LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Julia Schiele

Habitatmodelle für Brutvögel in Bayern

Bachelorarbeit
Betreuer: Prof. Dr. Torsten Hothorn
Institut für Statistik – Ludwig-Maximilians-Universität München

24. Oktober 2010

Danksagung

Diese Bachelorarbeit entstand am Institut für Statistik der Ludwig-Maximilians-Universität München in Zusammenarbeit mit dem Bayerischen Landesamt für Umwelt, Staatliche Vogelschutzwarte Garmisch-Partenkirchen, das mir die Daten für dieses interessante Thema zur Verfügung gestellt hat.

An dieser Stelle möchte ich mich bei meinen Betreuern Prof. Dr. Torsten Hothorn, Esther Herberich und Nikolay Robinzonov bedanken für die freundliche und engagierte Betreuung und die hilfreichen Gespräche und Anregungen.

Desweiteren geht mein Dank an meine Freunde nah und fern, die mich sehr unterstützt haben (sowohl auf kompetenter Fachebene als auch auf menschlich-emotionaler Ebene)

Und last but definitly not least an meine Mum, auf die ich mich immer verlassen kann.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

In einem Habitatmodell werden die verschiedenen Lebensräume von Tieren und Pflanzen untersucht, dabei sollen mögliche kausale Beziehungen zwischen den verschiedenen Habitateigenschaften und dem Vorkommen einer Art modelliert werden. Dies soll am Beispiel des Grauspechts und des Wespenbussards dargestellt werden.

Das Verbreitungsgebiet des Grauspechts erstreckt sich über weite Teile Zentral-, Nord- und Südosteuropas, sowie in einem breiten Gürtel bis an die pazifische Küste. Im Wesentlichen liegt die Nordgrenze des Verbreitungsgebietes im Übergangsbereich zwischen geschlossenem Nadelwald und aufgelockertem Laubmischwald, die Südgrenze verläuft in jenen Regionen, in denen die Baumsteppe in baumlose Strauch- und Buschsteppe übergeht. Innerhalb seines großflächigen und weiträumigen Verbreitungsgebietes ist der Grauspecht in Europa nirgendwo häufig. In Deutschland brüten etwa 15.000 Paare (Bauer und Berthold, 1997). Der Grauspecht gehört zu den recht schwer zu erfassenden Arten, da vor allem Einzelbrüter eine geringe Rufaktivität zeigen. Isolierte Reviere werden daher oft übersehen. Aus diesem Grunde unterliegen Bestandsangaben einer beträchtlichen Unschärfe. In Europa sind die Bestände zurzeit stabil beziehungsweise nehmen in einigen Staaten sogar leicht zu, ein Umstand, der möglicherweise aber ein Scheineffekt und nur auf die bessere Erfassung dieser Art in den letzten Jahren zurückzuführen ist.

Das Verbreitungsgebiet des Wespenbussards umfasst den größten Teil Europas sowie das südwestliche Sibirien. Der Wespenbussard bewohnt zumindest teilweise bewaldete Landschaften aller Art; bevorzugt werden Waldbereiche, die durch Lichtungen oder abwechslungsreiche Ränder strukturiert sind oder die in der Nähe zu abwechslungsreichen Feuchtgebieten liegen. Das regelmäßige Vorkommen reicht vom Flachland bis in die montane Stufe, höchste Brutnachweise erfolg-

ten in den Alpen auf etwa 1500 m. Nicht nur auf Grund der Tatsache, dass es sich um einen Zugvogel handelt und große Ähnlichkeit mit dem Mäusebussard besteht, sondern auch durch seine unauffällige Lebensweise ist es schwer, den Bestand der Wespenbussarde korrekt einzuschätzen. Für Deutschland wurden um das Jahr 2002 4000-4900 Paare angegeben. Die Meinungen über der Gefährdung des Wespenbussards gehen auseinander: weltweit gilt er als nicht gefährdet, auch in Deutschland gilt der Bestand insgesamt als ungefährdet (von Blotzheim *et al.*, 1989). In Bayern jedoch ist der Wespenbussard laut Roter Liste bereits als gefährdet eingestuft. Nicht nur deshalb ist es wichtig, möglichst gute Habitatmodelle zu erstellen, um die Lebensräume und damit die Arten besser schützen zu können.

Statistische Methoden für Artverbreitungsmodelle in der Biologie sind vielfältig. Bei bisherigen Methoden gibt es jedoch oft Schwierigkeiten mögliche nichtlineare Effekte, Interaktionen, Autokorrelationen oder Nicht-Stationarität in die Modellgleichungen aufzunehmen. Räumliche Autokorrelation erklärt sich dadurch, dass das Vorkommen einer Art durch räumliche Nähe anderer Tiere positiv oder negativ beeinflusst wird, ohne den Einfluss von Umweltvariablen zu beachten (Legendre, 1993). Speziell in Habitatmodellen muss davon ausgegangen werden, dass die modellierten Umwelteffekte zusätzlich über den Raum variieren. Dies wird in der Komponente der Nicht-Stationarität modelliert. Allerdings muss bisher mindestens einer, wenn nicht alle dieser eben erwähnten Effekte ignoriert werden, um überhaupt ein Modell schätzen zu können. Dabei sind die Konsequenzen für die Modellinferenz wie nicht unabhängig und identisch verteilte Residuen und damit verzerrte Schätzer und erhöhte Fehlerraten erster Art durchaus bekannt (Dormann *et al.*, 2007).

In Hothorn *et al.* (2010b) wird nun ein neuer Ansatz vorgestellt, der das eben

erwähnte Problem zu lösen versucht, indem die Einflüsse aller Variablen in eine globale und in eine lokale Komponente zerlegt werden. Dabei besteht die globale Komponente aus den Umweltvariablen wie Temperatur, Niederschlag, Bodennutzung. Sie bietet verschiedene Möglichkeiten, komplexere Strukturen, wie zum Beispiel Interaktionen oder nicht-lineare und nicht-additive Effekte zu modellieren. Die lokale Komponente umfasst die räumliche Autokorrelation und die Nicht-Stationarität der Umweltvariablen. Die effektive Variablenselektion durch Anwendung eines Boosting-Algorithmus führt zu einem sehr sparsamen Modell, das zusätzlich durch eine Stabilitätsselektion nur tatsächlich informative Variablen aufnimmt. Das Ziel dieser Arbeit ist, mit der Schätzmethode „Spatial Boosting“, ein Habitatmodell für das Brutverhalten von Grauspecht und Wespenbussard in Bayern zu erstellen.

2 Datenbeschreibung

2.1 Herkunft der Daten und Bearbeitung

Der Datensatz wurde vom Bayerischen Landesamt für Umwelt, Staatliche Vogelschutzwarte Garmisch-Partenkirchen zur Verfügung gestellt. Die Zielvariablen "Vorkommen von Grauspecht“ und "Vorkommen von Wespenbussard“ in Bayern sollen das Brutverhalten dieser Vögel darstellen. Dafür wurde die gesamte Fläche Bayerns aufgeteilt und als durchschnittlich 33.9 km^2 große Quadranten erfasst, in denen man das Brutverhalten bzw. das Vorkommen dieser Vögel mit Hilfe bestimmter beobachtbarer Kovariablen untersucht hat.

