fitted observations. Depending on the assumed flexibility of the effect and on the scale of measurement of a predictor, different types of base procedures can be chosen. A bivariate surface in the form of bivariate tensor product P–splines represents the baselearner for the spatial effect, whereas penalised regression splines are used for modelling the effects of the continuous ecological predictors. Categorial covariates are incorporated by linear baselearners.

In order to restrict the base procedures on weak learners, their complexity is limited to maximal four degrees of freedom. This uniform constraint on the model complexity additionally prevents a selection bias towards variables, for which the structure of the baselearner allows a higher degree of complexity. The restriction on the linear baselearners is realised by a ridge penalty.

For both boosting algorithms, the step size is chosen to be 0.1 and 1000 iterations are accomplished. Only for modelling the occurrence of the Swiss pine with GBM, a step size of 0.01 is selected, because otherwise, the algorithm would stop already after two iterations. The number of boosting iterations for the final boosting models of the data analysis are determined with the 10–fold cross–validation error by the elbow criterion. In the simulation study, the iteration with the minimum cross–validation error is chosen.

5.7. Profits and limitations of the boosting approach

A very attractive feature of boosting is, that, apart from an often superior predictive accuracy, also structural assumptions on the underlying mechanism can be included by the choice of the base procedure. Besides the ordinary GLMs or decision trees, also additive models and even survival models can be estimated (Hastie, Tibshirani and Friedman, 2001). Thus, boosting is a very flexible approach and has a broad range of application.

Even complex problems, in which the number of possible predictors exceeds the number of observations by far, can be solved with the integrated regularisation scheme. Therefore, the algorithm is computationally efficient, because the individual weak learners can be evaluated very fast. The implicit variable selection procedure in high dimensional covariate spaces avoids the multiple test problem, but its statistical properties

are still vague.

Unlike other black–box methods, boosting allows a slight insight into the modelled relationships: variable importance measures identify variables with a high impact and partial effects can be depicted.

On the one hand, the renouncement of statistical assumptions allows the application of boosting for many different problems, but, on the other hand information on the accuracy of the prediction, e.g. confidence intervals, cannot directly be obtained.

6. Simulation study

6.1. Introduction to the simulation study

For the proper use and the evaluation of statistical models in ecology, it is essential to understand their characteristics in this field of science. Besides a sensitivity analysis of fitted models, simulation studies are another approach for investigating the properties of statistical methods. Based on known data generating processes (DGPs), several data sets with varying characteristics are generated and analysed by the examined methods. The objective of the simulation study in this chapter is to gain an insight into the impact of different DGPs, sampling designs and analysis methods, when the different modelling strategies are applied.

For the purpose of selecting the appropriate model for any data analysis, Elith and Graham (2009b) highly recommend a deeper comprehension and a more detailed knowledge of the reasons of varying model performances, which has not yet been sufficiently explored for species distribution modelling. To reveal differences between the performance of the models, their capability to fit the true relationships should be measured. Therefore, Austin et al. (2006) emphasise the necessity of analysing artificial data, which was already realised e.g. by Austin et al. (2006), Reineking and Schröder (2006), Dormann et al. (2007) and Moisen and Frescino (2002).

Firstly, the discrepancies of the four methods are analysed concerning varying sampling designs. Hirzel and Guisan (2002) studied the impact of the sampling design on the estimations of generalised linear models and found out, that a spatialy regular grid improves the accuracy of the model. One aspect of this simulation study addresses to the question, whether also the data–directed models profit from a spatially balanced sample.

In addition to the comprehension of the underlying ecological relationship, the transferability of the resulting model on other locations as well as on hypothetical climate scenarios is of great interest (Fitzpatrick and Hargrove, 2009). Since often, the predictors for that purpose exceed the range of the data in which the model was calibrated, extrapolation properties and the involved uncertainties of the modelling techniques are analysed. While Thuiller et al. (2004) have already investigated some prediction characteristics of GAMs, also data–directed models are taken into account in this study.

Secondly, the structure of the data generating process is considered. Elith and Graham (2009b) explored different modelling techniques with a simulation study and they used an additive covering of effects in order to simulate the underlying mechanism. Since actually the real data generating process is mainly unknown, the differences between the

modelling techniques are analysed in this cahapter not only when an additive structure but also when a tree–like structure is presumed.

The third aspect of the simulation study investigates the impact of various predictor sets on the modelling strategies. On the one hand, it will be analysed, to what extent the disregard of a contributing predictor affects the estimation and whether a spatial trend is capable to compensate for it. On the other hand, the influence of several correlated predictors on the different variable importance measures of the techniques is focused.

6.2. Simulation setup

In this chapter the properties of the examined statistical methods will be investigated regarding three aspects by means of a simulation study. Different sampling designs as well as different analysis methods are considered. Furthermore, the three modelling techniques will be compared, when the data generating process varies.

Virtual species are generated and their true relationships to the environmental predictors are predetermined. The knowledge about the true DGP allows a precise comparison of model performance. The simulation procedure is arranged as follows and repeated 100 times:

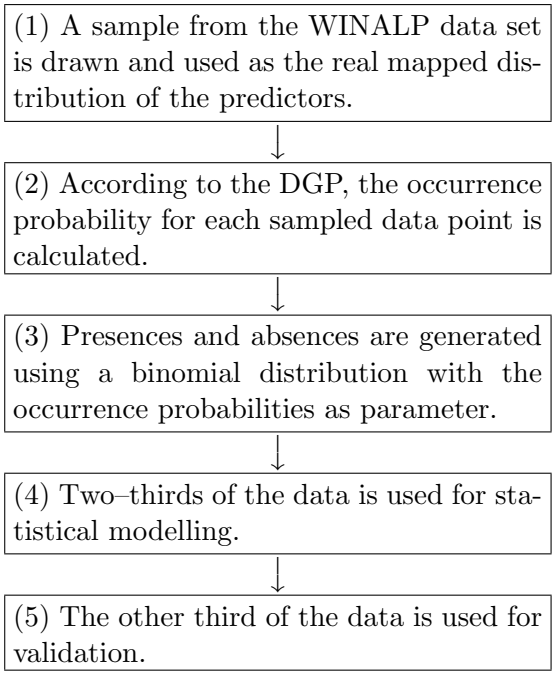


FIGURE 6.1.: *Simulation setup.*

The strategy of the simulation study is to start with an initial scenario, in which the predictors are sampled spatially unbalanced, the occurrence probabilities are simulated without a spatial effect and all contributing, environmental predictors are incorporated into the analysing model. Successively, one aspect is varied by fixing the other aspects according to the initial scenario, which results in the following overview of the scenarios of the simulation study:

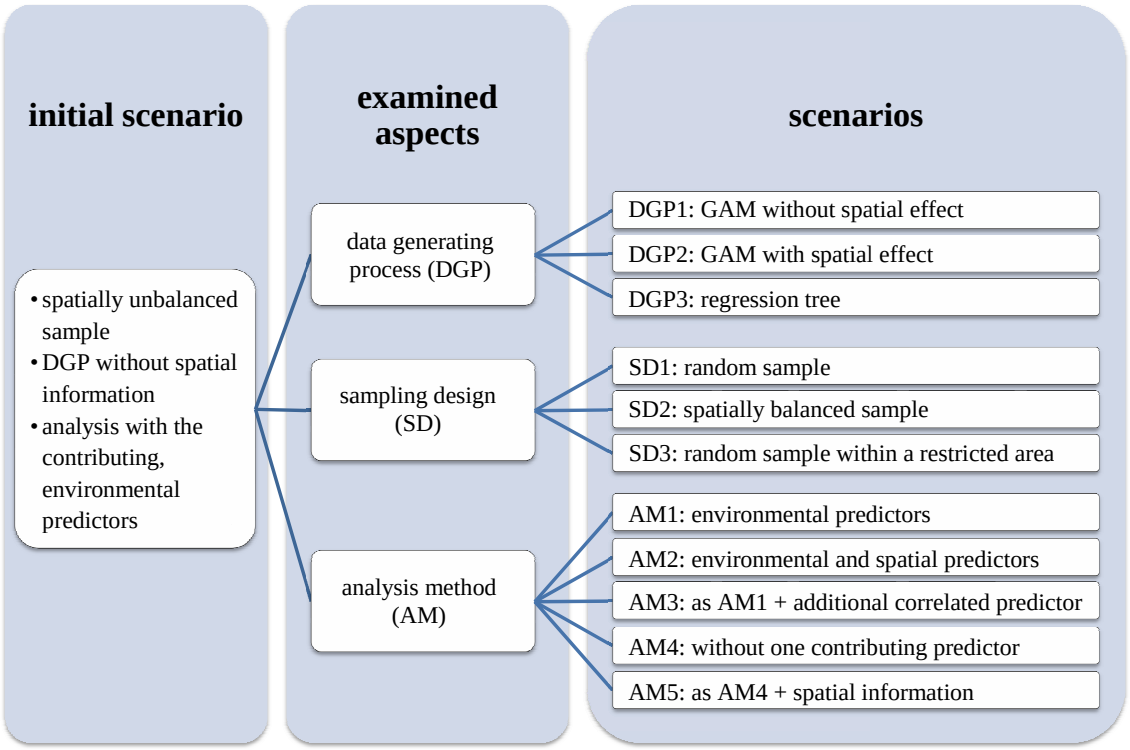


FIGURE 6.2.: *Outline of the simulation study.*

The first issue of the simulation study is, how the three models will behave, if the data in step (3) of the simulation setup is generated with varying DGPs. Besides a GAM with (scenario DGP2) and without a spatial effect (scenario DGP1), a decision tree (scenario DGP3) is considered as a data generating mechanism.

For each observation, the probability of the occurrence of the artificial tree species is calculated with

$$\mathbb{P}_{\text{DGP1}}(Y = 1) = \frac{1}{2}\left[\text{s}(z_{\text{G05\_20}}, z_{\text{P\_JJA}}) + \text{s}(z_{\text{P\_JJA}}) \cdot z_{\text{HYD\_UNIT}}\right]$$

if the tree occurrence depends only on ecological parameters (DGP1) and with

$$\mathbb{P}_{\text{DGP2}}(Y = 1) = \frac{1}{3}\left[\text{s}(z_{\text{G05\_20}}, z_{\text{P\_JJA}}) + \text{s}(z_{\text{P\_JJA}}) \cdot z_{\text{HYD\_UNIT}} + \text{s}(x, y)\right]$$

if a long–range spatial effect of unmeasured confounders, e.g natural and silviculturally guided dispersion, is assumed for DGP2.

The individual components of the above modelling equations are derived from simple regression models for the spruce with the covariates “growing degree days” (G05_20),

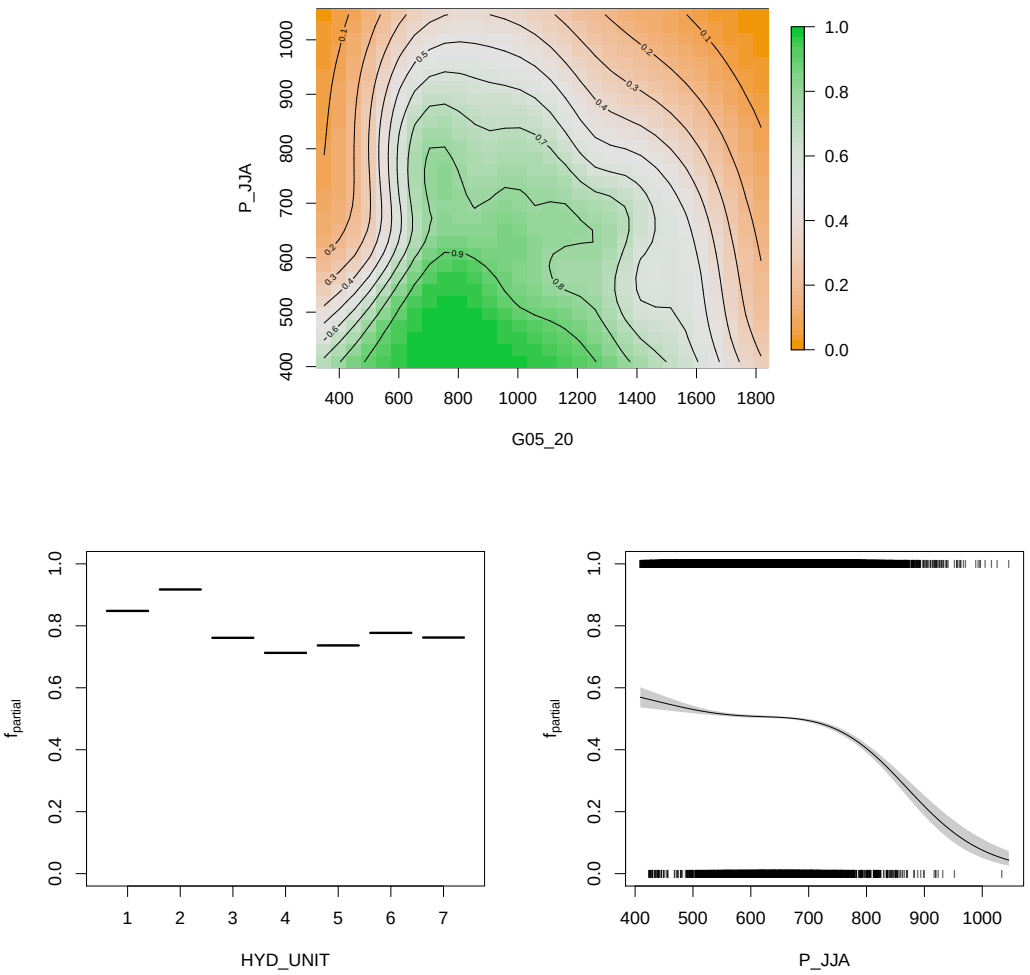


FIGURE 6.3.: *Real partial effects of the simulated tree species in DGP1 and DGP2.*

"precipitation in summer" (P_JJA), "hydrogen unit" (HYD_UNIT; as a factor variable) and optionally the coordinates (X,Y) from the WINALP data set. The response curves and surfaces of the individual components are considered as the truth and are depicted in figure 6.3 (the response surface for the coordinates is not illustrated). The data generating mechanisms for the scenarios DGP1 and DGP2 are not simple in order to imitate the quite complex structure of ecological relationships.

Furthermore, the behaviour of the modelling techniques is examined in scenario DGP3, if not an additive structure, but a tree structure will build up the DGP. The occurrence probabilities of the virtual tree species are simulated by a conditional inference tree on the WINALP data set for the distribution modelling of the spruce (cf. figure 6.4). Therefore, the predictors are again sampled from the empirical environmental variables, as described in step (1) of the simulation setup.

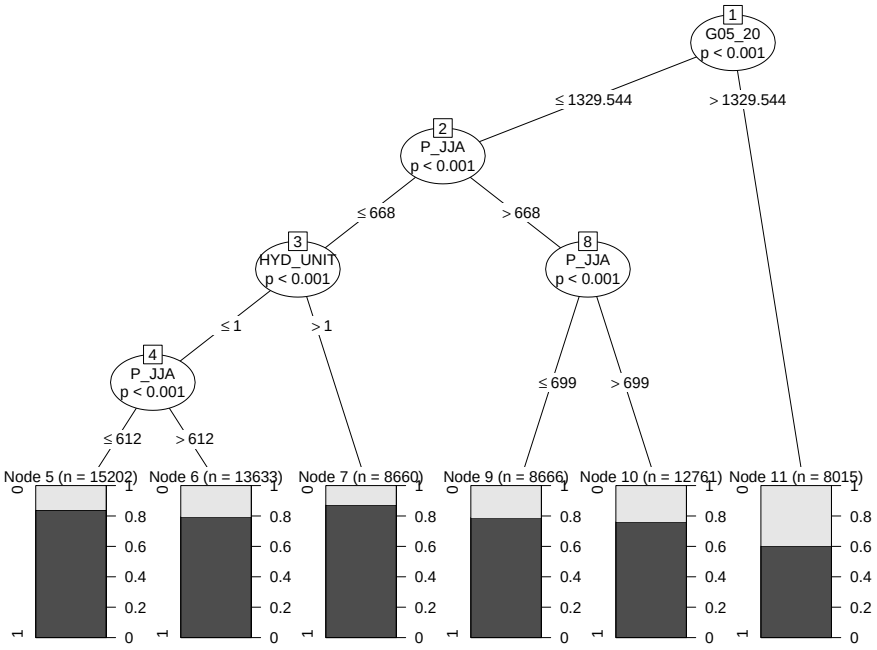


FIGURE 6.4.: *True tree structure for scenario DGP3.*

Secondly, three common types of sampling designs are compared:

- Spatially unbalanced observations
- Spatially balanced sample
- Spatially unbalanced sample within a restricted area, where the collected predictors do not cover the whole ecological niche

Spatially unbalanced covariables (SD1) will be achieved by sampling from the WINALP data set. For the spatially balanced sample (SD2), the observed area is divided into equally sized rectangles along the longitudes and latitudes. One observation is sampled from each rectangle and the predictors are used for the DGP.

Especially the last point is interesting for the examination of extrapolation properties of the GAM, the random forest and the boosting approach. Therefore, only the eastern ($\leq 45.25°$ longitude) observation points are used for the sample drawing (SD3), because this area is slightly colder. The test data sets are drawn from the remaining, warmer observations.

A further topic of the simulation study is the analysis with different predictor sets, which consist of

- only the non–spatial, environmental predictors, which are used in the data generating mechanism,
- spatial and environmental predictors, which are used in the data generating me-

chanism,

- environmental predictors from the DGP and an additional highly correlated variable,
- environmental predictors, but without one of the contributing predictors (G05_20) and
- environmental predictors without one of the contributing predictors, but with spatial information.

This part of the simulation study adresses to the question, to what extent the negligence of the spatial information (AM1) changes the results in comparison to allowing for this information (AM2) and whether additional variables, which partly correlate with the true influencing variables, have an effect on the estimation (AM3). Moreover it is explored, whether an additional spatial effect can compensate for the effect of a disregarded contributing predictor (AM4 and AM5).

The simulation study is accomplished by the Monte Carlo simulation technique, in which the expected value of a resulting specific value is approximated by averaging over several repeated samples. Here, the results of different modelling strategies and the scenarios are compared using the test data sets by means of the measure Δ_p^a, which quantifies the variation of the estimated occurrence probability around the true one on the logit scale:

$$\Delta_p^a = \frac{1}{n}\sum_{i=1}^{n} |\text{logit}(p_i) - \text{logit}(\hat{p}_i)| = \frac{1}{n}\sum_{i=1}^{n} \left|\log\left(\frac{p_i}{1-p_i}\right) - \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right| \quad .$$

The bias of the estimated probabilities is analysed with

$$\Delta_p = \frac{1}{n}\sum_{i=1}^{n} (\text{logit}(p_i) - \text{logit}(\hat{p}_i)) = \frac{1}{n}\sum_{i=1}^{n} \left(\log\left(\frac{p_i}{1-p_i}\right) - \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right) \quad .$$

Depending on the prevalence, deviations of the fitted occurrence probabilities from the true probabilities are measured in a different way and scaled to the whole set of the real numbers.

6.3. Effect of sample size

The sample size, which is used for a study within an appointed region, mainly determines the scale, on which structural differences can be identified. Comparing GAM, random forest and boosting, it will be analysed, to what extent the impact of sample size varies between the three methods.

In addition to the sample size of 4000, which is also used in most instances for the other scenarios, a smaller sample with $n = 2537$ and the twofold size of 8000 are explored.

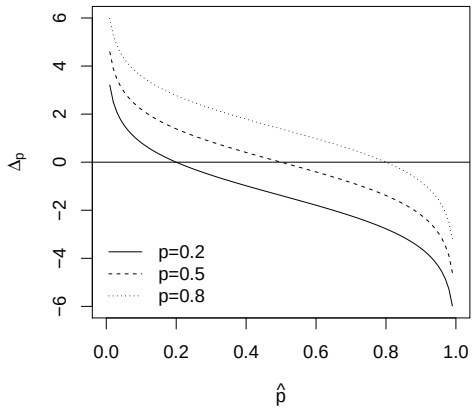


FIGURE 6.5.: *Illustration of the distance measure* Δ_p.

The peculiar size of the small sample is taken into account, because it is further required in chapter 6.4.1.

To quantify the dispersion of the estimations around the true occurrence probabilities, the sum of the absolute deviations between the logits of the true and the estimated probabilities, Δ_p^a, is considered, which identifies different properties:

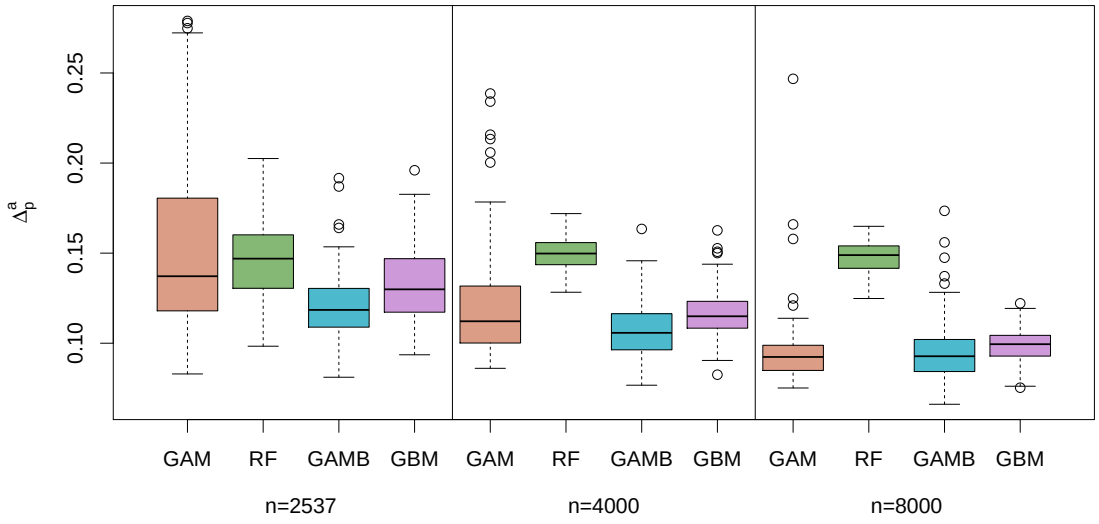


FIGURE 6.6.: *Absolute model performance for varying sample sizes.*

In contrast to random forest (RF), all other models reduce the deviation between truth and estimation with increasing sample size. Especially the generalised additive model profits from a large sample size. GAMBoost (GAMB) performs in all three cases slightly better than the boosted regression trees (GBM).

Tracking the performance of random forest along the "degree value days"–gradient for $n = 8000$ (cf. figure A.1), the difficulties of random forest with modelling the smooth response are detected. At the edges of the gradient, the true occurrence probabilities are overestimated, whereas in the middle part they are underestimated. Therefore, random forest delivers a slightly biased estimation on average with a high variation, which indicates, that the sample size does not suffice to detect the smooth relationship between the predictor variables and the response out of the available presence–absence data.

The resolution affects the accuracy of the estimation in a rather different way. A high sample size improves the results of GAM and the boosting approaches to different degrees.

6.4. Comparison of sampling designs

The suitability in terms of various sampling designs of the three modelling strategies is examined in this chapter. The unbalanced design (SD1) is compared to the spatially balanced design (SD2) and to the unbalanced design on a restricted area (SD3). The SD3 scenario focuses on the extrapolation properties of the models.

6.4.1. Balanced design

Depending on the studied issue, the question of the localisation of the observation sites arises additionally to the used resolution. Diverse sampling designs are possible, e.g. random, stratified, systematic or systematic–clustered designs. The performance of GAM, random forest and boosting is compared, by drawing a random subsample (scenario SD1) and a systematic, respectively a spatially balanced subsample of the data is drawn.

In order to generate a balanced sample, the study region is subdivided into 2537 equally sized units. In each case, one individual observation is drawn out of the units for the training data set. For a valid comparison of the balanced and the unbalanced design, only 2537 randomly sampled observations are used for SD1. The test data set for both scenarios is sampled at random.

Structural differences of the data sets in the two scenarios can be found by taking into account the spatial distribution of the sites, but also by regarding the comprised temperature range. The mean temperature range in SD2 is wider ($353 - 1815$ degree value days) than in SD1 ($432 - 1804$). Thus, a larger spectrum of the temperature gradient, which is especially expanded at the lower boundary, is analysed and the simulated observations are varying much more.

A look at the absolute deviations of the estimated occurrence probabilities from the true values shows results, which differ between the considered modelling strategies:

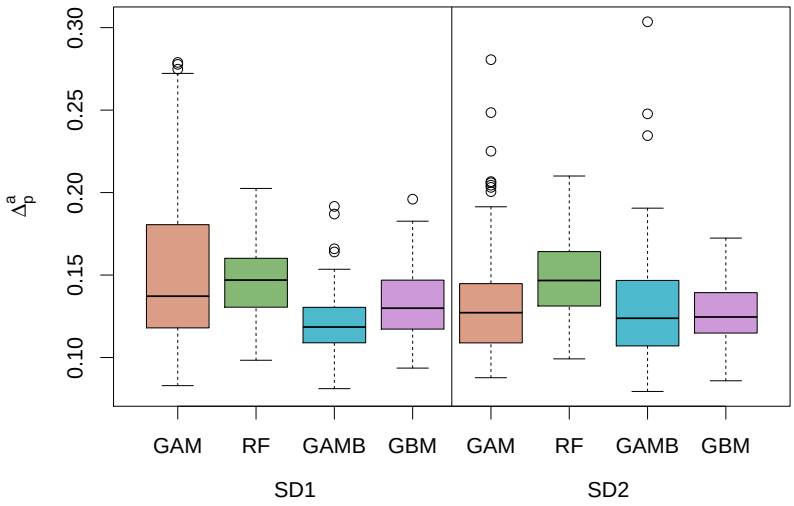


FIGURE 6.7.: *Absolute model performance for the unbalanced (SD1) and the spatially balanced (SD2) design.*

Using a spatially balanced design tends to change the predictive capacity of the analysed procedures, especially considering GAM and GAMBoost. The tree–based methods deliver similar results for both scenarios.

Modelling with GAM takes a remarkable advantage of the balanced design by raising the accuracy of the estimations. Apparently, GAMs profit from the higher variability of the data in the spatially balanced sample. Closer scrutiny of the averaged, absolute logit–residuals shows, that at the lower boundary of the "degree value days"–gradient a minor deviation from the truth is prevalent for the balanced design.

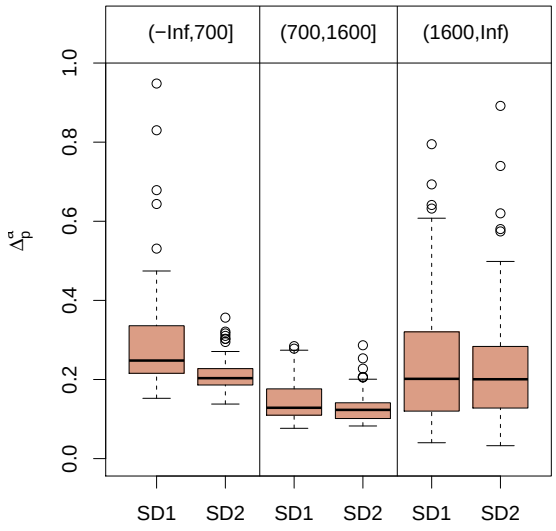


FIGURE 6.8.: *Performance Δ_p^a of GAM in scenarios SD1 and SD2 subdivided into intervals of the predictor "degree value days".*

Thus, the additional information of the balanced design is used to ameliorate the generalised additive model, especially at the boundaries of the predictor space.

The analogous illustration for GAMBoost (cf. figure A.2) shows a similar behaviour at the lower boundary like GAM, but in comparison to SD1 the scenario SD2 performs worse in the middle of the temperature gradient. Hence, the sparser information of the training data in SD2 for the central region of the predictor space downgrades the estimation in this area with an impact on the entire performance of the model.

6.4.2. Extrapolation

The projection of species distribution models often comes along with the exceeding of the scope, on which the models are calibrated. The arising uncertainties are the subject in this section.

Elith and Graham (2009b) explored the behaviour of GLM, boosted regression trees and random forests in terms of detecting the true response shape, identifying the real mapped suitabilities and extrapolation. The analysis of the extrapolation properties concentrates here on the description of the resulting response curves, whereas this part of the simulation study rather addresses to the issue of the predictive performance on extrapolated data.

Whereas in the initial scenario the sites are sampled from the entire study region (SD1) with averaged 1054 degree value days, only the western, slightly colder part (cf. figure 2.3) with averaged 997 degree value days is used for model building in scenario SD3. The models are validated on observations of the eastern part (averaged 1142 degree value days).

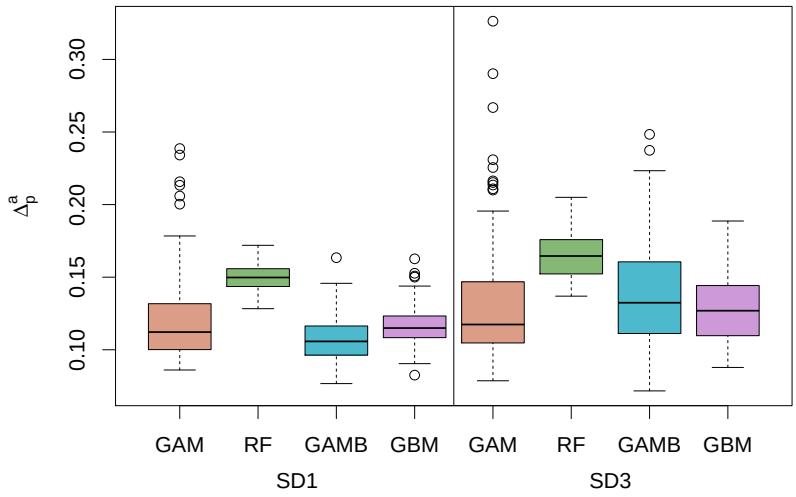


FIGURE 6.9.: *Absolute model performance on extrapolated data (SD3); performance of the initial scenario (SD1) as reference.*

Expectedly, figure 6.9 shows a minor estimation accuracy and also the variation of performance is higher for all models, if the test data set exceeds the range of the training data set. Whereas GAM reveals to some degree good extrapolation properties compared to the initial scenario, the estimations of the other models degrade.

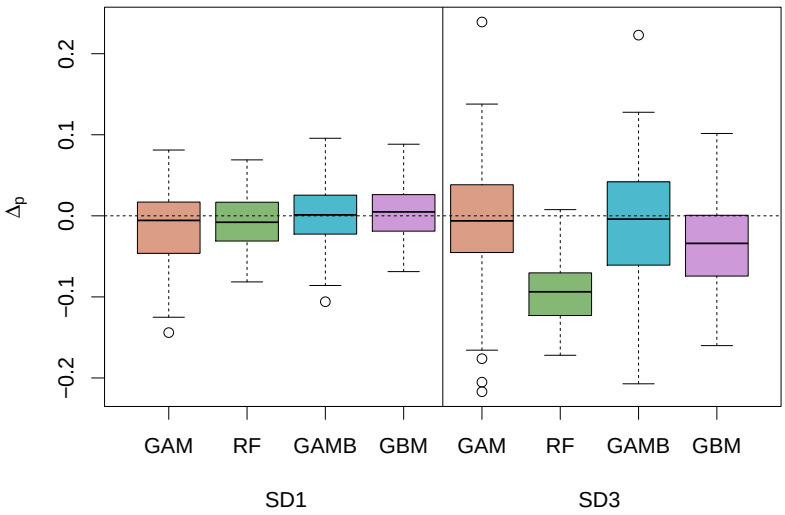


FIGURE 6.10.: *Mean model performance on extrapolated data (right); performance of the initial scenario (SD1) as reference (left).*

Considering Δ_p (cf. figure 6.10), the tree based methods random forest and GBM overestimate the occurrence probabilities, but the bias of random forest is more pronounced. The results of GAM and GAMBoost are comparable. On average, both modelling techniques are unbiased.

In order to explain the results of scenario SD3, some aspects of the extrapolation properties are investigated in detail.

GAM and GAMBoost continue their response curve smoothly to the extrapolated data and the direction of the surface is determined by the basis functions at the boundary of the training data. This appears to work quite well for GAM in this simulation study, mainly due to the low distance between the test data and the training data.

In contrast to GAMs, the quality of the data on the boundary highly influences the estimation in GAMBoost models, which is illustrated in the following figure by means of analysing one of the simulated data sets with and without an artificial outlier.

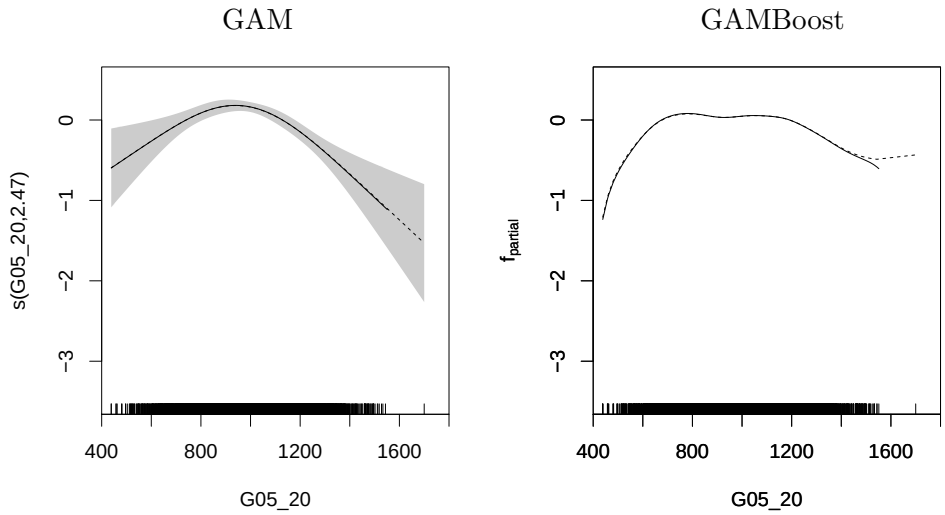


FIGURE 6.11.: *Influence of outliers on the estimation of GAM and GAMBoost; solid line: estimation without outlier, dashed line: estimation with outlier.*

If there are unrealistic values at the boundary of the training data, they will affect the GAMBoost modelling in this region, because the boosting algorithm concentrates with increasing boosting iterations on the observations, which are difficult to fit. Early stopping of the boosting algorithm, even slightly earlier than according to the cross–validation criterion on the training data, alleviates the problem a little.

This artificial example displays, that the GAMBoost rather than the GAM estimations at the boundary of the predictor space will significantly depend on the observations in this area, even if they are unrealistic. As the simulation study reveals, the strategy of GAM works better in comparison to GAMBoost.

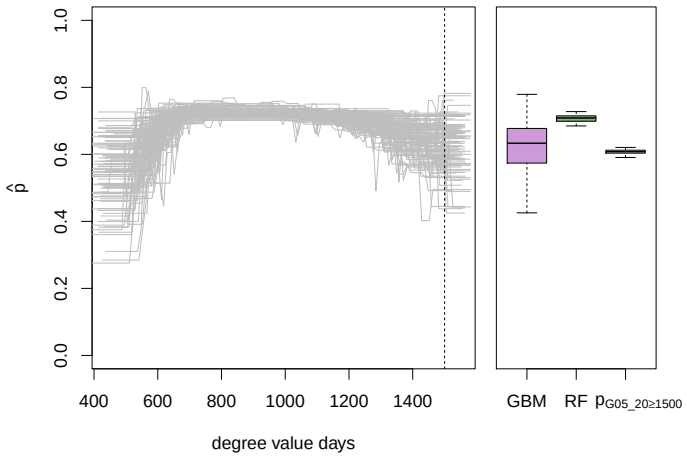


FIGURE 6.12.: *Extrapolation properties of GBM and random forest: partial dependence curves of GBM in scenario SD3 (left); partial dependence of GBM and random forest in scenario SD3 at* $G05\_20 = 1500$ *in comparison to mean occurrence probability if* $G05\_20 \geq 1500$ *in each simulated data set.*

Figure 6.12 provides a deeper insight into the shortcomings of the tree–based methods. The plot in the left part illustrates the individual developing of the partial dependencies of each simulated training data set along the G05_20 gradient, while on the right hand side, the two boxplots depict the distribution of the partial dependencies at the covariate "degree value days" fixed to 1500. As a reference, the mean simulated occurrence probability for observations with "degree value days" above 1500 is represented in the third boxplot.