Die Kovariablen setzen sich aus den Klima- und Bodennutzungsfaktoren zusammen. Die Klimavariablen stammen aus dem Projekt WorldClim, das sich zum

Ziel gesetzt hat, die wichtigsten Klimadaten für alle Regionen der Erde zu erfassen. Dazu wurden Auswertungen von Wetterstationen aus vielen verschiedenen Klimadatenbanken weltweit zusammengefasst. Eine ausführliche Beschreibung darüber findet man in Hijmans *et al.* (2005). In einer Auflösung von 0.93 km $\times$ 0.93 km = 0.86 km^2, umgangssprachlich auch 1 km^2-Auflösung genannt, stehen interpolierte Monatsdurchschnittsdaten zu den Niederschlagsmengen sowie Minimal-, Maximal- und Durchschnittstemperaturen pro Monat zur Verfügung. Daraus abgeleitet wurden 16 bioklimatische Variablen, die biologisch bedeutender sind, da man sie besser interpretieren kann. Sie beschreiben beispielsweise Jahrestrends, Saisonalität und Extremwerte sowie eventuelle limitierende Umweltfaktoren. Im Weiteren werden nur diese Bioclim-Variablen betrachtet. Sie sind in Tabelle 1 aufgeführt.

Variable	Name	Messniveau
Jahresdurchschnittstemperatur	bio1	metrisch
Tagestemperaturspanne	bio2	metrisch
Isothermalität	bio3	metrisch
Temperatur-Saisonalität	bio4	metrisch
Maximaltemperatur des wärmsten Monats	bio5	metrisch
Minimaltemperatur des kältesten Monats	bio6	metrisch
Jahrestemperaturspanne	bio7	metrisch
Durchschnittstemperatur des feuchtesten Quartals	bio8	metrisch
Durchschnittstemperatur des trockensten Quartals	bio9	metrisch
Durchschnittstemperatur des wärmsten Quartals	bio10	metrisch
Durchschnittstemperatur des kältesten Quartals	bio11	metrisch
Jahresniederschlag	bio12	metrisch
Niederschlag im feuchtesten Monat	bio13	metrisch
Niederschlag im trockensten Monat	bio14	metrisch
Niederschlags-Saisonalität	bio15	metrisch

Tabelle 1: Bioklimatische Variablen von WorldClim.

Die Daten beruhen hauptsächlich auf Messungen der Jahre 1960 bis 1990. Nur wenn in diesem Zeitraum zu wenige Messungen vorlagen, wurde die Zeitspanne

auf die Jahre 1950 bis 2000 ausgedehnt.

Der zweite Teil der Einflussvariablen stammt aus dem CORINE LandCoverProjekt CLC2000, das die europäische Umweltagentur EEA in Zusammenarbeit mit dem European Topic Centre for Terrestrial Environment (ETC-TE) ins Leben gerufen hat. Durch das Projekt sollten einheitliche und vergleichbare Daten über die Oberflächenstruktur und die Art der Bodenbedeckung in Europa gesammelt werden (Deutsches Zentrum für Luft-und Raumfahrt e.V., 2005). Aus Satellitenbildern im Maßstab 1:100.000 wurden zum ersten Mal im Jahr 1990 die 44 verschiedenen Landnutzungsklassen in einer Auflösung von 100 m $\times$ 100 m eingeteilt, wobei in Deutschland nur 37 Klassen relevant sind. Der vorliegende Datensatz enthält Beobachtungen aus dem Jahr 2000, mit 21 verschiedenen Klassen sowie zwei Zusammenfassungen für die Kategorien Wald und Wasser. Drei Variablen (Deponien, Gletscher, Verkehr) wurden von Beginn an ausgeschlossen, da sie nur sehr geringe Ausprägungen aufwiesen. So ergeben sich insgesamt 20 Bodennutzungsvariablen, die in Tabelle 2 aufgeführt sind. Diese Variablen beschreiben den jeweiligen Anteil der Bodennutzung in dem hektargroßen Feld. Wenn es nicht genügend Ausprägungen pro metrischer Variable gab, wurde sie eingeteilt in die Kategorien: 1 = vorhanden, 0 = nicht vorhanden. Die Variable Stadt wurde in drei Kategorien eingeteilt.

Zusätzlich liegen für alle Quadranten die Koordinaten im Gauß-Krüger-System und die Höhe über Normalnull als Variable vor. Aus der Höhe wurde die standardisierte Höhe mit der Formel $\frac{\text{Höhe}-\min(\text{Höhe})}{\max(\text{Höhe})}$ berechnet. Die Höhe geht als Kovariable bei den Umweltvariablen in das Modell ein, wohingegen die standardisierte Höhe in die Berechnung der Nicht-Stationarität einbezogen wird. Da die Kovariablen aus beiden Quellen in verschiedenen Auflösungen vorlagen, wurden

Variable	Messniveau
Wald (SAWald)	metrisch
Wasser (SAWasser)	binär: 1 = vorh., 0 = nicht vorh.
Abbauflächen	binär: 1 = vorh., 0 = nicht vorh.
Acker	metrisch
Deponien	nicht verwendet
Felsen	binär: 1 = vorh., 0 = nicht vorh.
Gletscher	nicht verwendet
Heiden und Moore (HeidenMoore)	binär: 1 = vorh., 0 = nicht vorh.
Industrie	binär: 1 = vorh., 0 = nicht vorh.
Komplex	metrisch
Laubwald	metrisch
Mischwald	metrisch
Moore	binär: 1 = vorh., 0 = nicht vorh.
Nadelwald	metrisch
Obst	binär: 1 = vorh., 0 = nicht vorh.
Stadt	kategorial: 0 = nicht vorh., 1 = Anteil < 0.1, 2 = Anteil > 0.1
Sumpf	binär: 1 = vorh., 0 = nicht vorh.
Verkehr	nicht verwendet
Waldrandgebiet (WaldrandGeb)	binär: 1 = vorh., 0 = nicht vorh.
Fließende Gewässer (WasserFl)	binär: 1 = vorh., 0 = nicht vorh.
Stehende Gewässer (WasserSteh)	binär: 1 = vorh., 0 = nicht vorh.
Weinbau	binär: 1 = vorh., 0 = nicht vorh.
Wiesen	metrisch

Tabelle 2: Bodennutzungsvariablen von CORINE.

jeweils Durchschnittswerte für die ca. 40 km^2 großen Quadranten gebildet.

2.2 Deskriptive Analyse

In Abbildung 1 sind die Verteilungen einiger bioklimatischer Variablen dargestellt: „Jahresdurchschnittstemperatur“ (bio1) in Grad Celcius (multipliziert mit 10), „Isothermalität“ (bio3) in Prozent und „Jahresniederschlag“ (bio12) in Millimeter. Die übrigen bioklimatischen Variablen befinden sich in Anhang A.1.