As the partial dependency curves of the GBM indicate, the tree–based methods random forest and GBM will be continued with constant values, if the region of the data, which is used for modelling, is exceeded. In particular, the variation at the edges of the gradient attracts attention, which is presented with the first boxplot. In comparison to that, the boxplot of the random forest shows a variation on a significantly fewer extent, which derives from the heuristic of the method to average preferably unbiased trees. However, the validation of the random forests on extrapolated data is considerably more biased than by applying GBM.

The trend of GBM and random forest to overestimate extrapolated test data derives from the unrealistic, constant continuation of the response curves, whereas the boosted trees perform better than the averaged trees. The simulation study demonstrates fairly better extrapolation properties of GAM.

6.5. Comparison of data generating processes

Usually the complex interactions in ecological systems are known only fractionally. Assuming an additive overlay of the individual covariate effects is a widespread approach, but also a tree–like structure is imaginable. Each model concentrates on specific aspects of the underlying structure. Because of being always a simplification of reality, detecting the entire complexity of the relations is often not possible.

For that reason, the properties of GAM, random forest, GAMBoost and GBM based on different data generating processes, are examined in terms of their predictive accuracy. In addition to the initial scenario (DGP1), in which the data is generated by a generalised additive model with three environmental predictors, a GAM with an additional spatial effect (DGP2) and a classification tree are analysed with the three environmental predictors "degree value days", "precipitation" and "hydrogen unit".

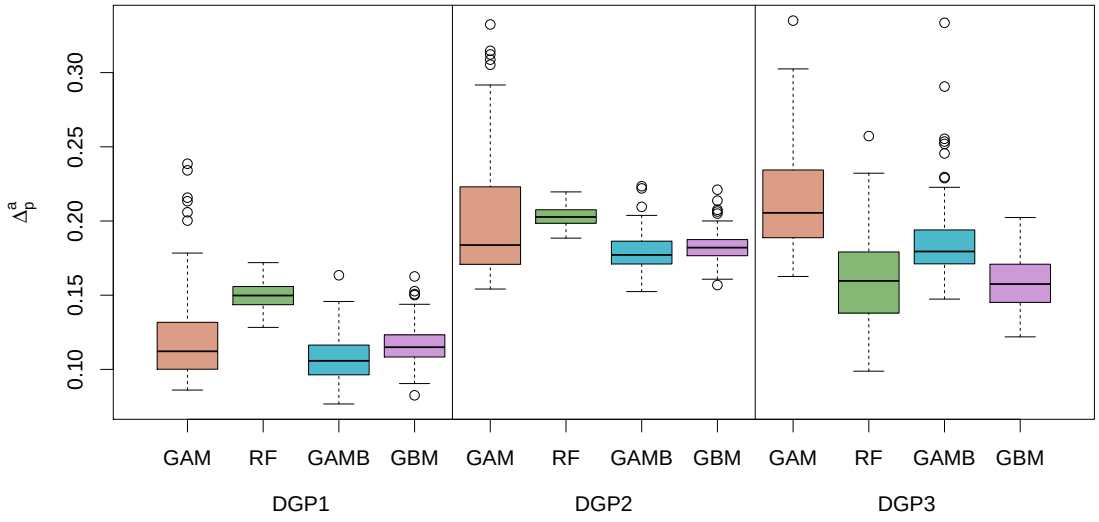


FIGURE 6.13.: *Absolute model performance for different DGPs: GAM with only environmental predictors (DGP1), GAM with environmental predictors and spatial effect (DGP2), classification tree with environmental predictors (DGP3).*

Disregarding the spatial effect causes fewer precise estimations for all models. The considerably higher variation in the performance of GAM is to note, which will be absent, if the other techniques are conducted. GAMBoost looses its slight advantage over GBM.

As expected, accounting for the spatial effect in DGP2 with GAM and GAMBoost enhances the estimation, especially, if GAMBoost is used (cf. figure 6.14). The improvement is not excessively high because apparently, the sample size of 4000 does not satisfy to detect the complex additive structure of DGP2 more appropriately.

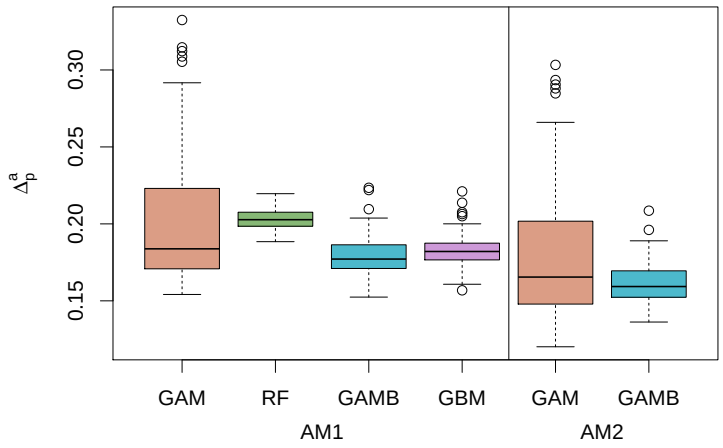


FIGURE 6.14.: *Absolute model performance for the comparison of scenario AM1 with AM2 based on DGP2.*

If the data is generated by a tree mechanism, as expected, the tree–based models random forest and GBM will perform better than GAM and GAMBoost, which have difficulties with this kind of structure.

The behaviour of GBM is remarkable, because this approach delivers good results for tree–like as well as for smooth, additive structures. Although it does not directly model smooth response curves, the additive structure of the boosting algorithm combined with a tree as baselearner, is appropriate for modelling data generated from GAM as well as data with a tree structure.

6.6. Comparison of analysis methods

It is a matter of common knowledge, that disregarding one or more contributing predictor variables biases the estimation. Because it is not possible, to identify all influential variables in the complex interdependencies of ecology, it must be assumed, that not all of them can be comprised by data collection.

The aim of this chapter is the examination of different analysis methods, each varying in the utilised predictor set for the analysis. Based on the data generating process DGP1, i.e. without spatial effect, the data is analysed with the different predictors.

Because commonly spatial position is not included in tree based models, the scenarios, which contain spatial information, i.e. AM2 and AM5 respectively, are only evaluated for GAM and GAMBoost.

The characteristics of the modelling techniques with different analysis methods are investigated in terms of predictive accuracy and the behaviour of the corresponding variable importance measures.

6.6.1. Predictive accuracy

First of all, the predictive accuracy is examined, when DGP1 is used for data generating and the evaluation is carried out with AM1, AM2 and AM3. Here, the question whether an additional, in reality not directly contributing, but highly correlated variable, affects the accuracy of the estimation arises.

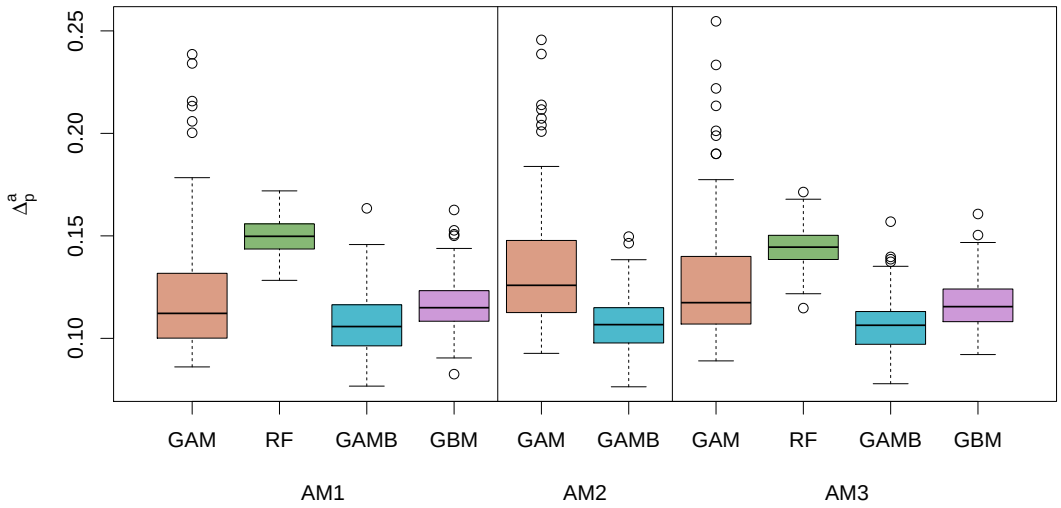


FIGURE 6.15.: *Absolute model performance for the comparison of scenario AM1 with AM2 and AM3 based on DGP1.*

A decline in accuracy of GAM can be discovered, if variables, which actually do not influence the data generating process, e.g. coordinates or a further environmental covariate, are used for the analysis. However, the relevance of the correlated predictors in the various approaches is of greater interest and examined in the next section.

But the results will change, if the data generating process is assumed to contain a spatial effect: GAM and GAMBoost can be improved by including the coordinates as predictors (cf. figure 6.14). This finding is really not surprising, but it poses the question of whether allowing for the coordinates can compensate for the disregard of spatial acting, contributing predictors. Scenarios AM4 and AM5 address to that issue.

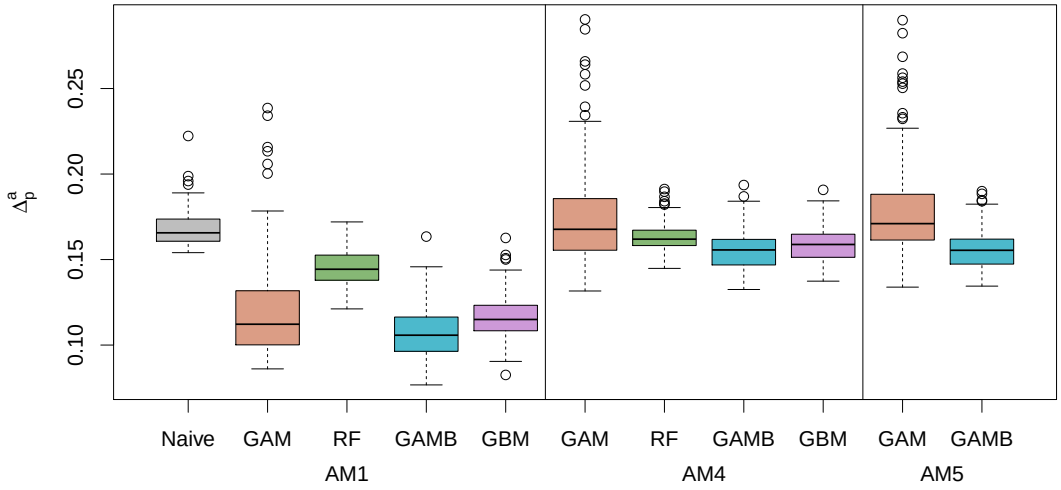


FIGURE 6.16.: *Absolute model performance for the comparison of scenario AM1 with AM4 and AM5 based on DGP1.*

The results show, that, if a contributing predictor is not included, the accuracy of the predictions will decrease for all models to the scope of naive estimation calculated through the prevalence of the respective test data set.

The additional inclusion of a spatial effect cannot offset the negligence of an important, spatial acting covariate neither if GAM nor if GAMBoost is applied. Either the spatial effect is too smooth to comprise the rather small–scaled coverage of the temperature variable or 4000 observations are not sufficient to model the temperature effect spatially.

6.6.2. Variable importance

A major topic of the application of statistical methods is the identification of driving forces of a process. Often the relevant predictors have to be detected from many possibly important ones. Therefore, each modelling technique provides evidence, which can be quantified by various variable importance measures as theoretically introduced in the chapters before.

The challenge is, that each modelling technique measures the variable importance in a different way and thus, the measures emphasise various aspects. Taking into account the importance measures of several models reveals a differentiated view on the relevance of the individual predictors keeping their diverseness in mind.

Within the underlying analysis, GAM and GAMBoost include only the main effects of the environmental predictors and interactions are left out. Consequently, the corresponding importance measures quantify the relevance of the predictors disregarding their interactive potential. The variable importance measures of GAM and GAMBoost are confined to characterise the strength of the smooth influence of each single predictor conditioned on the others.

In contrast to the importance measure for GAMBoost, the p–value of a generalised additive model allows for the uncertainty of the estimation existing especially at the boundaries.

Since random forest and GBM also model interactive effects, these approaches assess the relevance of a predictor in a more general way. However, if random forest is used, the importance of a covariate in a very elaborate construct will be evaluated, whereas for GBM the relevance is determined in terms of interactions with the maximal depth.

For a more profound understanding of the properties of the different importance measures, their behaviour is examined under various analysis methods with simulated data. Therefore, only the ranks of the importance measures are taken into account, because the absolute values are difficult to compare and, in the case of random forest, not interpretable. Mean inverted ranks are used to determine the importance depicted in figure 6.17.

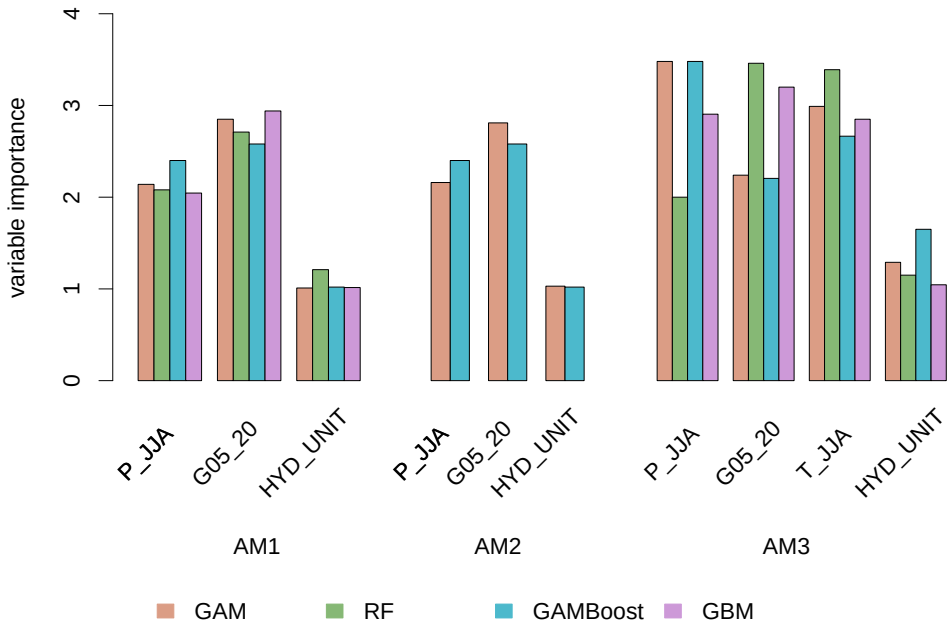


FIGURE 6.17.: *Mean inverted ranked variable importance for AM1, AM2 and AM3 (the most important predictor receives the most points).*

Analysing the data of the initial scenario with the contributing, environmental covariates (AM1) delivers a similar pattern of variable importance for the different approaches. For GAMBoost, the advantage of G05_20 compared to P_JJA is not very evident. The reason thereof is, that the importance measure for GAMBoost is slightly more affected by individual, extreme values at the boundaries of the gradient.

Furthermore, the importance of the environmental covariates will not change, if additionally to the true contributing predictors a smooth spatial effect is included in the analysis (AM2), because the spatial effect contributes only marginally to the model.

The main issue of scenario AM3 is the impact of two highly correlated predictors on the variable importance measures. The results of figure 6.17 show quite different characteristics for the four models.

GAM and GAMBoost similarly evaluate the importance of the variables. On average, precipitation is the most important predictor. But a closer look reveals, that in the most cases one of the two temperature variables is more fundamental than P_JJA, whereas the other temperature variable possesses only a marginal effect on the model. Relating to GAM and GAMBoost, only one of the correlated predictors has a major impact on the resulting model.

In comparison to that, both temperature variables are equally important for random forest and like in the initial scenario temperature is more important than precipitation. Thus, the random forest model gains information from both correlated predictors.

The GBM procedure selects in each iteration two variables for a tree, which improve

the model at best. The algorithm does not recognise any significant differences between the variables P_JJA, G05_20 and T_JJA in AM3 and a further look at the results of the single simulated data sets does not reveal any structure for the variable importance. Apparently, GBM provides only rough information about the relevance of the predictors, if they are correlated.

6.7. Concluding remarks

Finally, the results from the simulation study will be evaluated and summarised.

Generalised additive model

The high variability in the GAM estimations for the simulated data sets can be lowered by an increased sample size, which leads to comparably good predictions.

Furthermore, the estimations are improved by a sample design with a higher diverseness in the data, e.g. by a spatially balanced design. In particular, the extrapolation properties of the GAM should be emphasised. The predictive accuracy is slightly superior to the other modelling techniques. Apparently, it is a good strategy to extrapolate the estimation with the smooth continuation of the response surface.

However, if the simulated data exhibits a tree–like structure, the generalised additive model performs rather poor. This is due to the structure of the data itself, but also due to the fact, that only predetermined interactions can be modelled. The true contributing interactions will have to be previously identified by experts, which is especially demanding and often impossible, if high dimensional interactive terms exist.

Inducing a multicollinearity problem through an additional, highly correlated covariate, tends to downgrade the predictive accuracy of the model. The inaccurate and instable estimations of the effects must not be ignored in the application of a GAM with correlated predictors.