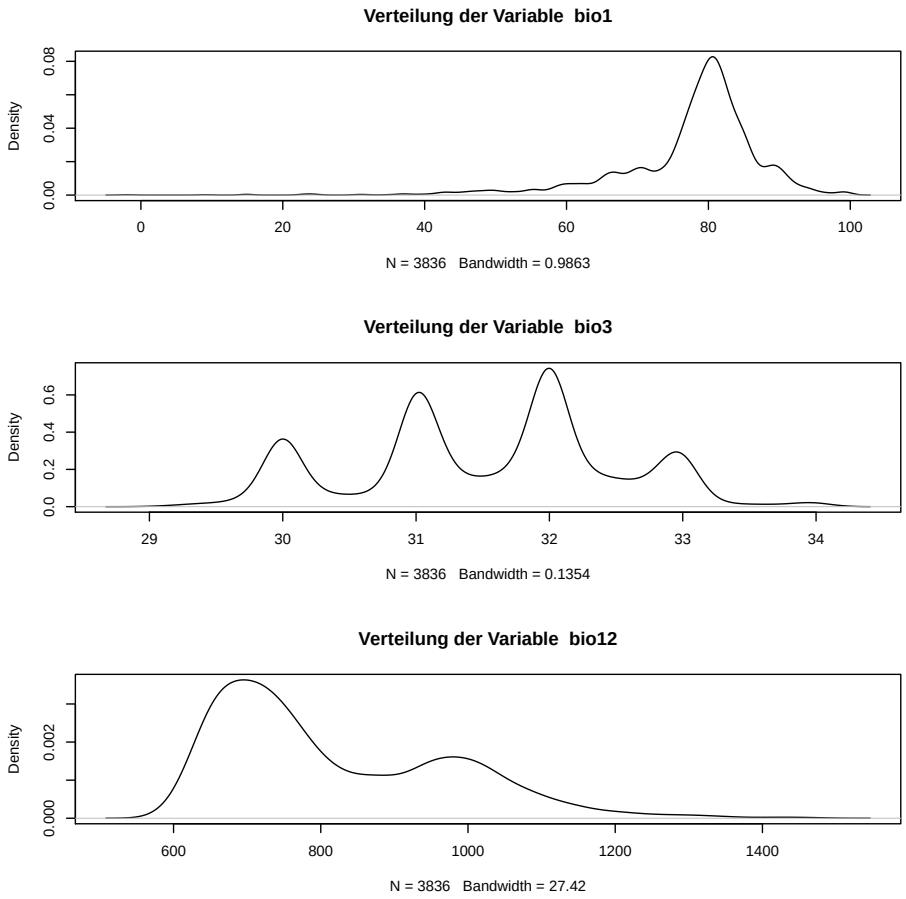


Abbildung 1: Verteilung ausgewählter bioklimatischer Variablen: „Jahresdurchschnittstemperatur“ (bio1), „Isothermalität“ (bio3), „Jahresniederschlag“ (bio12).

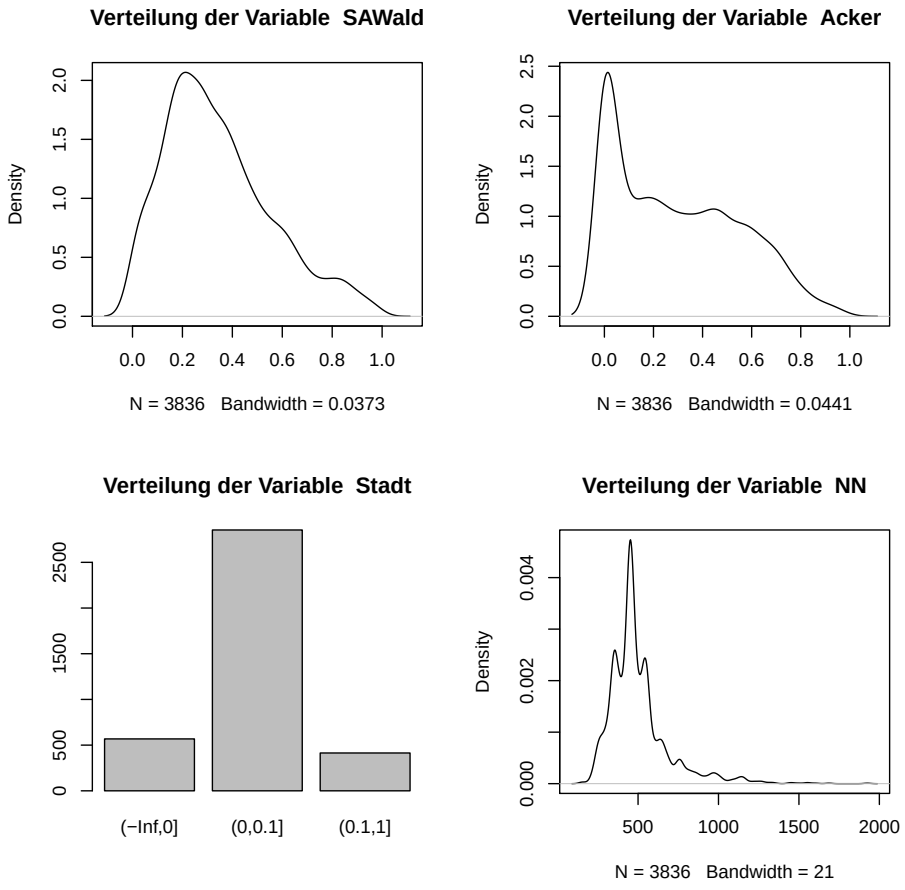


Abbildung 2: Verteilung ausgewählter Bodennutzungsvariablen: „Waldanteil“ (SAWald), „Ackeranteil“ (Acker), „Stadtanteil“ (Stadt), „Höhe“ (NN).

In Abbildung 2 sind beispielhaft die Verteilungen einiger Bodennutzungsvariablen abgebildet: „Waldanteil“ (SAWald), „Ackeranteil“ (Acker), „Stadtanteil“ (Stadt) und „Höhe“ (NN). Die übrigen Variablen sind in Anhang A.1 dargestellt. Auffallend ist, dass die meisten Bodenvariablen eine sehr linkssteile Verteilung haben, es gibt also wenig Beobachtungen, die einen hohen Anteil an der jeweiligen Bodennutzung aufweisen. Das bedeutet auch, dass die Quadranten sehr heterogen sind und es wenige Grids gibt, die von einer Bodennutzung dominiert werden.

Abbildung 3: Vorkommen des Grauspechts in Bayern.

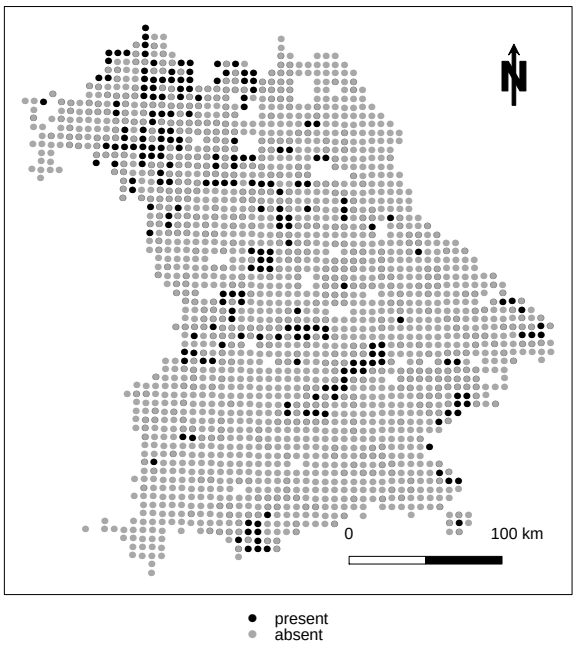


Abbildung 4: Vorkommen des Wespenbussards in Bayern.