Random forest

Random forest has difficulties with predicting the complex, additive structure in the simulated data set. Even an increased sample size of 8000 does not provide better results and obviously, the contained information of presence–absence response does not suffice for an accurate model.

The predictive performance of random forest even deteriorates, if the model is used for extrapolation, because the response surface is constantly continued beyond the range of the observed data.

However, random forest will render comparably good results, if the underlying structure is a regression tree. Thus, modelling high–dimensional, especially tree–like interactions with implicit variable selection is the strength of random forest.

The incorporation of an additional predictor, which provides no further information, does not affect the predictive accuracy of random forest and exhibits a similar importance as predictors, which measure the same factor.

Boosted generalised additive models

In most of the investigated scenarios, GAMBoost belongs to the best models. The predictive accuracy of the boosted version of generalised additive models tends to be less varying for the different scenarios in comparison to GAM.

Another difference between GAM and GAMBoost originates from the extrapolation properties. Because the response surface is not continued as smooth, the predictions of GAMBoost on the extrapolated data sets are even worse. Furthermore, GAMBoost models do not profit from a balanced sample.

Tree–like structures can be predicted more precisely than with GAM, but, as expected, less accurately than with tree–based methods.

If a smooth spatial effect is included in the data generating process, the predictions of GAMBoost are slightly more precise than with GAM, even though the degrees of freedom are restricted.

The implicit variable selection of the GAMBoost algorithm chooses mainly one out of several highly correlated predictors. This is also obvious in the variable importance measure.

Boosted regression trees

The boosted regression trees represent an appealing approach for modelling smooth, additive structures as well as tree–like mechanisms. If the data is generated with a smooth response surface, GBM is in the scope of GAM and GAMBoost or only slightly worse. In the simulated data set, the ability of recognising tree–like structures is similar to random forest.

The constant continuation of the response surface on extrapolated data strongly varies between the different simulation data sets. Although the absolute deviation of the extrapolated predictions from the true occurrence probabilities is comparable to GAM and GAMBoost, the estimations are biased.

During the boosting iterations, those predictors are selected, which enhance the model with trees of a predefined depth at most. Hence, variable importance for GBM evaluates rather the relevance of the predictors for interactions, than for the

main effects. However, the importance measure will show an insecure behaviour, if the predictors are highly correlated.

Referring to GAM and GAMBoost, the addition of a spatial effect to the model cannot compensate for the disregard of an important predictor. Two reasons are suggested: Firstly, the neglected predictor contributes to the data generating process in interaction with another covariate. Secondly, the result could be due to a very small–range spatial action of the predictor, which cannot be covered with a smooth, quite long–range spatial effect.

All in all, GAM has its main advantage in the attractive extrapolation properties. Depending on the data generating process the performance of the models varies, whereas GBM seems to be the most general model. Especially the different variable importance measures can identify various aspects and characteristics of the underlying mechanism.

7. Species distribution modelling of forest communities

After the theoretical description of the investigated methods and the examination of some properties with a simulation study, the application and evaluation of GAM, random forest and boosting for the spatial prediction of forest communities are the matter of interest in this chapter.

Therefore, model calibration is accomplished with a proportion of 75% of the original data set; the remaining part of the data is used for validation.

7.1. Development of expert models with GAMs

In chapter 3 different approaches for improving the simple generalised additive model were introduced. Their profit is now explored with the application to the WINALP data set and the parsimonious expert models are developed.

As seen in the simulation study, GAMs have difficulties with the identification of tree–like data structures. Thus an approach, which includes a correction factor allowing for complex interactions, is considered. Furthermore, models with weighted observations are calculated in order to imitate a balanced design, in which particularly the boundaries of the parameter space are weighted stronger. To account for long–range spatial autocorrelation, models with a smooth spatial effect are examined. Calibration is conducted with the GCV criterion:

	Ash		**Spruce**		**Swiss pine**	
	without Int.	with Int.	without Int.	with Int.	without Int.	with Int.
Model 1	0.4604	0.4588	0.3707	0.3695	0.0041	0.0041
Model 2	0.3855	0.3781	0.3338	0.3286	0.0010	0.0009
Model 3	0.5001	0.4952	0.3817	0.3799	0.0123	0.0107
Model 4	0.4776	0.4670	0.3683	0.3617	0.0032	0.0030

TABLE 7.1.: *GCV values for different modelling techniques: Model 1: unweighted GAM, with coordinates; Model 2: weighted GAM, with coordinates; Model 3: unweighted GAM, without coordinates; Model 4: weighted GAM, without coordinates.*

Table 7.1 illustrates, that models, which include a smooth spatial effect, provide a better

GCV value for all tree species, which indicates, that a spatial effect exists. Thus, the allowance for unmeasured confounders ameliorates the model.

The inclusion of interactions through an additional factor renders just a marginal advantage. Expressed in per cent, the Swiss pine profits from the additional interaction factor most.

Weighting the observations also improves the models according to the GCV-criterion. However, raising the influence of observations at the boundary of the parameter space through weighting means, that the estimation focuses more just on these regions of the parameter space, which has often little impact on the estimation, because normally either presences or absences are prevailing there. It is suggested, that the improvement of GCV of the weighted models results rather from an artificially increased importance of small residuals than a better model fit. In addition, the ecology expert appraised the predictions of the weighted models as quite unrealistic.

Taking everything into account, the sparse hypotheses–directed models are established without weighting observations and also without an interaction factor, but with allowing for a spatial effect in the models for the ash and the spruce; for the Swiss pine also the spatial effect is not included.

In order to select relevant predictors, the decisions of the expert were guided by plausibility of the resulting response shapes as well as by the importance of the variables in the variable selection procedure with the GCV criterion. The following environmental predictors are selected for the expert models:

ash	T01_20, TGBS
spruce	T_JJA, P_JJA
Swiss pine	T_JJA, P_JJA

7.2. Comparison of model performance

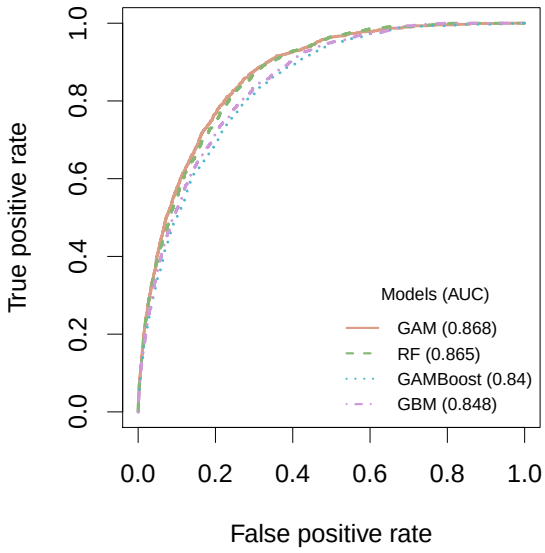
After calibrating the models of each estimation strategy, their predictive performance is compared by means of the discriminatory power. The discrimination ability of the individual models is validated with the widely used, threshold– and scale–independent AUC (area under the receiver operation characteristics) criterion on the external test data set.

The ROC (receiver operation characteristics) curves and the corresponding AUC values for each individual tree species are illustrated in figure 7.1. For a further evaluation of the classifiers, confusion matrices (cf. table 7.2) are assessed by the prevalence of the examined species as threshold (Liu et al., 2005).

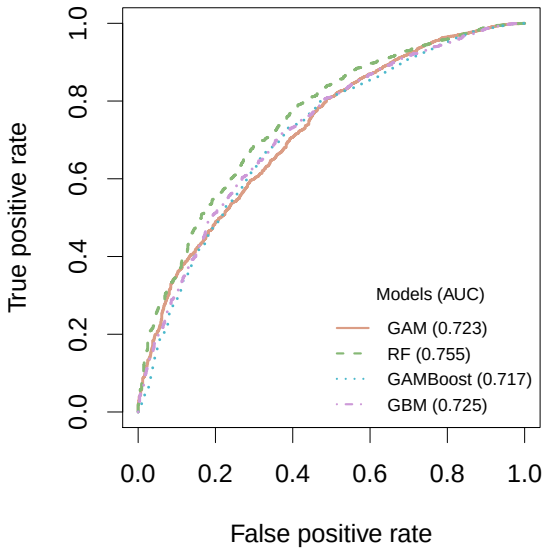
Similar to the findings of Guisan et al. (2007), it is quite evident, that the discrimination power varies more between the different tree species than between the modelling strategies. The reason for that is, that the ROC curve is mainly influenced by the

distribution of the tree species along the examined gradients and thus, according to Lobo, Jiménez-Valverde and Real (2007), the specialisation of a species is reflected. In general, the curves are rather close to each other, but some tendencies are visible.

ash



spruce



Swiss pine

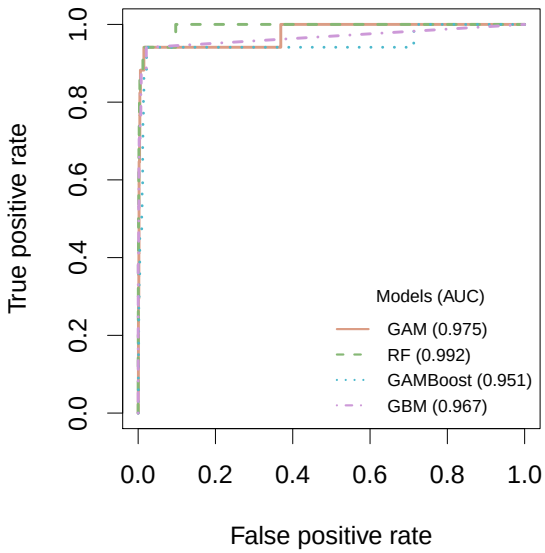


FIGURE 7.1.: *Model validation with ROC–curves.*

The predictive capacity of random forest stands in the forefront regarding all tree species. Although the GAMs are calculated with just a few predictors, they perform in the range of the other models. For all three tree species, the random forest approach discriminates the data marginally better than the two boosting methods, which deliver quite the same results.

	ash		**spruce**		**Swiss pine**	
	0	1	0	1	0	1
GAM						
0	9380	3061	508	262	12866	956
1	244	1154	4725	8344	1	16
RF						
0	10586	1855	239	531	13693	129
1	471	927	952	12117	2	15
GAMBoost						
0	9684	2757	513	257	12086	1736
1	381	1017	4285	8784	1	16
GBM						
0	9056	3385	485	285	13554	268
1	272	1126	3757	9312	2	15

TABLE 7.2.: *Confusion matrices; rows: true abundances, columns: estimated abundances.*

Modelling the ash with GAMs and random forest provides comparably good results in terms of AUC (0.868 and 0.865 respectively), but the misclassification rate of GAM (0.239) exceeds the rate of random forest (0.168). Random forest differs from the other techniques due to its high specificity, but it possesses the lowest sensitivity of all models. The boosting approaches indicate a slightly worse discrimination power, with AUC = 0.84 for GAMBoost and AUC = 0.848 for GBM.

Analysing the spruce, the discrimination quality of GAM (AUC=0.723) is not as good as the performance of random forest (AUC=0.755) and similar to the boosting methods (GAMBoost: AUC=0.717; GBM: AUC=0.725). Especially the presences can be well predicted with random forests, whereas the other approaches have issues with their detection.

The barely observed Swiss pine is difficult to model, because only information from 68 presences is available. Again, the random forest (AUC=0.992) discriminates best, followed by GAM (AUC=0.975), GAMBoost (AUC=0.967) and GBM (AUC=0.936). However, if the estimated occurrence probabilities are dichotomised, the prediction accuracy is higher for the tree based methods than for GAM and GAMBoost.

All in all, the random forests exhibit the highest discrimination level and tend to predict only the more frequent class quite well.

7.3. Variable importance

Since ecological relationships are highly interdependent and influenced by many conditions, it is essential to identify the mainly relevant factors in order to construct reasonable and parsimonious models. Therefore, variable importance measures provide decision support.

In section 6.6.2, the properties of the variable importance measures are outlined and investigated with a closer look at the impact of correlated predictors. The order of variable importance for the WINALP data is determined analogous to the simulation study: the higher the rank of importance the higher the score. The maximum score is 16.

Before analysing the results, it is to mention, that the predictors for modelling GAMs are preselected and in contrast to the simulation study, the relevance of the predictors of the GAMs is now evaluated through the sequence of removal in the backward selection procedure. Furthermore, the boosting approaches implicitly select variables, because not all predictors contribute to the model. As the results for the three tree species on the subsequent pages show, GAMBoost excludes more variables than GBM.