In den Abbildungen 3 und 4 erhält man einen groben Überblick darüber, wo Grauspecht und Wespenbussard in Bayern überall leben. Offensichtlich ist das Vorkommen weder bei Grauspecht noch bei Wespenbussard homogen in Bayern verteilt. Grauspechte haben ein deutlich größeres Brutgebiet als Wespenbussarde. Beide Arten kommen vermehrt im Raum Garmisch-Partenkirchen und Unterfranken vor, der Grauspecht zusätzlich im westlichen Oberbayern, sowie im Grenzbereich zwischen Mittelfranken und Schwaben.

2.3 Besonderheiten der Daten

Im Folgenden soll ein Habitatmodell für Grauspecht und Wespenbussard zur Untersuchung des Brutverhaltens bzw. Vorkommens von Brutvögeln erstellt werden. Bei der Modellanpassung an die vorliegenden Daten gibt es einige Punkte, die beachtet werden sollten. Zum einen beinhaltet der Datensatz eine große Menge an Kovariablen. Ein Hauptziel der Anpassung ist es also herauszufinden, welche Kovariablen von Bedeutung sind, und auf diese Weise die Modellkomplexität so weit wie möglich zu reduzieren. Zum anderen sollte man darauf achten, dass es aufgrund des Raumes Abhängigkeiten zwischen den einzelnen Beobachtungen geben kann. In diesem Fall würde das Vorkommen des Grauspechts oder des Wespenbussards in einem Quadranten, die Wahrscheinlichkeit dafür, dass sich auch im Nachbarquadranten Grauspechte und Wespenbussarde befinden, erhöhen, obwohl dies durch die beobachteten Umweltvariablen nicht vorhergesagt würde. Dieses Phänomen bezeichnet man als räumliche Autokorrelation (Legendre, 1993). Für die Modellierung wurde die so genannte Methode „Spatial Boosting“ ausgewählt, die im Folgenden erläutert werden soll. Die genauen Details sind nachzulesen bei Hothorn *et al.* (2010b).

3 Methoden

3.1 Generalisierte Additive Modelle

Die generalisierten additiven Modelle (GAM) stellen eine Art Erweiterung zu den generalisierten linearen Modellen (GLM) dar, zu dem das Logitmodell gehört. Ziel ist es durch das generalisierte additive Modell den Einfluss der Kovariablen auf den Response flexibel durch eine Funktion zu modellieren. Um die Schätzung dieser Funktion zu vereinfachen, unterstellt man für den Einfluss zusätzlich eine additive Struktur, das heißt man geht davon aus, dass sich der Prädiktor η_i additiv aus glatten eindimensionalen Funktionen der einzelnen Kovariablen zusammensetzt:

$$\eta_i = f(\mathbf{x}_i, s_i) = \sum_j f_{(j)}(x_{ij}, s_i).$$

Bei den vorliegenden Daten handelt es sich bei jeder einzelnen Vogelart um eine Binärvariable $y_i \in \{0, 1\}$ als Response, wobei y_i = 1, falls die Vogelart im Quadrant i vorkommt, y_i = 0 falls nicht.

Für die binäre abhängige Größe wird eine Bernoulli-Verteilung $B(1, \pi_i)$ angenommen, deren Wahrscheinlichkeit $\pi_i = \mathbb{P}(y_i|\mathbf{x}_i, s_i)$ über eine logit-Funktion $\pi_i(f) = logit^{-1}(f(\mathbf{x}_i, s_i))$ modelliert wird.

Die Wahrscheinlichkeit, dass eine Vogelart an einem Punkt s, abhängig von den Umweltvariablen $\mathbf{x} = (x_1, \ldots, x_p)$ vorkommt, entspricht also der inversen Logit-Funktion an der Stelle der Regressionsgleichung.

3.2 Die Methode des Spatial Boosting

Bei hochdimensionalen Datensätzen sind übliche Schätzverfahren, wie zum Beispiel penalisierte Schätzung nicht mehr anwendbar. Es kommt zu numerischen

Rechenproblemen. Boosting ist ein möglicher Algorithmus zur Schätzung hochdimensionaler Regressionsmodelle für additive Prädiktoren. Das iterative Anpassen einzelner schwacher Schätzer führt zu einem insgesamt numerisch guten Schätzergebnis und überzeugt durch seine effektive Variablenselektion. Beim Spatial Boosting werden die Kovariablen in eine globale und eine lokale Komponente aufgeteilt. Die globale Komponente beachtet hierbei ausschließlich die Umweltvariablen sowie mögliche lineare oder nicht-lineare Effekte und Interaktionsterme. Ein rein globales Modell würde annehmen, dass die Effekte der Umweltvariablen fest und universal sind. Bei Auftreten von Nicht-Stationarität variieren diese Effekte jedoch mit dem Raum. Die lokale Komponente beschreibt daher die räumliche Autokorrelation als Funktion $f_s(s)$ nur abhängig vom Raum. Die Nicht-Stationarität wird als Funktion $f_{ns}(\mathbf{x}, s)$ in Abhängigkeit vom Raum und den Umweltvariablen modelliert. Durch die lokale Komponente erhält man eine Schätzung der unbeobachteten Heterogenität, die durch räumliche Autokorrelation oder nicht-stationäre Effekte verursacht wird. Dies ist deshalb von Bedeutung, da man davon ausgehen muss, nicht alle tatsächlichen Einflussvariablen erfasst zu haben. Die Annahme der Unabhängigkeit von $Y_i|\mathbf{x}_i$ kann aber nur getroffen werden, wenn alle Kovariablen gegeben sind. Deswegen werden die restlichen nicht erfassten Kovariablen sozusagen zu einem räumlichen Effekt der unbeobachteten Heterogenität zusammengefasst. Dies ist bei den meisten der bisher verwendeten Verfahren nicht der Fall.

Durch die Zerlegung hat die Regressionsfunktion, die in die Modellgleichung einfließt, folgende Form:

$$f(\mathbf{x}, s) = \underbrace{f_{env}(\mathbf{x})}_{global} + \underbrace{f_{ns}(\mathbf{x}, s) + f_s(s)}_{lokal} \tag{1}$$

Mit dieser Modellzerlegung wird auch die Variabilität in drei Komponenten zerlegt: die Variabilität erklärt durch die Umweltvariablen ($f_{env}(\mathbf{x})$), Variabilität, die von räumlicher Autokorrelation verursacht wird ($f_s(s)$) und die Variabiliät verursacht durch nicht-stationäre Umwelteffekte, das heißt zusätzlich räumlich variierende Effekte der Umweltvariablen ($f_{ns}(\mathbf{x}, s)$).

3.2.1 Beschreibung der Modellkomponenten

Da das Modell vom Raum abhängig ist, ist es nur auf das betreffende Untersuchungsgebiet anwendbar. f_{env} kann hingegen für Prognosen außerhalb Bayerns genutzt werden, da in diesem Term die räumlichen Effekte herausgerechnet werden und somit die Prädiktionen nicht verzerrt werden. Der Term kann auf zwei Arten modelliert werden: Die einfachste Möglichkeit ist ein parametrischer Ansatz mit dem linearen Prädiktor $f_{env}(\mathbf{x}) = \mathbf{x}^T\beta$, wobei β der zu schätzende Vektor der Regressionskoeffizienten ist. Eine bisher genutzte Möglichkeit, hier die Autokorrelation miteinzubeziehen, ist zum Beispiel die Spezifizierung einer Arbeitskovarianz in Generalized Estimating Equations (GEE) (Dormann *et al.*, 2007). Eine andere Möglichkeit der Modellierung ist ein nonparametrischer Ansatz mit additiven glatten Funktionen, also $f_{env}(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j)$, wobei $\mathbf{x} = (x_1, \ldots, x_p)$. In jeder einzelnen Kovariable kann so ein möglicher nicht-linearer Effekt auf flexible Weise geschätzt werden. Komplexere Modelle erlauben zusätzlich Interaktionen, wie zum Beispiel Random Forests oder Boosted Regression Trees. $f_s(s)$ stellt eine glatte zweidimensionale Oberflächenfunktion dar, die die unbeobachtete Heterogenität, eingeführt durch lokale Einflüsse, modelliert. So werden räumliche Autokorrelationsmuster erkannt. $f_{ns}(\mathbf{x}, s)$ repräsentiert die räumliche Nicht-Stationarität.

3.2.2 Modellanpassung durch Spatial Boosting

Die Modellanpassung wird durch die Minimierung der negativen Log-Likelihood der zugrunde liegenden Verteilung durchgeführt. Das Vorkommen der Grauspechte bzw. der Wespenbussarde folgt einer $B(1, \pi_i)$-Verteilung mit $\pi_i = \mathbb{P}(y_i|\mathbf{x}_i, s_i)$ und $\pi_i(f) = logit^{-1}(f(\mathbf{x}_i, s_i))$ Damit ist die negative Log-Likelihood-Funktion

$$\hat{f} = \underset{f}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(y_i, \pi_i(f))$$

mit

$$\rho(y_i, \pi_i(f)) = -y_i \log(\pi_i(f)) - (1 - y_i) \log(1 - \pi_i(f))$$

als Beitrag einer Beobachtung zur Gesamt-Log-Likelihood.

Die Funktion $\hat{f}$, die die Verlustfunktion minimiert, wird mit einem Componentwise Functional Gradient Descent Boosting-Algorithmus geschätzt. Für Modelle der Form (1) können auch Methoden wie MCMC-Algorithmen (Fahrmeir *et al.*, 2004), (Kneib *et al.*, 2008) oder penalisierte Schätzung von generalisierten additiven Modellen verwendet werden. Diese Methoden sind jedoch rechenaufwändig und auf Daten mit einer geringen Zahl an Einflussvariablen oder einer kleinen bis mittleren Beobachtungszahl ausgelegt und es gibt keine effizienten Verfahren der Variablenselektion. Auf diese Weise würden unbedeutende Parameter das finale Modell unnötig komplex machen. Die Modellinferenz hat hier aber vor allem die Selektion von informativen Parametern zum Ziel. Falls keine räumliche Autokorrelation vorliegt, sollte auch die Modellkomponente $f_s(s)$ nicht in das Modell aufgenommen werden, das heißt $f_s(s) \equiv 0$ und genauso $f_{env}(\mathbf{x}) \equiv 0$, falls kei-

ne der Umweltvariablen einen Einfluss hat. Hier ist man allerdings mehr an den Effekten der einzelnen Umweltvariablen, also an dem Ergebnis $f_j(x_j) \equiv 0$ interessiert, was bedeutet, dass die Variable x_j keinen Einfluss auf das Vorkommen von Grauspecht und Wespenbussard hat. Der Idealfall wäre ein globales Modell, in das nur wenige Umweltkomponenten aufgenommen werden.

Componentwise Functional Gradient Descent Boosting-Algorithmus

Für den Componentwise Functional Gradient Descent Boosting-Algorithmus wird $\hat{f} \equiv 0$ als konstantes Modell initialisiert. Im ersten Schritt werden die Residuen für das aktuelle Modell berechnet. Unter dem Residuum versteht man hier den negativen Gradienten u_i der Verlustfunktion ρ, berechnet für jede Beobachtung y_i:

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

Nun wird diejenige Basisprozedur g_{j^*} ($f_j(x_j)$, f_{ns} oder f_s) ausgewählt, welche die Residuen am besten beschreibt, das heißt die Summe der quadrierten Differenz zwischen Residuen und Modellkomponente minimiert:

$$j^* = \underset{1 \leq j \leq p}{\operatorname{argmin}} \sum_{i=1}^{n} (u_i - \hat{g}_j(x_i))^2.$$

Nur diese Komponente wird aktualisiert mit zum Beispiel 10% der Prädiktionen (Schrittweite ν) und zum aktuellen Modellfit hinzugefügt:

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

Für alle anderen Komponenten gilt:

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

Anschließend werden die Residuen wieder neu berechnet und die entsprechende Modellkomponente aktualisiert. Diese Schritte werden wiederholt, bis eine vorher festgelegte Anzahl von Iterationen durchgeführt wurde. Das finale Modell $\hat{f}$ setzt sich zusammen aus der Summe aller gefitteten Modelle der einzelnen Komponenten $\hat{f}_{env}$, $\hat{f}_{ns}$ und $\hat{f}_s$. Die mathematischen Details werden von Bühlmann und Hothorn (2007) und Kneib *et al.* (2007) beschrieben.

Basisprozedur

Die sogenannte Basisprozedur, die auch als Baselearner bezeichnet wird, bestimmt, wie die Residuen gefittet werden. Die Wahl der Baselearner ist entscheidend, da sie festlegen, in welcher Form die einzelnen Modellkomponenten in das finale Modell eingehen. Für f_{env} kommen lineare Modelle, Smoothing-Splines, univariate P-Splines oder Regressionsbäume in Frage. Wobei letztere Methode genau mit den Boosted Regression Trees übereinstimmt. Für f_s werden die Baselearner als bivariater Tensorprodukt P-Spline gewählt, was einer glatten zweidimensionalen Oberflächenfunktion entspricht. Für die nicht-stationäre Komponente f_{ns} bietet sich ein Produkt eines Tensorprodukt P-Splines mit einer Umweltvariable x_j an. Interaktionen können zum Beispiel über lineare Terme von Produkten berücksichtigt werden oder, wenn man noch flexibler sein möchte, über zwei- oder dreidimensionale glatte Funktionen.

Wie bereits erwähnt, wurden metrische Umweltvariablen mit nur wenigen Ausprägungen kategorisiert, so dass nun zwei unterschiedliche Variablentypen vorliegen. Für die stetigen Variablen wurden als Baselearner penalisierte Regressionssplines (mit sechs Freiheitsgraden) verwendet und für die faktorisierten Variablen einfache lineare Modelle, die über Ridge-Regression (Parameter π bestimmt durch sechs Freiheitsgrade) geschätzt wurden.

3.2.3 Modellwahl und Variablenselektion

Es gibt sechs verschiedene Grundmodelle, die alle möglichen Einflussszenarien beschreiben, indem sie verschiedene Restriktionen an die einzelnen Modellkomponenten stellen (Tabelle 3).

Modell	$f_{env}(\mathbf{x})$	$f_{ns}(\mathbf{x}, s)$	$f_s(s)$
Spatial	$\equiv 0$	$\equiv 0$	
Additive	$\sum_{j=1}^{p} f_j(x_j)$	$\equiv 0$	$\equiv 0$
Add/Spatial	$\sum_{j=1}^{p} f_j(x_j)$	$\equiv 0$	
Tree/Spatial		$\equiv 0$	
Add/Vary	$\sum_{j=1}^{p} f_j(x_j)$		
Tree/Vary			

Tabelle 3: Modellrestriktionen.