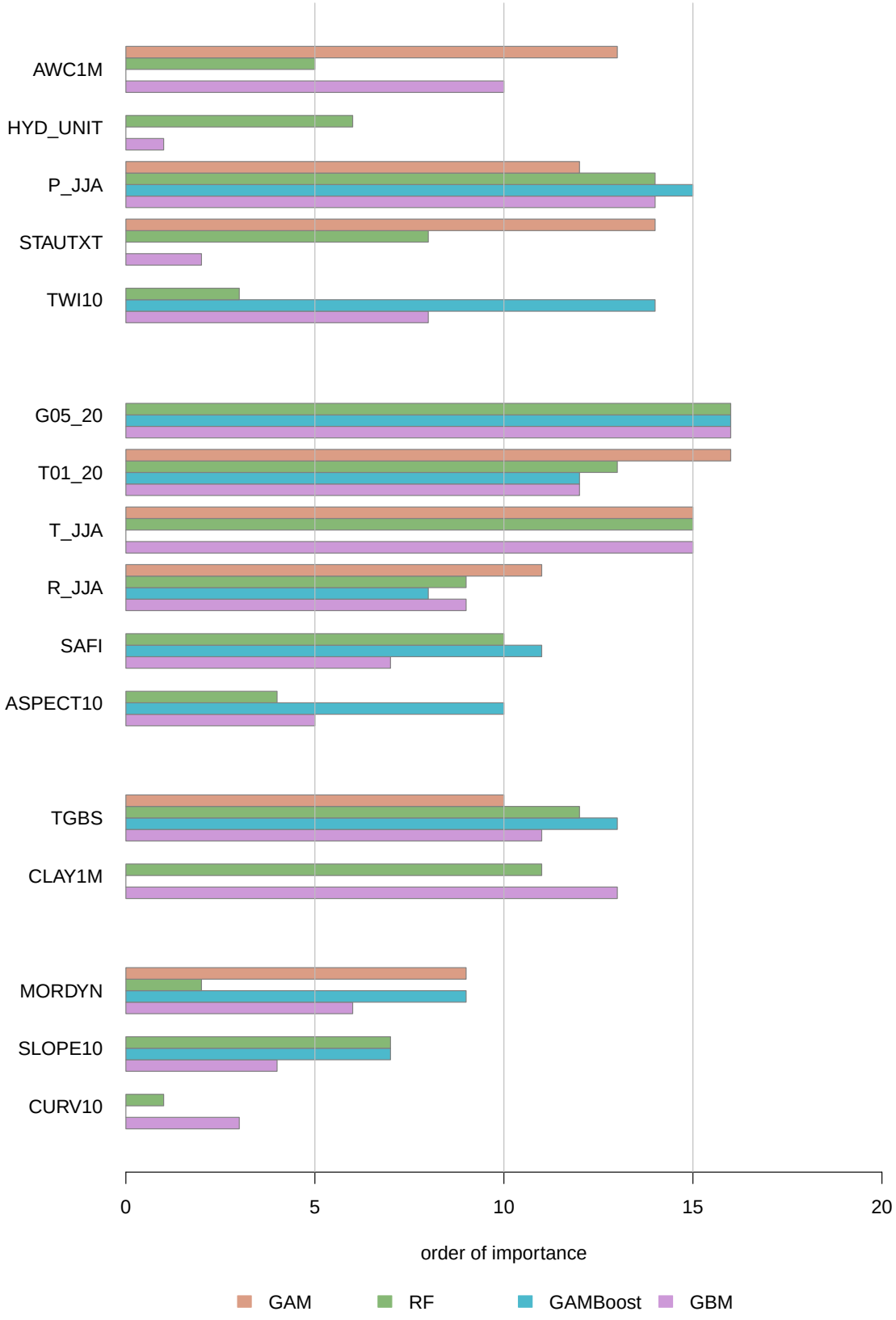


FIGURE 7.2.: *Variable importance: ash.*

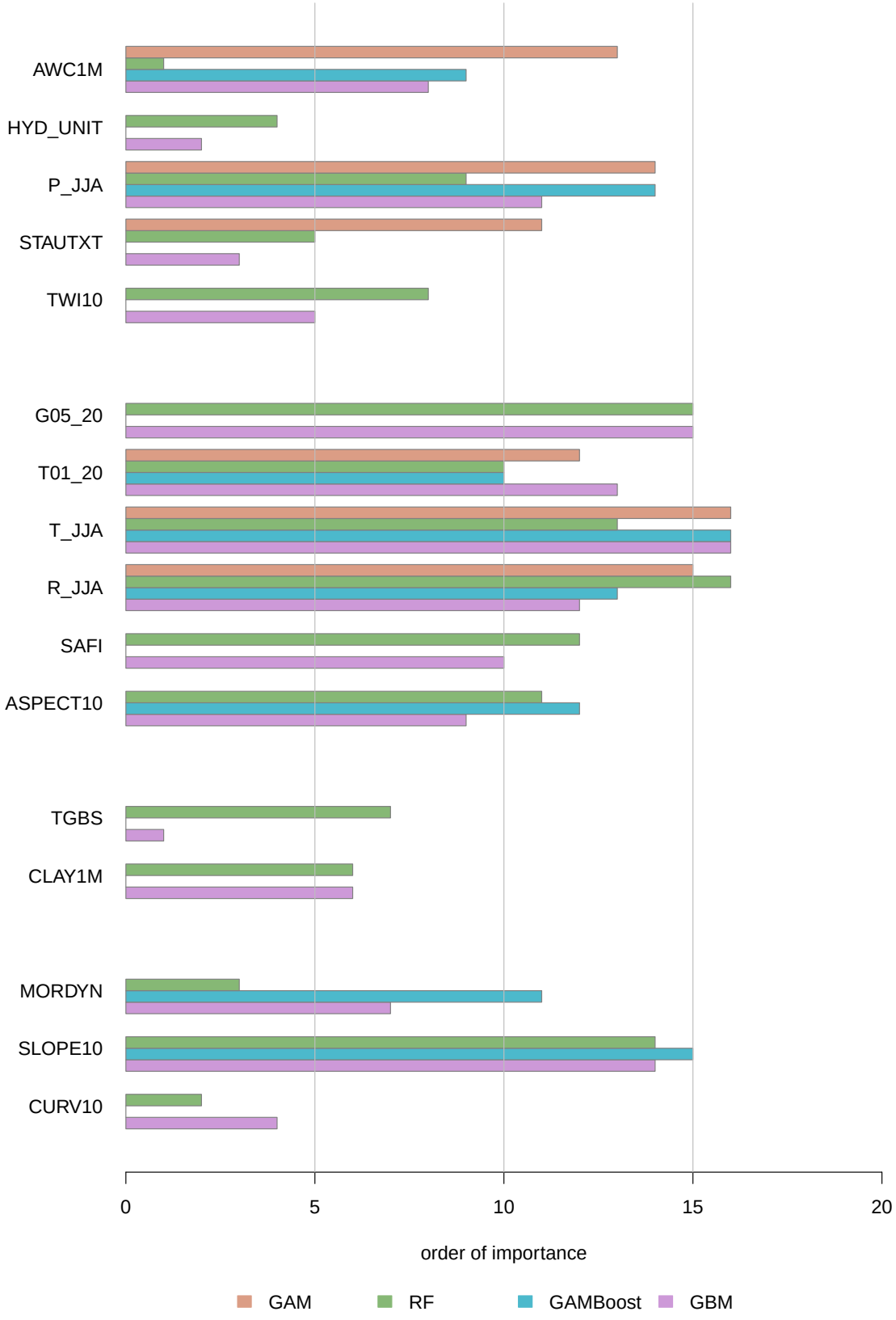


FIGURE 7.3.: *Variable importance: spruce.*

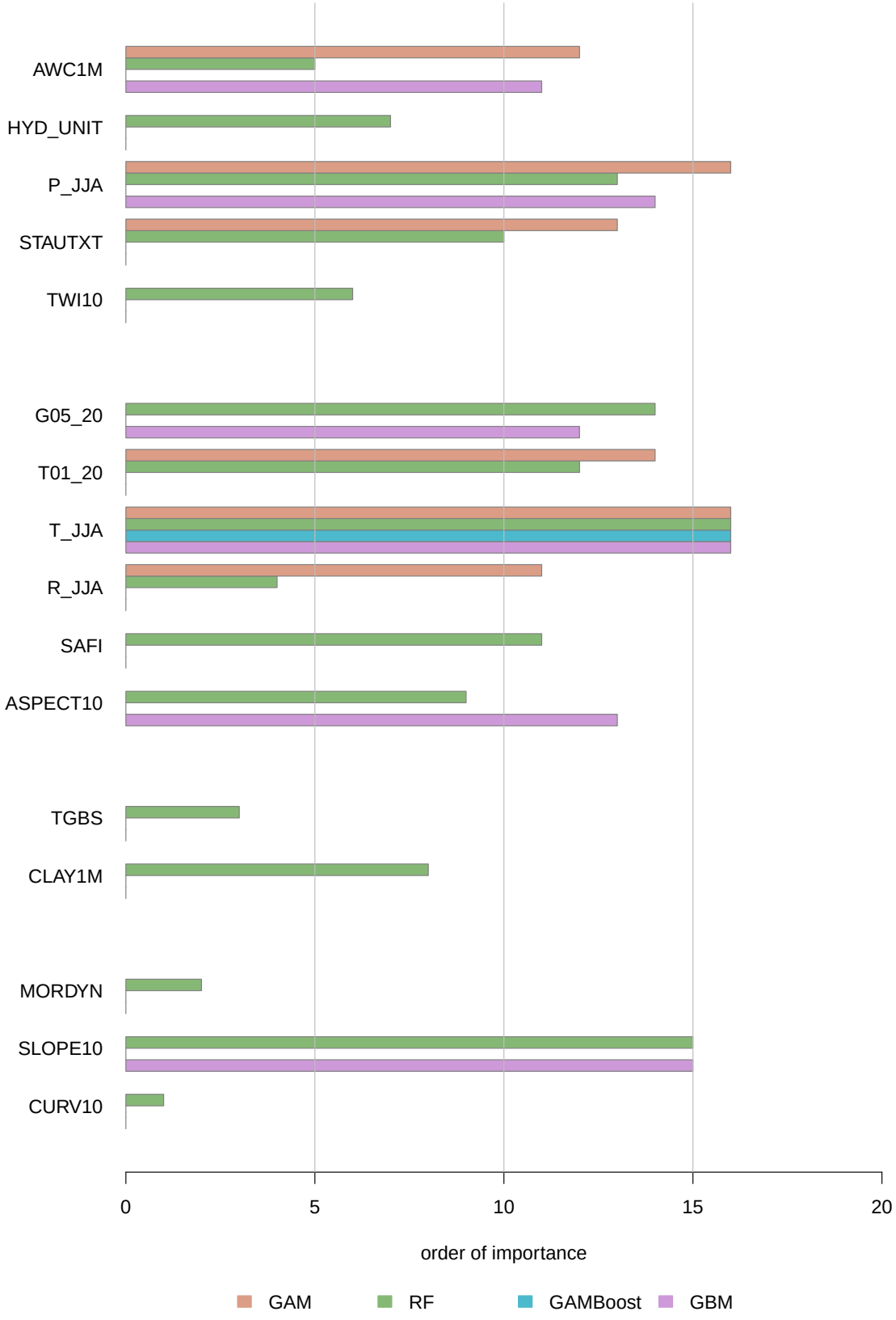


FIGURE 7.4.: *Variable importance: Swiss pine.*

Water

Precipitation (P_JJA) is the most important water variable for modelling the three tree species, whereas water supply seems to be more relevant for an ash than for a spruce and a Swiss pine. The pretty high importance of AWC1M in the GBM models indicates the relevance of this variable for two–way interactions.

Water logging is measured with two variables: STAUTXT and HYD_UNIT. The results of the importance analysis show, that all modelling techniques evaluate the importance of STAUTXT higher than HYD_UNIT.

The topographic wetness index (TWI10) is rather significant in the GAMBoost model for the ash. Thus, the smooth main effect of TWI10 provides, additionally to the precipitation in summer, valuable information about the water requirement of the ash.

Energy

Since two of the temperature variables, i.e. G05_20 and T_JJA, are strongly correlated, GAMBoost selects only that predictor, which is relevant for the model, as similarly seen in the simulation study. It is to note, that the absolute importance value of T01_20, which is also correlated with G05_20 and T_JJA, is marginal in the GAMBoost model of the ash. The most important temperature variable is identified in accordance to GBM.

For the spruce models, it is not clear, whether G05_20 or T_JJA is of higher relevance, whereas "temperature in summer" (T_JJA) is the major predictor for the models of the Swiss pine and "degree value days" for the ash models. T01_20 is the least contributing temperature predictor for all data–directed modelling techniques.

The importance of "radiation in summer" (R_JJA) appears to be tree–specific: its contribution to the spruce models is higher than for the other trees, in particular the Swiss pine.

The "slope aspect favourability index" (SAFI) and the "exposition" (ASPECT) play a minor role in modelling species distributions. Other energy variables are more important.

Nutrients

The predictors, which describe the nutrient balance of the soil are merely relevant for modelling the ash. Similarly to the appraisement of the ecological experts, the depth gradient of base saturation (TGBS) influences the distribution of the ash.

Geomorphodynamics

The slope angle (SLOPE10) of an observation point seems to have a fundamental influence on the distribution models of the spruce and the Swiss pine. Since GAMBoost does not recognise SLOPE10 as a contributing predictor for the latter tree species, it is suggested, that this variable is mainly important for interactions.

In opposition to the experts' opinion, the data–directed models indicate a higher impact of MORDYN on the models for the spruce than for the ash.

7.4. Response curves

Besides the predictive performance and the variable importance of the species distribution models, their ability to reflect the true environmental relationships is a matter of interest. For a more profound understanding of the models, the environmental niche is described along ecological gradients with response curves. The univariate case described in section 2.3 is now extended to multiple predictors.

Because of the large amount of predictor variables, the description of the response curves is limited to the most important variables. All predictors of the sparse expert models and the two most important variables of the data–directed models are analysed. A direct comparison between all modelling techniques is not possible, because the models differ in the predictor set as well as in variable importance.

To achieve comparable response curves for the four modelling strategies, the partial effects on the scale of the linear predictor are centred at zero and transformed on the scale of the response. The resulting curves for the ash, the spruce and the Swiss pine are depicted on the following pages.