Das Modell Spatial, das nur den lokalen Einfluss misst und alle anderen Komponenten auf Null setzt, wäre das beste Modell, wenn keine der erhobenen Umweltvariablen Einfluss auf den Response hat. Wenn dagegen nur diese Umweltvariablen Einfluss haben ohne räumliche Variation und dabei die einzelnen Variablen additiv und ohne Interaktionen auf den Response wirken, wäre das Modell Additive das richtige. Add/Spatial modelliert einen additiven Effekt der Umweltvariablen sowie einen zusätzlichen räumlichen Effekt ohne Nicht-Stationarität oder Interaktionen zu berücksichtigen. Mit Regressionsbäumen als Baselearner für f_{env} können Interaktionen besser modelliert werden, ansonsten ist das Modell Tree/Spatial gleich dem Vorherigem. Am komplexesten sind die letzten beiden Modelle, die damit auch die größte Flexibilität bieten: Add/Vary modelliert wieder additive Effekte für f_{env} und erlaubt gleichzeitig räumliche Autokorrelation und Nicht-Stationarität. Dies ist auch bei Tree/Vary der Fall. Dort sind zusätzlich Interaktionen bei den Umweltvariablen erlaubt, was insgesamt heißt, dass über-

haupt keine Restriktionen an die Modellkomponenten gestellt werden. Aus diesen sechs Grundmodellen wird für die vorliegenden Daten das beste Modell ausgewählt (vgl. Kapitel 4.1).

Die eigentliche Modellwahl wird in zwei Schritten durchgeführt. Für jedes der sechs oben genannten Modelle wird die ideale Iterationszahl bestimmt. Diese ergibt sich als m_{stop} mit dem minimalen empirischen Risiko, berechnet mit Bootstrap- und Kreuzvalidierungsverfahren. Eine andere Möglichkeit wäre, m_{stop} durch das Informationskriterium nach Akaike (AIC), das korrigierte AIC oder das Bayesianische Informationskriterium (BIC) zu bestimmen. Da es sich aber um einen hochdimensionalen Datensatz handelt, ist die Berechnung über Bootstrap und Kreuzvalidierung am geeignetsten. Die Wahl des idealen Stoppkriteriums hat den Zweck, Overfitting zu vermeiden. Im zweiten Schritt wird mit der neu bestimmten optimalen Anzahl an Boosting-Schritten die Modellanpassung wiederholt. Die sechs Modelle werden anhand der negativen Log-Likelihood verglichen. Die beste Modellanpassung hat dasjenige Modell, das in wiederholten Bootstrapstichproben die kleinste negative Log-Likelihood hat (vgl. Abbildungen 5 - 8).

Zudem muss auch die Schrittweite ν festgelegt werden. Für bisherige Probleme schien die Wahl dieser Schrittweite von eher geringer Bedeutung zu sein, solange sie klein genug gewählt wird, um den Effekt des aktuellen Fits zu dämpfen. Eine kleinere Schrittgröße bedeutet typischerweise eine größere Anzahl an Iterationsschritten und somit mehr Berechnungszeit, wobei sich die Prädiktionsgenauigkeit im Allgemeinen nicht verschlechtert. Aus diesem Grund genügt es meist den Parameter ν „ausreichend klein“ zu wählen (Bühlmann und Hothorn, 2007). Daher wurde bisher die Schrittweite oft auf den Wert $\nu = 0.1$ festgelegt. In der Auswertung dieser Arbeit stellte sich heraus, dass ein weiteres Verringern der

Schrittgröße die Ergebnisse für die vorliegenden Daten nicht nennenswert verbessern kann (vgl. Kapitel 4.1).

Da immer nur eine Modellkomponente pro Iterationsschritt angepasst wird, führt eine kleine Anzahl an Iterationen zu einem sparsamen Modell. Somit ist diese Methode eine sehr gute Möglichkeit der Variablenselektion. Zusätzlich wird für das beste Modell eine Stability Selection angewandt, um sicher zu stellen, dass tatsächlich nur einflussreiche Variablen und Komponenten aufgenommen werden und man keine Effekte interpretiert, die in Wirklichkeit gar nicht bestehen. Dazu wird die empirische Wahrscheinlichkeit berechnet, wie oft die Variable in Teildaten ausgewählt wird (Meinshausen und Bühlmann, 2010). Variablen, deren Wahrscheinlichkeit größer einem festgelegten Grenzwert sind, gelten als einflussreich, wobei das Signifikanzniveau α eingehalten wird. Auf diese Weise erhält man ein Modell, das so komplex wie nötig, aber so einfach wie möglich ist.

4 Ergebnisse

4.1 Statistische Analyse

Für den vorliegenden Datensatz wurde das Boosting-Verfahren für alle sechs vorher spezifizierten Modelle mit zwei verschiedenen Schrittgrößen $\nu = 0.1$ und 0.05 durchgeführt. Wie bereits in Abschnitt 3.2.2 erwähnt wurde, spielt die Wahl der Schrittgröße ν eine untergeordnete Rolle. Da bereits die Hyperparameter für die Glättung jedes Baselearners und die optimale Anzahl an Iterationen über Kreuzvalidierung oder ähnliches bestimmt werden müssen, wird der Parameter ν der Einfachheit halber vorgegeben, um eine weitere Verkomplizierung des Algorithmus zu vermeiden. Die nachfolgenden Boxplots (Abbildungen 5 bis 8) zeigen

daher für alle Modelle mit dem optimalen m_{stop} die Out-of-Bootstrap negative Log-Likelihood für mehrere Bootstrapstichproben und für die zwei verschiedenen Schrittgrößen ν; zuerst für den Grauspecht, dann für den Wespenbussard.

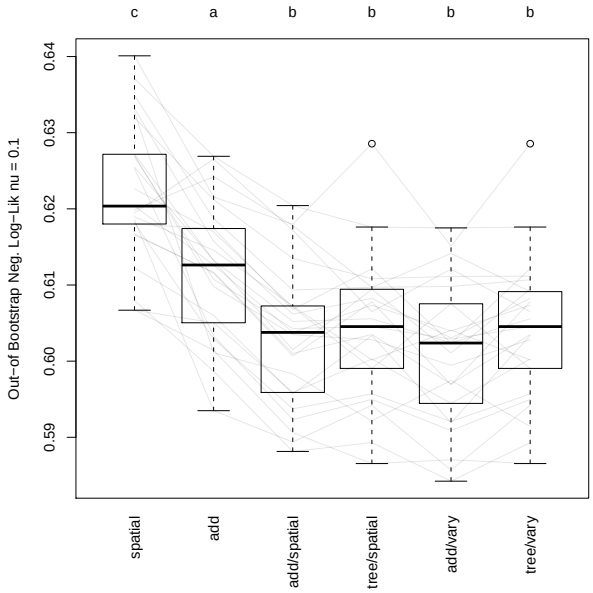


Abbildung 5: Out-of-Bootstrap Neg. Log-Likelihoods $\nu = 0.1$ Grauspecht.

Wie bereits angenommen bewirkt die Schrittgrößenverkleinerung hier wenn überhaupt nur eine minimal kleine Veränderung bzw. Verbesserung bei der Modelgüte, deshalb wird im Folgenden nur mit der Schrittgröße ν = 0.10 verfahren.
Als Vergleichsmethode für die Fragestellung, welche der Modelle sich signifikant im Mittelwert der negativen Log-Likelihood unterscheiden, wurde ein multipler Vergleich nach Tukey gemacht. Die Buchstaben über den Boxplots geben an, welche Modelle die gleiche Modellgüte haben und welche sich unterscheiden. Modelle mit gleichem Buchstaben haben hier die gleiche Modellgüte.

Bei beiden Werten von ν hat die Zielvariable Wespenbussard immer das Modell

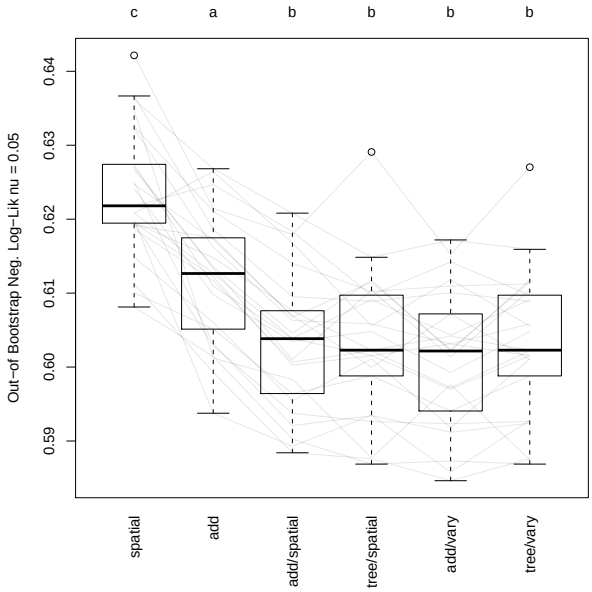


Abbildung 6: Out-of-Bootstrap Neg. Log-Likelihoods $\nu = 0.05$ Grauspecht.

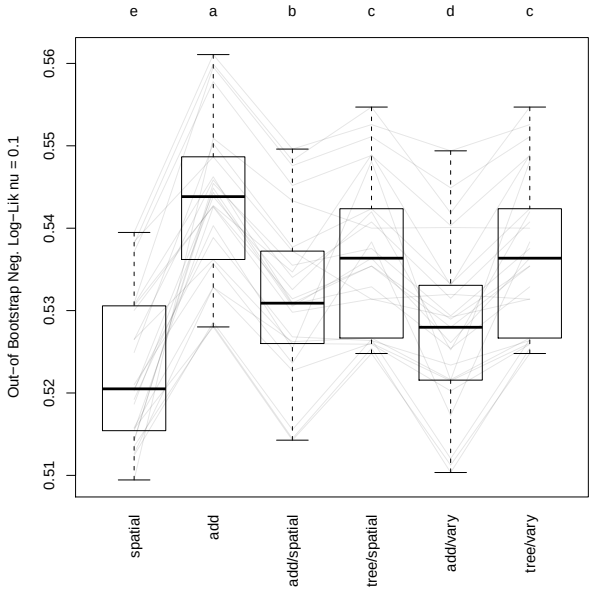


Abbildung 7: Out-of-Bootstrap Neg. Log-Likelihoods $\nu = 0.1$ Wespenbussard.

Spatial als beste Modellanpassung. Dies ist ein Hinweis darauf, dass ein großer räumlicher Effekt besteht und f_s eine dominierende Modellkomponente ist.

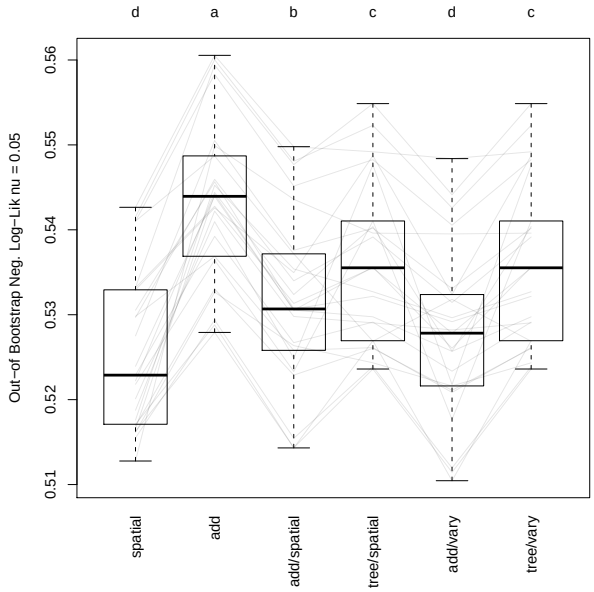


Abbildung 8: Out-of-Bootstrap Neg. Log-Likelihoods $\nu = 0.05$ Wespenbussard.

Für den Grauspecht sind die Modelle Add/Spatial, Tree/Spatial, Add/Vary und Tree/Vary die besten, da sie die kleinste negative Log-Likelihood aufweisen. Aufgrund der gleichen Modellgüte, die die vier nach Tukey haben, entscheidet man sich bei der weiteren Interpretation für das weniger komplexe Modell Add/Spatial, welches zusätzlich zum räumlichen Effekt additive Einflüsse der Umweltvariablen miteinbezieht.

Tabelle 4 zeigt, welche Variablen für die Schrittgröße ν = 0.10 im Add/Spatial-Modell für den Grauspecht ausgewählt wurden.
Man kann davon ausgehen, mit diesem Modell die bestmögliche Anpassung an die Daten gefunden zu haben. Die weitere Interpretation beschränkt sich auf die Modelle Spatial für den Wespenbussard und Add/Spatial für den Grauspecht jeweils mit der Schrittgröße $\nu = 0.10$.

Modell	Schrittgröße ν	Ausgewählte Variablen
Add/Spatial	0.10	Acker, Komplex, Mischwald, Nadelwald, Wiesen, bio4, Höhe, SAWasser, Laubwald, bspatial

Tabelle 4: Selektierte Variablen für Grauspecht.

4.2 Interpretation

Im Modell Spatial für den Wespenbussard wird die gesamte Heterogenität nur anhand der räumlichen Verteilung erklärt. In Abbildung 9 sind die relativen Unterschiede des Vorkommens für den zentrierten räumlichen Effekt dieses Modells gezeichnet. Man sieht, dass besonders im Raum Garmisch-Partenkirchen und im westlichen Unterfranken das Vorkommen der Wespenbussarde wahrscheinlicher ist als im restlichen Bayern, die Wespenbussarde aber auch in Mittelfranken, Zentral- bis Niederbayern auftreten können. Dies deckt sich auf den ersten Blick mit der beobachteten Verteilung in Abbildung 4.

Die Abbildung 10 zeigt die gefitteten Werte, die das erhöhte Vorkommen im Raum Garmisch-Partenkirchen und im westlichen Unterfranken richtig wiedergeben, jedoch für das vereinzelte Vorkommen (im Vergleich zu den Ballungsgebieten bei Garmisch-Partenkirchen und in Unterfranken) in dem Bereich Mittelfranken, Zentral- bis Niederbayern keine Wespenbussarde prognostiziert bzw. schätzt.