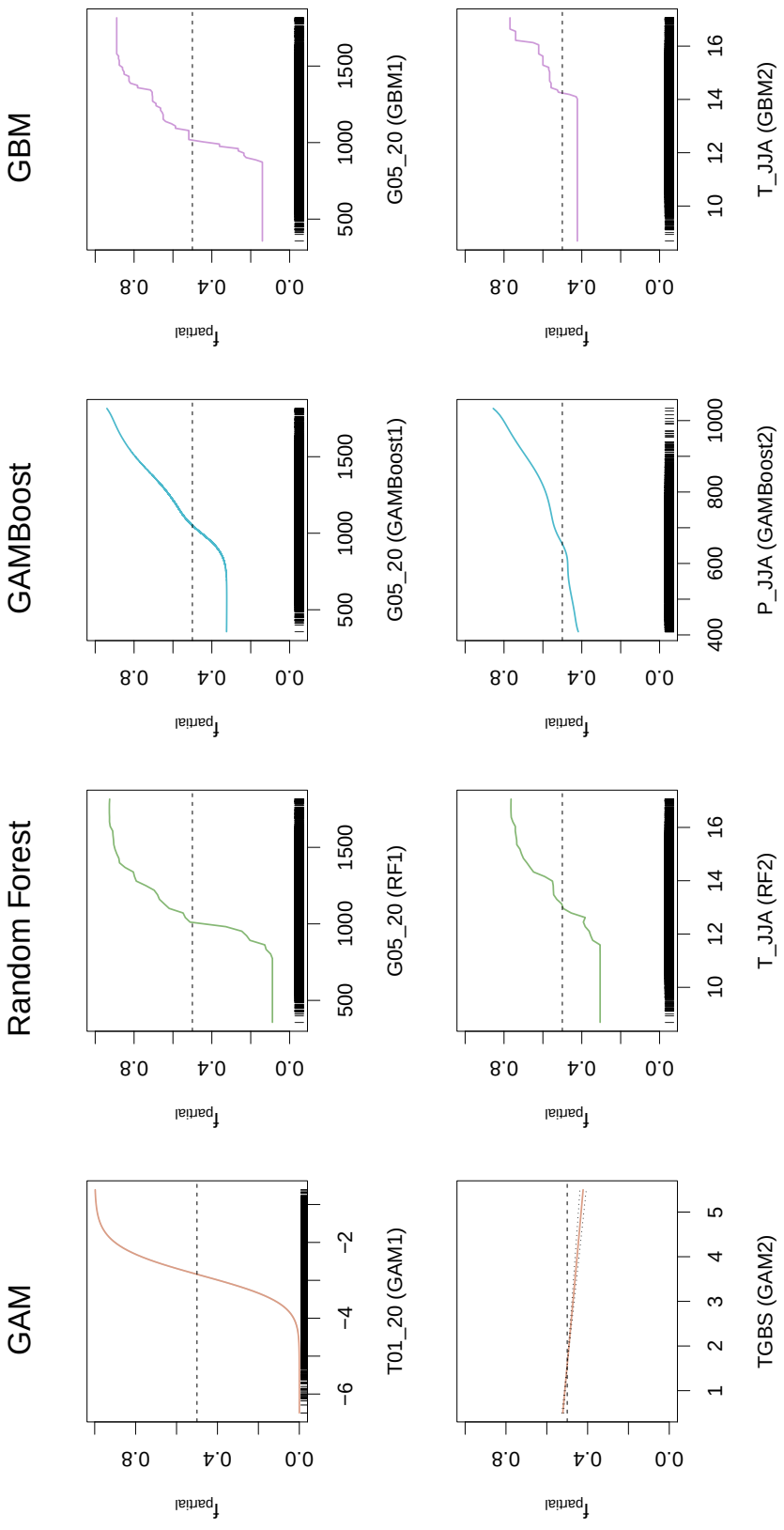


FIGURE 7.5.: *Response curves: ash.*

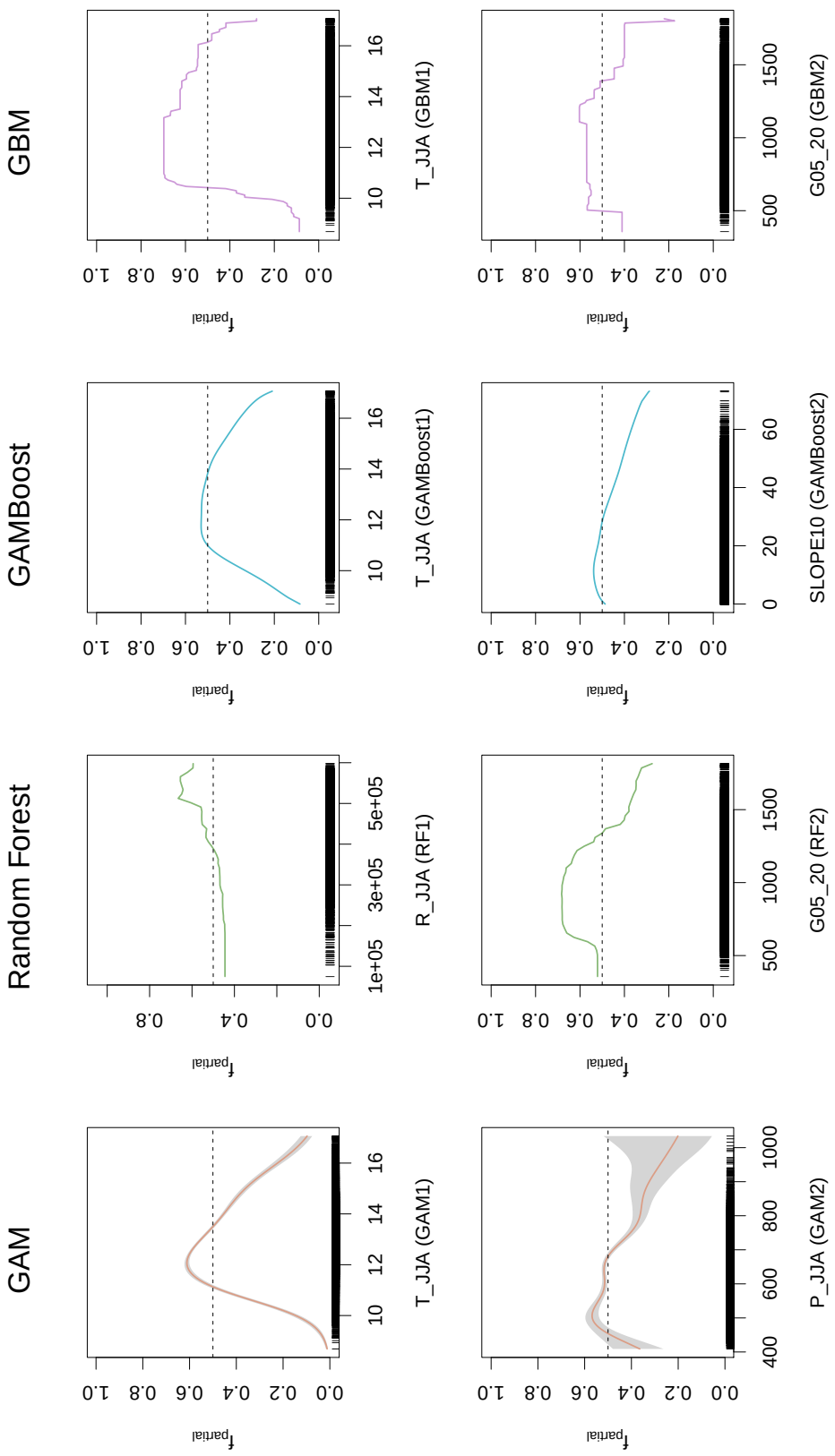


FIGURE 7.6.: *Response curves: spruce.*

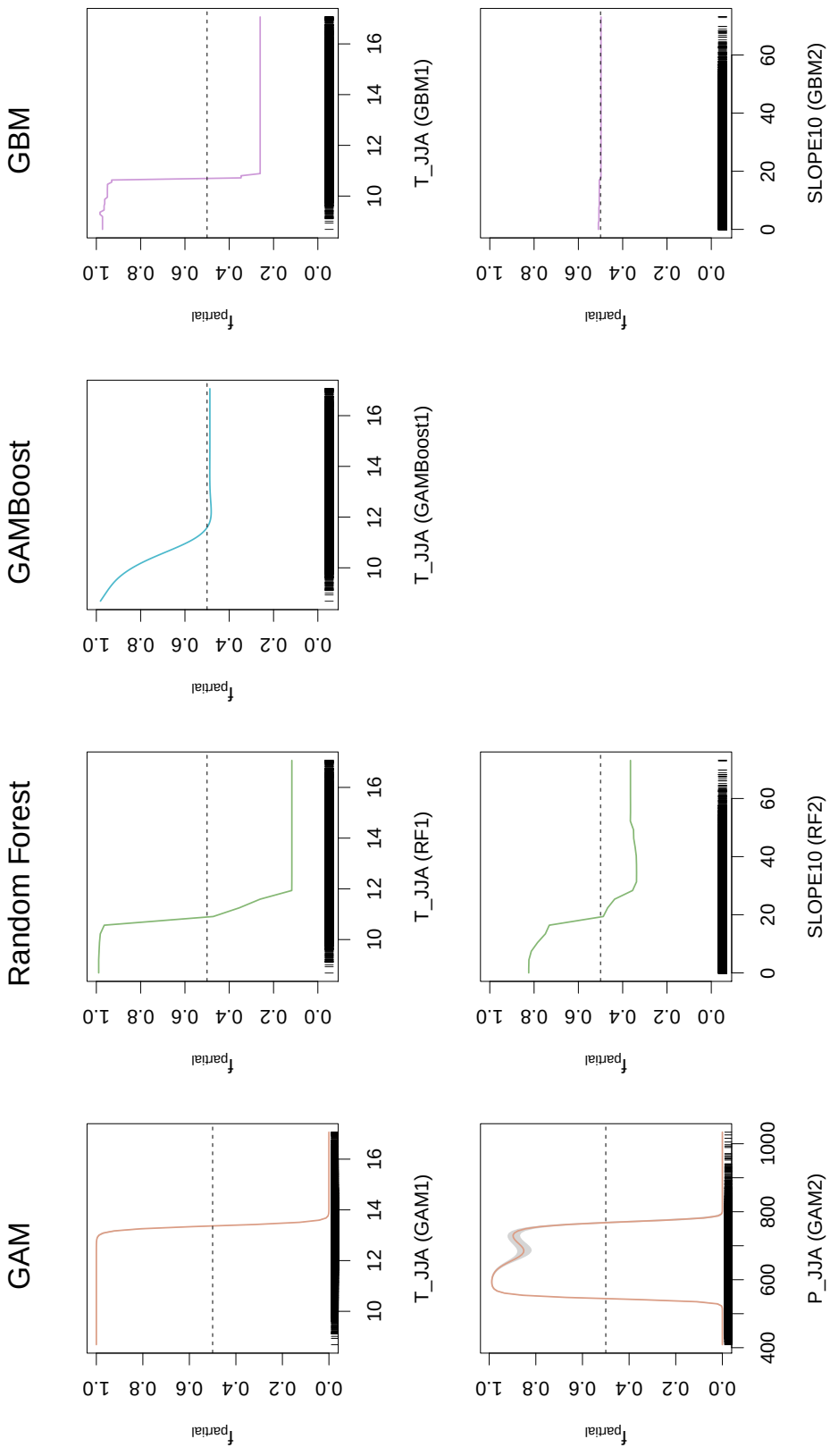


FIGURE 7.7.: *Response curves: Swiss pine.*

Ash

The data–directed models for the ash are mainly influenced by the predictor "growing degree days" (cf. figure 7.5: RF1, GAMBoost1, GBM1). All these models indicate an increasing species occurrence above 900 growing degree days. Below this boundary, the response curve is at a constant level. The upper limit for good growing conditions along this gradient is not reached, but saturation is visible for random forest and GBM above 1500 growing degree days. The response curve of the GAMBoost model is not that steep and on the whole, the effect is smaller.

Considering the partial effects of "temperature in summer" in the random forest and the GBM approach (cf. figure 7.5: RF2, GBM2), a right–shift of the increasing effect for GBM in comparison to random forest is evident. This means, that conditioned on the other predictors, T_JJA provides only at the upper boundary additional information for the GBM model. Perhaps another predictor describes the effect of the middle part of the underlying factor better, which could be a reason for the shifted curve.

The response shape of "January temperature" in the hypotheses–directed GAM (cf. figure 7.5: GAM1) resembles the curves of G05_20 of the data–directed models. The ash prefers high temperatures, above -3.5°C and the right edge is truncated.

The requirements of the ash regarding the nutrient supply is contained in the generalised additive model (cf. figure 7.5: GAM2). If the soil is highly saturated with bases, more ashes will occur.

Moreover, the growing limits in terms of the demand for precipitation are not reached within the underlying data, as the GAMBoost model depicts (cf .figure 7.5: GAM-Boost2). However, an increasing precipitation heightens the occurrence of the ash.

Spruce

Considering the effect of "temperature in summer" for the distribution of the spruce (cf. figure 7.6: GAM1, GAMBoost1, GBM1), the response curves are not truncated, slightly right–skewed and display the best growing conditions at about 11 to 13°C. The data–directed approaches reveal a broad top, which is not recognised with the GAM. The influence of temperature on the growth of the spruce seems to be stronger in the modelling with GAM and GBM.

The variable "growing degree days" is an important predictor for random forest and GBM (cf. figure 7.6: RF2, GBM2). Both models show a declining response curve for high values of the gradient, from about a value of 1200 degree value days. The positive effect decreases also at the left boundary and GAMBoost as well as GBM reach a level edge. G05_20 plays a slightly minor role for GBM than for the random forest model.

The characteristics of the spruce distribution in terms of precipitation are described with GAM (cf. figure 7.6: GAM2). The right–skewed curve reveals, that the upper limit of the ecological niche is covered by the WINALP data, whereas at the lower

boundary the truncated response curve tends to exhibit a negative effect. Best growing conditions are at about 500mm total precipitation in summer.

According to the random forest model, spruces prefer locations with rather high radiation (cf. figure 7.6: RF1). An upper limit is not clearly comprised within the data. Furthermore, the analysis of the slope gradient with the GAMBoost model (cf. figure 7.6) identifies a negative relationship between the slope of a location and the growth of spruces.

Swiss pine

The driving force of the distribution models of the Swiss pine is the temperature in summer. The shape of the response curves (cf. figure 7.7: GAM1, RF1, GAMBoost1, GBM1) is indeed similar for the four different types of models: low temperatures have a positive effect on the occurrence, whereas high temperatures are disadvantageous. However, the breaking point is considerably higher for the GAM than in the data–directed models. The effect of T_JJA is strongest pronounced for random forest and GAM.

The slope of an observation point (cf. figure 7.7: Rf2, GBM2) is a relevant predictor for the tree–based methods random forest and GBM and thus, the variable is mainly important for interactions. Both models indicate, that the Swiss pine prefers rather plain regions with a slope below 20%, but the partial effect of SLOPE10 in GBM is marginal and almost invisible.

As also the descriptive analyses in chapter 2 demonstrate, the ecological niche of the Swiss pine regarding precipitation, which is investigated with the GAM, is entirely described within the data set (cf. figure 7.7: GAM1). Good growing conditions for the Swiss pine will be provided, if the precipitation amount lies between 550 and 750mm.

7.5. Response surfaces

In order to account for unmeasured confounders and to prevent spatially autocorrelated residuals, spatial information in the form of a smooth effect of the coordinates is included in GAM and GAMBoost, as described theoretically in section 3.6.

The main difference between the two approaches with regard to the modelling of the spatial effect is, that the available degrees of freedom are estimated by cross–validation in the generalised additive model, whereas they are restricted for the GAMBoost model. Due to this limitation, the base procedure of the coordinates is not favoured, because of the higher number of degrees of freedom in comparison to the environmental variables. Thus, the predictor, which provides most information with at most four degrees of freedom is chosen in each boosting iteration.

Figure 7.8 depicts the spatial effects of GAM and GAMBoost for the ash and the spruce, in which the effects are centred and transformed to the response scale, similarly to the response curves of the environmental predictors. In consequence of implausibility, the spatial effect in the expert model for the Swiss pine is removed. Moreover, the spatial information is irrelevant in the GAMBoost model.

The results for the two ash models are quite different: The spatial effect is substantially higher and more varying in the generalised additive model, whereas in the GAMBoost model only a very weak effect exists.

The effects for the spruce differ mainly in their intensity, but the tendencies are quite similar. The eastern part of the study region is disadvantageous, whereas the western part demonstrates a fairly positive effect.

The main reason for the differing results originates in the various predictors, on which the models are based. This implicates, that the spatial confounders, which are modelled with the spatial effect, have also a different structure.

The benefit of the smooth effect of the coordinates to alleviate spatial autocorrelation has to be evaluated by means of analysing the spatial structure of the residuals. However, the quantification of spatial autocorrelation within large data sets raises computational difficulties, as described in section 7.7.

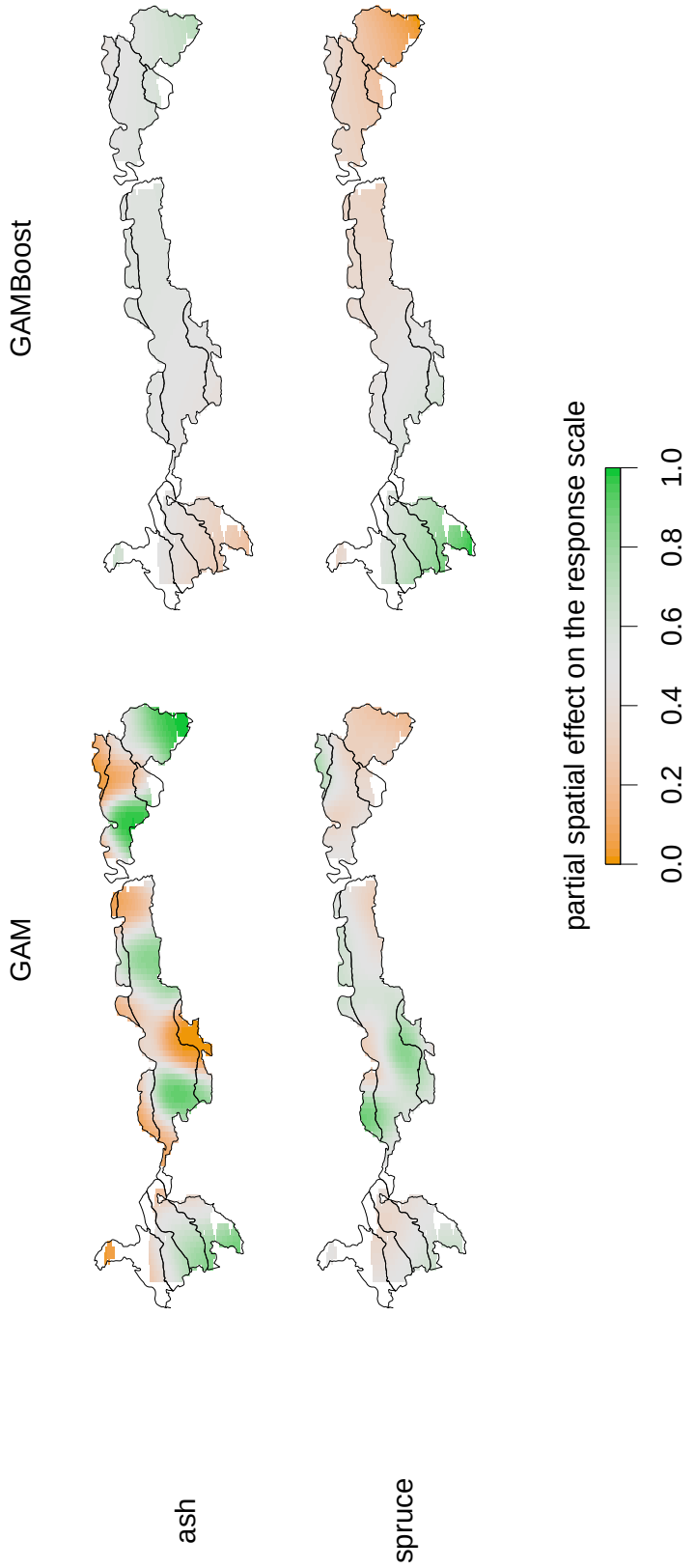


FIGURE 7.8.: *Spatial response surfaces: GAM and GAMBoost for the ash and the spruce.*

7.6. Prediction

The current phase of the WINALP project aims at the modelling of forest types in order to construct species suitability maps. For this purpose, it is essential to compare the different modelling techniques concerning the spatial demonstration of the predictions.

Exemplarily, the ash models are considered and illustrated, because the information content of probability maps for very seldom or very frequent tree species is low. The graphics for the spruce and the Swiss pine can be found in the appendix (figures: A.4, A.5, A.6, A.7). It is to note, that the colouring in the illustrations varies for the individual tree species and that the categories do not have equal distances.

Since the ash is a tree species, which prefers the warmer valleys, all predictive maps reveal a higher occurrence probability in these regions and the valleys are well to distinguish.

Whereas the generalised additive model and also the boosting models show a quite smooth predictive distribution of the ash, the map produced by the random forest is patchier. When the spruce and the Swiss pine are analysed, the random forest predictions do not show any spatially structured distribution at all.

Furthermore, the random forest estimates more extremely. This means, that medium occurrence probabilities only appear seldom.

For the ash model, the GAMBoost method calculates considerably higher values for the locations with lower probabilities compared to the other approaches . The tendency of GAMBoost to differ in the prediction of the prevailing class, i.e. presences or absences, can also be recognised in the models for the spruce and the Swiss pine. Additionally, residual analyses also detect these characteristics.

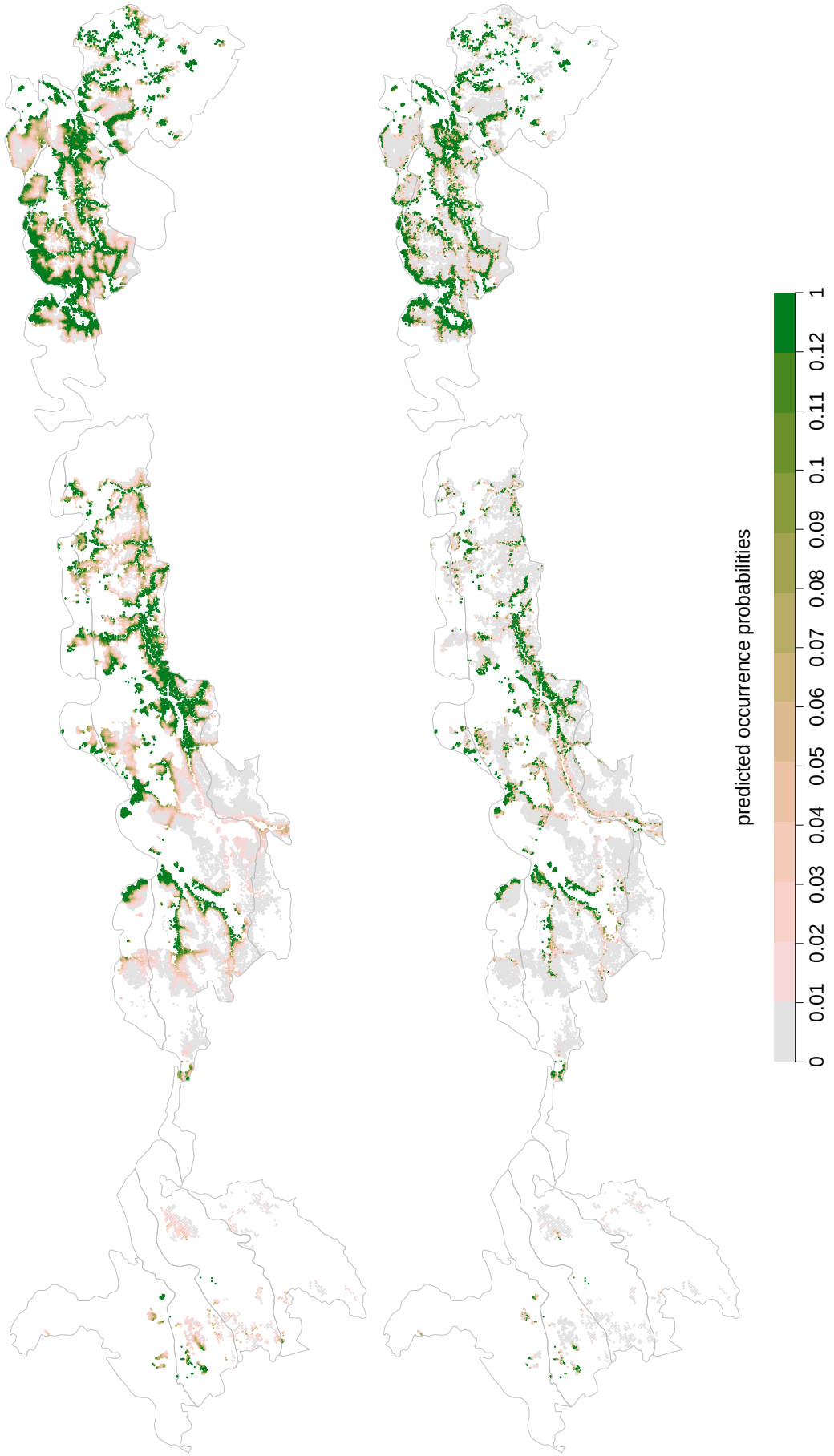


FIGURE 7.9.: *Spatial predictions for the ash: GAM (above) and RF (below).*

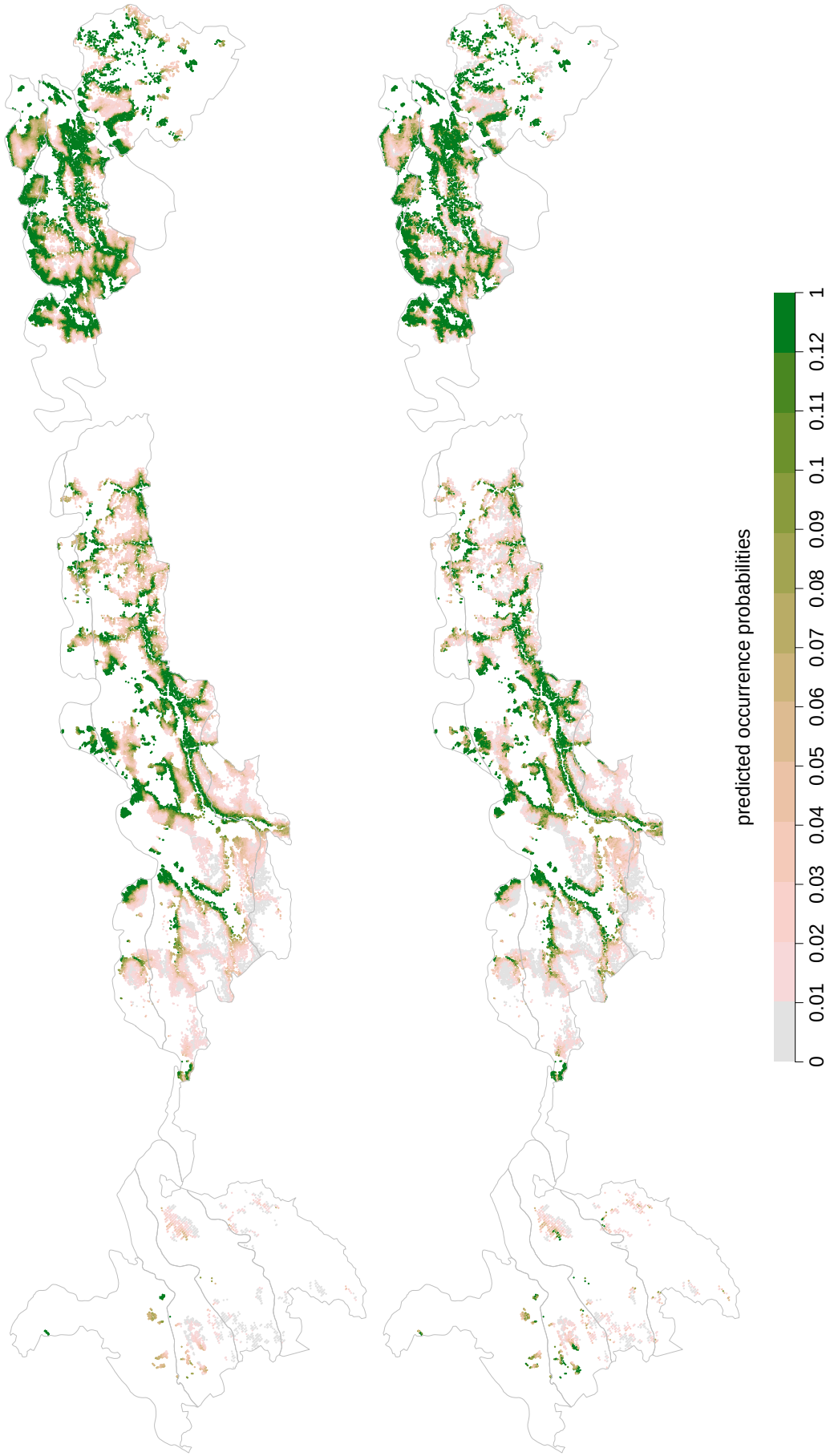


FIGURE 7.10.: *Spatial predictions for the ash: GAMBoost (above) and GBM (below).*

7.7. Limitations of the analysis

At the end of the data evaluation, some limitations of the applied techniques are reviewed.

Spatial autocorrelation in GAM

As mentioned in section 3.6, the allowance for spatial autocorrelation is essential to conform the requirement of independent errors in the regression model. Therefore, a smooth spatial trend of the coordinates is included in the GAM.

Though, looking at the residuals of the model, e.g. for the ash, a light, small–scaled spatial structure is identifiable:

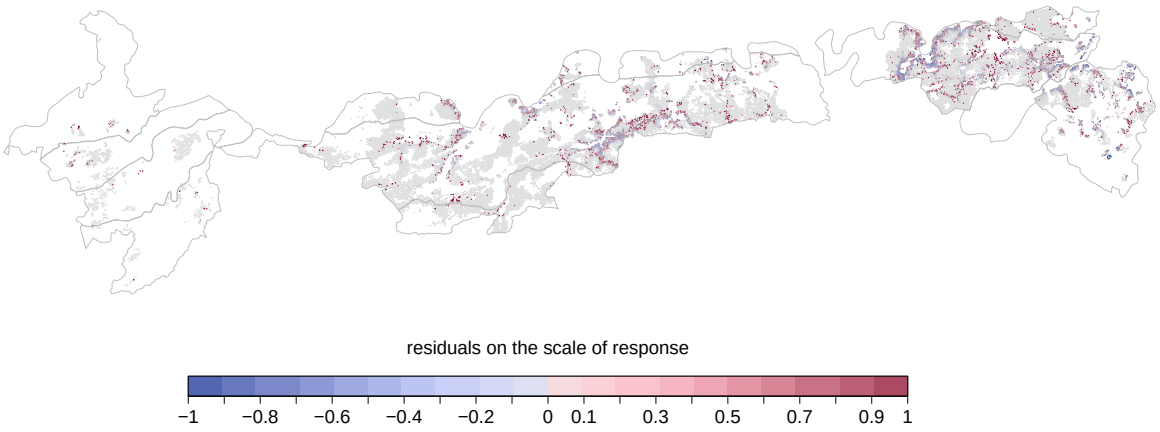


FIGURE 7.11.: *Residuals for the expert model of the ash.*

Obviously, the spatial trend is not able to remove spatial structure on a small range, as it is also suggested in Dormann et al. (2007).

Commonly applied approaches for exploring spatial autocorrelation include variograms and correlograms, in which the similarity between the residuals is depicted against the corresponding distance of the observation points. But also parametric methods, like kriging, are often used.

In ecology, Moran's I is widely used to quantify the spatial dependency of values, for instance of residuals (e.g. Legendre, 1993; Kühn, 2007). However, it must be pointed out, that Moran's I depends on the sample size. On that account, the absolute values of this measure do not indicate the strength of spatial autocorrelation. A permutation based test of significance on Moran's I provides more precise information on the existence of spatial autocorrelation, but also evokes a multiple test problem, if several lags are considered.

In addition to that, the calculation of Moran's I and also the parametric kriging approach is computationally demanding, because high–dimensional distance matrices have to be established. For that reasons, the quantification of spatial autocorrelation remains an open issue in the analysis of the large WINALP data set.

AUC

The area under the receiver operating characteristic curve measures the predictive performance only in terms of discriminatory power, while predictive accuracy is not incorporated. In particular, this becomes evident for the two boosting approaches: GBM and GAMBoost differ only slightly according to AUC, but the prediction maps of the GAMBoost models show higher occurrence probabilities, especially for the more prevalent class.

Lobo, Jiménez-Valverde and Real (2007) criticise the indiscriminate use of AUC and mention the influence of irrelevant thresholds and the equal treatment of sensitivity and specificity as further reasons. Thus, the quality of the models should not only be linked to the AUC criterion and the ROC curves, but also to other aspects, like plausibility of the results or residual diagnostic.

Does occurrence probability equal habitat suitability?

Using occurrence probabilities derived from statistical models for probability maps is a common practice in ecology. However, these maps only reflect the expected probability, with which a particular tree can be found under the given conditions of the site. It is questionable, to what extent this information renders a statement on the potential suitability of a site for the growing of a tree species.

7.8. Concluding remarks

The utilisation of statistical methods for the analysis of the WINALP data provides valuable insights into their adequacy for describing the habitat of tree species. Different aspects of the models were examined with the result, that the suitability of the methods varies depending on the considered aspect.

Generalised additive model

Indeed, the attempts to account for interactions and to imitate a balanced sample by weighting are abandoned because of a too small benefit and implausibility, but based on expert knowledge parsimonious and in terms of discriminatory power competitive models are established.

Due to the aim of accomplishing sparse models, it is necessary to account for the unmeasured and in particular for the disregarded predictors by the inclusion of a spatial effect. Thus, also the models are improved according to the GCV criterion, though the entire spatial structure of the data cannot be explained.

The importance of environmental predictors influencing the ecological mechanisms of a species' niche is difficult to evaluate with a generalised additive model, because the order of removal in a backward selection procedure serves only as a rough guide. Furthermore, the relevance in the analysis of the WINALP data depends on the preselection of the variables, which was necessary in order to prevent a multicollinearity problem.

Huge effort of the ecological expert was required in order to build models with ecologically reasonable response shapes. In comparison to the data–directed models, the resulting partial effects tend to be more extreme and also the shapes differ to some extent. Thus, the response curves of the GAM do not display the underlying structure of the data precisely.

Random forest

Since the theoretical concept of random forest is designed for discrimination, the predictive performance in terms of AUC is for all tree species among the best in comparison to the other modelling techniques. But another consequence of concentrating mainly on discrimination is, that the spatial predictions are quite uneven, in particular the pattern is spotted, if one class is very predominant.

Especially the variable importance measure for random forests provides valuable information on the relevance of the predictors in the complex model including also high–dimensionally interactive relationships and can be used as a guideline for the selection of important predictors for a parametric model. The application of random forest based on conditional inference trees properly allows for categorial variables in the tree building process.

As a consequence of the aggregation proceeding in the random forest algorithm, the response curves are comparatively smooth. The partial effects are highly informative, but their calculation is computationally very intensive.

Boosted generalised additive models

Besides the GAMBoost models tend to perform slightly worse in terms of discriminatory power, the smooth response curves and the resulting variable importance measure provide useful information on the smooth main effects within the data generating mechanism. Therefore, the implicit variable selection procedure includes only one of highly correlated predictors, namely the predictor with the higher informational content regarding the main effect.

In addition to that, GAMBoost offers an opportunity to account for unmeasured confounders with a smooth spatial effect, which will be only preferred to environmental predictors, if, based on the same number of degrees of freedom, more variation can be explained.

The prediction maps have a similar structure compared to GAM, but the estimations tend to be less extreme for the more prevalent class.

Boosted regression trees

The predictive performance of boosted regression trees on the WINALP data set is comparable to GAMBoost. The response curves display a similar shape, but are wigglier, especially for less important predictors. The reason thereof is, that weak decision trees are used as baselearners, which produce piecewise constant predictions.

The relevance of the predictors in a boosted regression tree reflects the importance of interactions of the predefined depth, which renders further information in addition to the rather overall importance of random forest and the importance of the main effects provided by GAMBoost.

The structure of the predictive map resembles the corresponding map of the generalised additive model and the values seem to be slightly higher.

8. Summary and perspectives

8.1. Concluding reflection of the acquired results

Statistical models for the description of the ecological niche of species are of great interest for the WINALP project and also generally in ecology, because the spatial predictions serve as species suitability maps and provide an important base for site–specific forest management.

A hypotheses–directed approach in terms of a generalised additive model is contrasted to data–directed strategies, particularly random forest and boosting. The models are evaluated on real data as well as with a simulation study.

The selection of an appropriate model depends on the structural assumptions regarding the underlying mechanisms, but also on the properties of the model. Because the relationships between species and environments are often largely unknown and vary between the different tree species, the comparison of several models related with expert knowledge is essential.

The imitation of a stratified sample by weighting seems to improve the fit of GAMs, but the quantification of the profit and the interpretation is questionable.

Moreover, in accordance to Hirzel and Guisan (2002), a spatially balanced sample ameliorates the accuracy of GAMs. However, this tendency is not detected by the usage of the data–directed approaches.

Including a spatial effect in order to allow for unmeasured confounders enhances the species distribution models with GAMs for the WINALP data, but does not remove the entire spatial structure and cannot compensate for the disregard of a small–scaled predictor.

The predictive capacity of GAMs within the range of the training data is comparable to the data–directed approaches including the good discriminatory power of random forest. The properties of GAMs for extrapolated data are favourable and more precise estimations are delivered in comparison to the data–directed models.

Since the data–directed approaches make fewer assumptions on the structure of the model, they are more appropriate for reflecting the real relationships in the data, especially with regard to partial effects.

A differentiated insight into the habitat properties of the examined tree species is reflected by the importance measures of the various data–directed models: the relevance of the predictors concerning overall and interactive relevance as well as the impact of

the main effects can be analysed. However, the exploration of the influence of correlated predictors on the importance measure of the boosted regression trees displayed unclear properties.

The results show, that on the one hand, the incorporation of expert knowledge offers models with comparably good predictive power, and that on the other hand, data–directed models are also able to provide insights into the data generating process.

8.2. Outlook

Based on this work, a recalibration of the actual parametric species distribution models of the WINALP project with the acquired results on the variable importance, the response shapes and the properties of the models is possible. Data–directed models turned out to render valuable information on the habitat of the species and thus, they are worthwhile to be applied to the further tree species in the project.

Future studies within the WINALP project address to the incorporation of data from southern Germany and Europe on a smaller resolution in order to comprise a larger range of the ecological niche of an examined species and to account for supra–regional effects. Multilevel models will be accomplished by including the predictions of the higher levels into the models for the Northern Alps.

Several suggestions were proposed to account for spatial autocorrelation in regression models, including wavelet analysis (Carl and Kühn, 2008), autoregressive models, spatial generalised linear mixed models and spatial generalised estimating equations (Dormann et al., 2007). However, the usage of these methods is restricted either to data on a regular grid or to linear effects.

The last–mentioned deficit can be offset by using polynomial predictor terms, whose degrees are estimated through a GAM without allowing for spatial autocorrelation or by applying general additive mixed models. Large computational effort in order to create $n \times n$ correlation matrices excludes the application of these approaches so far, even for a sparse matrix.

Current statistical research concerns the improvement of the variable selection properties of boosting. Bühlmann and Hothorn (2010) suggest an advanced boosting algorithm, “Twin Boosting”, in order to obtain sparser models.

A. Appendix

A.1. Table of data files for R–code

On the appended CD, the programming code for the data evaluation and the simulation study is provided. In the following, the contents will be listed.

- ▷ 01_DataManagement.r: *Data management*
- ▷ 01_TestTrain.r: *Division of the data in test and training data set*
- ▷ 02_DescriptiveAnalysis.r: *Descriptive analysis of the data set*
- ▷ 03_ GamModels.r: *Modelling using GAM*
- ▷ 03_ GamCalc.r: *Calculation of the predictions of GAM*
- ▷ 04_ RfModels.r: *Modelling using random forest*
- ▷ 04_ RfCalc.r: *Calculation of the predictions and the variable importance of random forests*
- ▷ 05_ BoostingModels_GamB.r: *Modelling using boosting: GAMBoost*
- ▷ 05_ BoostingModels_Gbm.r: *Modelling using boosting: GBM*
- ▷ 05_ BoostingCalc.r: *Plotting of model selection, calculation of predictions and variable importance of boosting*
- ▷ 06_ DataSim.r: *Generating of simulation data sets*
- ▷ 06_ SimModels_Gam.r: *GAM using simulated data*
- ▷ 06_ SimModels_Rf.r: *Random forests using simulated data*
- ▷ 06_ SimModels_GamB.r: *GAMBoost using simulated data*
- ▷ 06_ SimModels_Gbm.r: *GBM using simulated data*
- ▷ 06_ SimCalc_Gam.r: *Simulation study: predictions and variable importance of GAM*
- ▷ 06_ SimCalc_Rf.r: *Simulation study: predictions and variable importance of random forest*
- ▷ 06_ SimCalc_GamB.r: *Simulation study: predictions and variable importance of GAMBoost*

- 06_ SimCalc_Gbm.r: *Simulation study: predictions and variable importance of GBM*
- 06_ SimResults.r: *Results for chapter 6*
- 07_ GamResults.r: *Results for chapter 7.1*
- 07_ ResultsPerf.r: *Results for chapter 7.2*
- 07_ ResultsImp.r: *Results for chapter 7.3*
- 07_ ResultsResp.r: *Results for chapters 7.4 and 7.5*
- 07_ ResultsPred.r: *Results for chapter 7.6*
- 07_ ResultsResid.r: *Results for chapter 7.7*
- AuxiliaryRoutines:
  - 00_Coding.r: *Coding of variables and levels*
  - 00_FormulaCov: *Diverse auxiliary functions for the creation of model formulas*
  - 00_Shape.r: *Function for adding topographical information to a spatial plot*
  - 01_DivideTestTrain.r: *Function for the division of the data in test and training data set*
  - 02_DescHelpers.r: *Diverse auxiliary functions for the descriptive analysis*
  - 03_Weighting.r: *Function for the calculation of the product weights*
  - 03_GAM.r: *Function for the computation of the GAMs*
  - 03_InteractMod.r: *Function for adding an interaction factor to a GAM*
  - 03_VarSel.r: *Function for variable selection with GCV*
  - 04_CForestResponse.r: *Function for calculating the partial effects of a random forest*
  - 05_VarImpGamB.r: *Function for calculating the variable importance of GAMBoost*
  - 06_Cv.r: *Function for creating a cross–validation matrix for GAMBoost*
  - 06_SimResultsHelpers.r: *Functions for illustrating the results of the simulation study*
  - 07_MyVisGam.r: *Modification of vis.gam()*
  - 07_GamCol.r: *Modification of plot.gam()*
  - 07_Pred.r: *Function for plotting predictions*
  - 07_Resid.r: *Function for plotting residuals*

A.2. Additional graphics: Simulation study

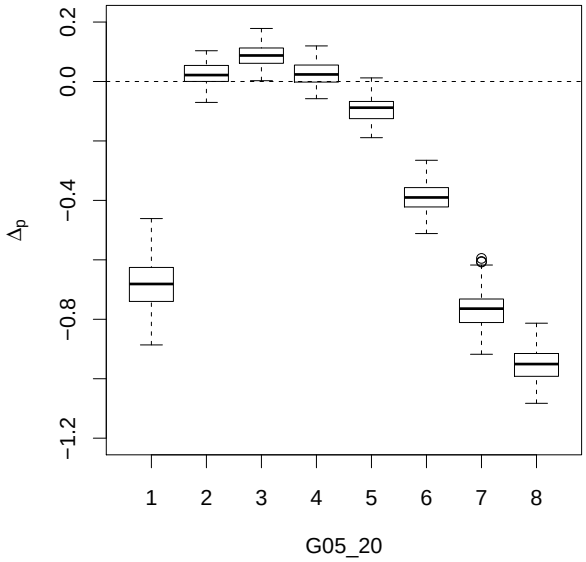


FIGURE A.1.: *Performance of Random Forest in the initial scenario subdivided into intervals of "degree value days":* $(-\infty, 600], (600, 800], \ldots, (1600, 1800], (1800, \infty)$.

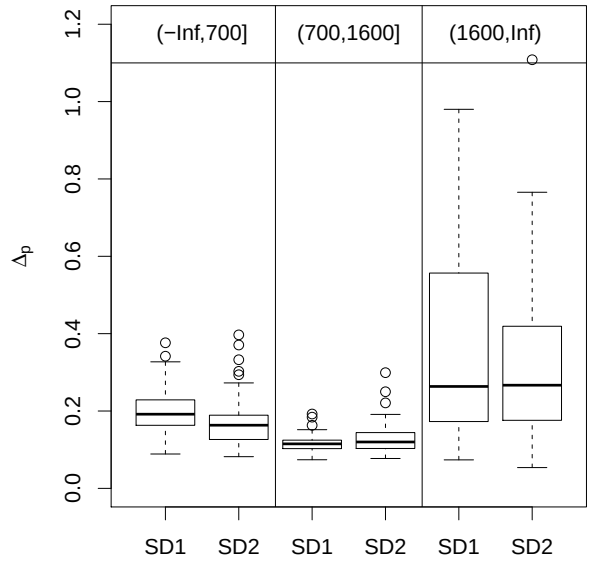


FIGURE A.2.: *Performance* Δ_p^a *of GAMBoost in scenarios SD1 and SD2 subdivided into intervals of the predictor "degree value days".*

A.3. Additional graphics: Data analysis

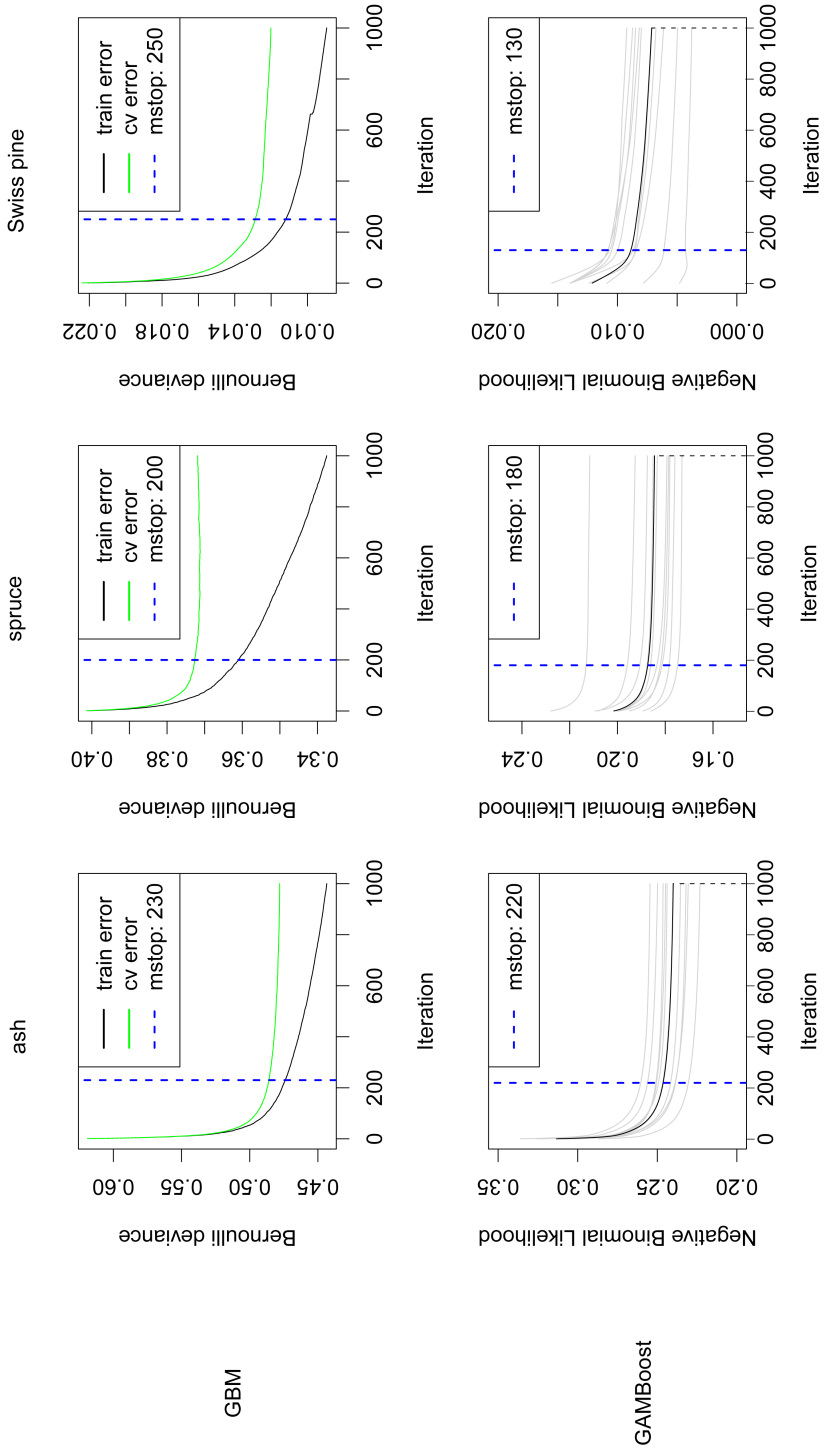


FIGURE A.3.: *Tuning parameter selection: Boosting.*

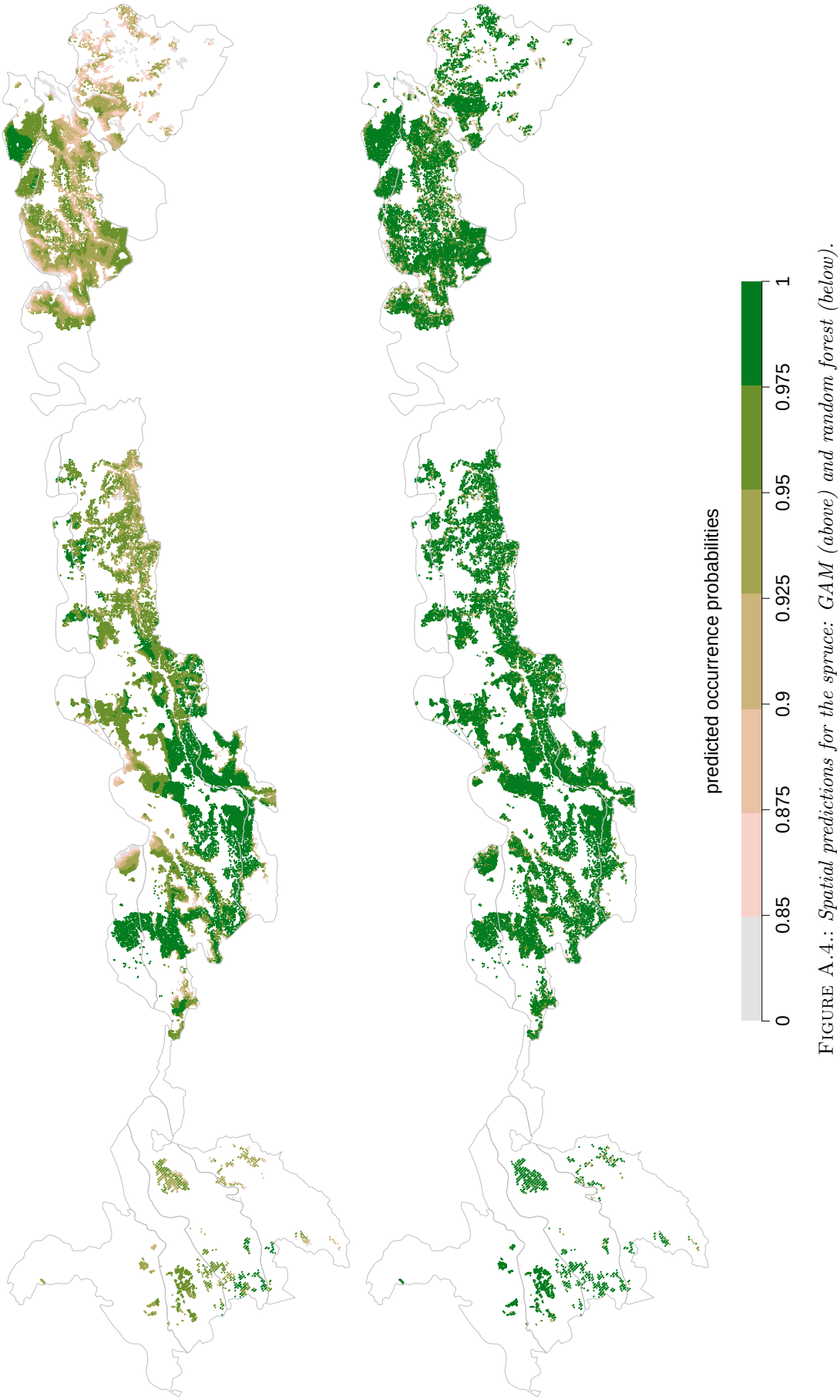


FIGURE A.4.: *Spatial predictions for the spruce: GAM (above) and random forest (below).*

FIGURE A.5.: *Spatial predictions for the spruce: GAMBoost (above) and GBM (below).*

FIGURE A.6.: *Spatial predictions for the Swiss pine: GAM (above) and random forest (below).*

FIGURE A.7.: *Spatial predictions for the Swiss pine: GAMBoost (above) and GBM (below).*

Bibliography

Araújo, M. and Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography* **16**, 743–753.

Austin, M. (1987). Models for the analysis of species' response to environmental gradients. *Vegetatio* **69**, 35–45.

Austin, M. (1999). A silent clash of paradigms: some inconsistencies in community ecology. *Oikos* **86**, 170–178.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* **157**, 101–118.

Austin, M., Belbin, L., Meyers, J., Doherty, M., and Luoto, M. (2006). Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. *Ecological Modelling* **199**, 197–216.

Austin, M. and Smith, T. (1989). A new model for the continuum concept. *Vegetatio* **83**, 35–47.

Bartlett, P. and Traskin, M. (2007). Adaboost is consistent. *Journal of Machine Learning Research* **8**, 2347–2368.

Bickel, P. J. and Ren, J.-J. (2001). The bootstrap in hypothesis testing. *Lecture Notes-Monograph Series* **36**, 91–112.

Breiman, L. (1996a). Bagging predictors. *Machine Learning 24*(2), 123–140.

Breiman, L. (1996b). Out-of-bag estimation. Technical Report, Department of Statistics, University of California at Berkeley, CA, USA.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics 26*(3), 801–849.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation 11*(7), 1493–1517.

Breiman, L. (2001a). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science* **16**, 199–231.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International Group.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22*(4), 477–505.

Bühlmann, P. and Hothorn, T. (2010). Twin boosting: improved feature selection and prediction. *Statistics and Computing 20*(2), 119–138.

Bühlmann, P. and Yu, B. (2003). Boosting with the l_2 loss: Regression and classification. *Journal of the American Statistical Association* **98**, 324–339.

Carl, G. and Kühn, I. (2008). Analyzing spatial ecological data using linear regression and wavelet analysis. *Stochastic Environmental Research and Risk Assessment* **22**, 315–324.

Caswell, H. (1988). Theory and models in ecology: a different perspective. *Ecological Modelling* **43**, 33–44.

Cutler, D., Edwards Jr., T., Beard, K., Cutler, A., Hess, K., Gibson, J., and Lawler, J. (2007). Random forests for classification in ecology. *Ecology 88*(11).

De'ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology 88*(1), 243–251.

De'ath, G. and Fabricius, K. (2000). Classification and regression trees a powerful yet simple technique for ecological data analysis. *Ecology 81*(11), 3178–3192.

Dormann, C., McPherson, J., AraÃºjo, M., Bivand, R., Bolliger, J., Carl, G., Davies, R., Hirzel, A., Jetz, W., Kissling, D., Kühn, I., Ohlemüller, R., Peres-Neto, P., Reineking, B., Schröder, B., Schurr, F., and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30**, 609–628.

Elith, J. and Graham, C. (2009a). Do they? how do they? why do they differ? on finding reasons for differing performances of species distribution models. *Ecography* **32**, 66–77.

Elith, J. and Graham, C. (2009b). Do they? how do they? why do they differ? on finding reasons for differing performances of species distribution models. *Ecography* **32**, 66–77.

Elith, J., Leathwick, J., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology 802-813*(77).

Fahrmeir, L., Kneib, T., and Lang, S. (2007). *Regression - Modelle, Methoden und Anwendungen*. Springer Berlin Heidelberg New York.

Fitzpatrick, M. and Hargrove, W. (2009). The projection of species distribution models and the problem of non–analog climate. *Biodiversity and Conservation* **18**, 2255–2261.

Freund, Y. and Schapire, R. (1995). A decision-theoretic generalization of on–line learning and an application to boosting. *Proceedings of the Second European Conference on Computational Learning Theory*. Springer, Berlin.

Friedman, J. (1991). Multivariate adaptive regression splines. *Annals of Statistics* **19**, 1–141.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*(5), 1189–1232.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics 28*(2), 337–407.

Gauch Jr., H. and Whittaker, R. (1972). Coenocline simulation. *Ecology 53*(3), 446–451.

Guisan, A., Edwards, T., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* **157**, 89–100.

Guisan, A., Zimmermann, N., Elith, J., Graham, C., Phillips, S., and Peterson, A. (2007). What matters for predicting the occurrences of trees: techniques, data, or species´ characteristics? *Ecological Monographs 77*(4), 615–630.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer–Verlag.

Hirzel, A. and Guisan, A. (2002). Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling* **157**, 331–341.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2009). mboost: Model–based boosting. R package version 1.1-4.

Hothorn, T., Hornik, K., Strobl, C., and Zeileis, A. (2009). party: A laboratory for recursive partytioning. R package version 0.9-999.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*(3), 651–674.

Huisman, J., Olff, H., and Fresco, L. (1993). A hierarchical set of models for species response analysis. *Journal of Vegetation Science* **4**, 37–46.

Hurlbert, S. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs 54*(2), 187–211.

James, G. and Hastie, T. (1997). Generalizations of the bias/variance decomposition for prediction error. pp. Dept. Statistics, Stanford Univ., Stanford, CA, Tech. Rep.

Johnson, J. and Omland, K. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution* **19**, 616–625.

Kearns, M. and Valiant, L. (1994). Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the Association for Computing Machinery* **41**, 67–75.

Krebs, C. (1985). *Ecology. The experimental analysis of distribution and abundance* (4 ed.). Harper and Row, New York.

Kühn, I. (2007). Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions* **13**, 66–69.

Lawler, J., White, D. Neilson, R., and Blaustein, A. (2006). Predicting climate–induced range shifts: model differences and model reliability. *Global Change Biology* **12**, 1568–1584.

Leathwick, J., Elith, J., and Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling* **199**, 188–196.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology 74*(6), 1659–1673.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association 101*(474), 578–590.

Liu, C., Berry, P., Dawson, T., and Pearson, R. (2005). Selecting thresholds of occurrence in the prediction of species distribution. *Ecography* **28**, 385–393.

Lobo, J., Jiménez-Valverde, A., and Real, R. (2007). Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* **17**, 145–151.

Ludwig, J. and Reynolds, J. (1988). *Statistical ecology: a primer on methods and computing*. John Wiley and Sons.

MacArthur, R. (1972). *Geographical Ecology: Patterns in the Distribution of Species*. Harper and Row, New York.

Maggini, R., Lehmann, A., Zimmermann, N., and Guisan, A. (2006). Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography* **33**, 1729–1749.

Marx, B. and Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis* **28**, 193–209.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

Meynard, C. and Quinn, J. (2007). Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography* **34**, 1455–1469.

Mohler, C. (1983). Effect of sampling pattern on estimation of species distributions along gradients. *Vegetatio 54*(2), 97–102.

Moisen, G. and Frescino, T. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling* **157**, 209–225.

Oksanen, J. and Minchin, P. (2002). Continuum theory revisited: what shape are species responses along ecological gradients? *Ecological Modelling* **157**, 119–129.

Olden, J., Lawler, J., and Poff, N. (2008). machine learning methods without tears: a primer for ecologists. *The Auarterly Reviw of Biology 83*(2), 171–193.

Oppel, S., Strobl, C., and Huettmann, F. (2009). Alternative methods to quantify variable importance in ecology. Technical Report, Department of Statistics, University of Munich.

Prasad, A., Iverson, L., and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9**, 181–199.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reineking, B. and Schröder, B. (2006). Constrain to perform: Regularization of habitat models. *Ecological Modelling* **193**, 675–690.

Ridgeway, G. (2007). gbm: Generalized boosted regression models. R package version 1.6-3.

Strobl, C., Boulesteix, A.-L., and Augustin, T. (2007). Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis 52*(1), 483–501.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9*(1), 307.

Strobl, C., Boulestix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8–25.

Thuiller, W., Araújo, M., and Lavorel, S. (2003). Generalized models vs. classification tree analysis: Predicting spatial distributions of plat species at different scales. *Journal of Vegetation Science* **14**, 660–680.

Thuiller, W., Brotons, L., Araújo, M., and Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography* **27**, 165–172.

Toft, C. (1990). Reply to seaman and jaeger: An appeal to common sense. *Herpetologica 46*(3), 357–361.

Wood, S. (2006). *Generalized Additive Models - An Introduction with R*. Chapman & Hall/CRC.

Wood, S. (2009a). mgcv: Generalized additive model selection. http://www.maths.bath.ac.uk/ sw283/.

Wood, S. (2009b). mgcv: Multiple smoothing parameter estimation by gcv or ubre. http://www.maths.bath.ac.uk/ sw283/.

Wood, S. and Augustin, N. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling* **157**, 157–177.

Yee, T. and Mitchell, N. (1991). Generalized additive models in plant ecology. *Journal of Vegetation Science* **2**, 587–602.

List of Figures

Erklärung

Hiermit bestätige ich, Veronika Fensterer, dass ich die vorliegende Diplomarbeit selbständig und ohne Benutzung anderer als den angegebenen Hilfsmitteln angefertigt habe.

München, den 1. Juni 2010 Veronika Fensterer