Für den Grauspecht wird das Modell Add/Spatial angenommen, in das die Effekte der Umweltvariablen als additive glatte Funktionen aufgenommen werden. Es hat eine vergleichbar gute Modellanpassung. Als einflussreiche Kovariablen ergeben sich durch Stability Selection die zehn Kovariablen Ackergebiet, Komplexe Parzellenstruktur, Mischwald, Nadelwald, Wiesen, Temperatur-Saisonalität (bio4), Höhe, prozentualer Anteil an Wasser, Laubwald und die räumliche Kom-

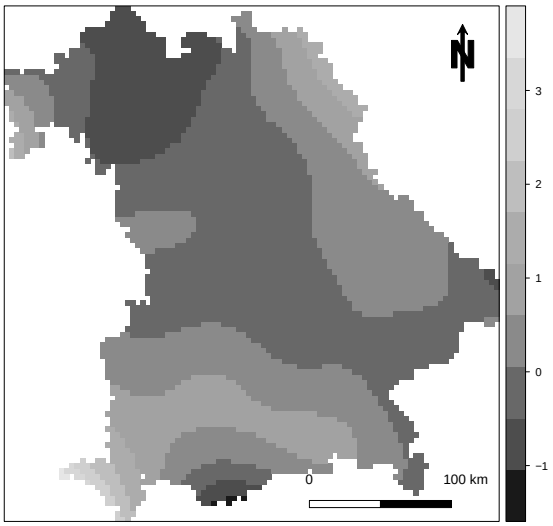


Abbildung 9: Geschätzter räumlicher Effekt im Modell Spatial Wespenbussard.

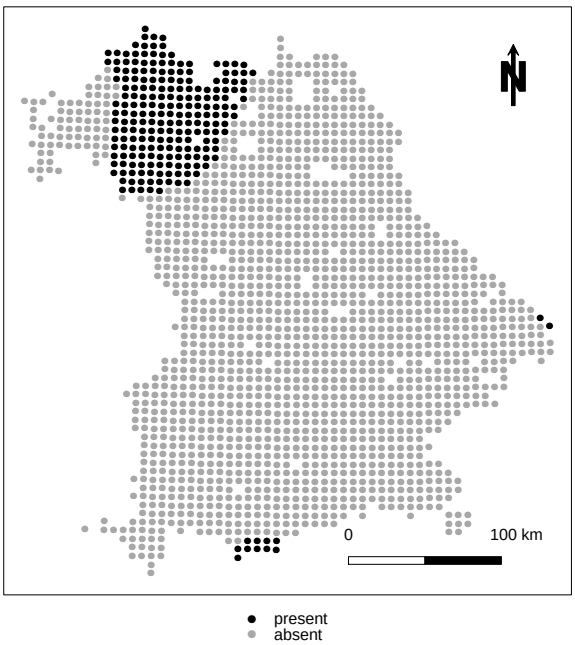


Abbildung 10: Gefittetes Vorkommen der Wespenbussarde im Modell Spatial.

ponente. Zuerst überprüft man, wieviel Variabilität überhaupt durch die einzelnen Modellkomponenten erklärt wird.

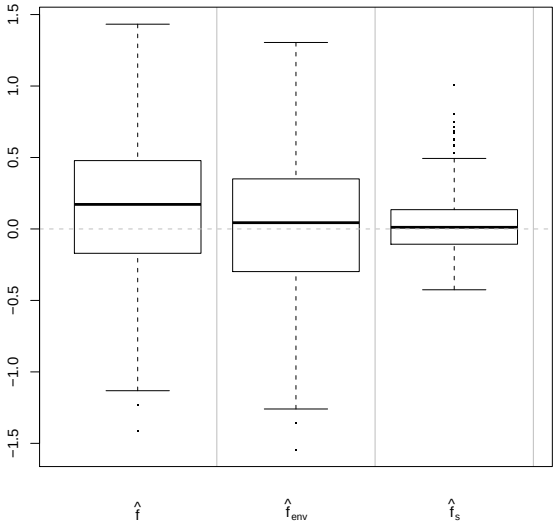


Abbildung 11: Zerlegung der erklärten Variabilität für die einzelnen Modellkomponenten Modell Grauspecht (gefittete Werte auf der Log-Skala).

In Abbildung 11 wird deutlich, dass der Hauptteil der Variabilität durch die Umweltvariablen ($\hat{f}_{env}$) erklärt wird. Die weitere erklärende Größe ist die räumliche Komponente $\hat{f}_s$, deren Einfluss im Vergleich zur Umweltkomponente jedoch viel geringer ist.

Die geschätzten Effekte der einzelnen Umweltvariablen f_{partial} lassen sich so interpretieren, dass sich die Chance auf das Vorkommen des Grauspechts bei Konstanthalten aller anderen Einflussvariablen multiplikativ um den Faktor $\exp(f_{\text{partial}})$ ändert.
In den nachfolgenden Grafiken bedeutet ein geschätzter Effekt größer als Null einen positiven Einfluss und dementsprechend ein geschätzter Effekt kleiner als Null einen negativen Einfluss.

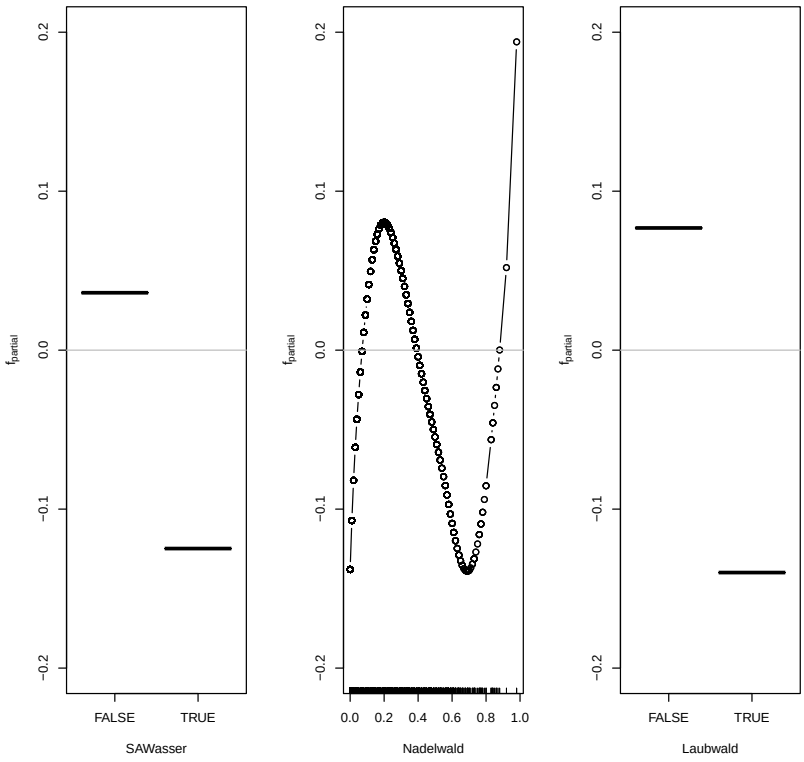


Abbildung 12: Partielle Effekte der Umweltvariablen „Wasseranteil“ (SAWasser), „Nadelwaldanteil“ (Nadelwald), „Laubwald“ (Laubwald) - Modell Grauspecht.

In Abbildung 12 sieht man die geschätzten partiellen Effekte für die Variablen „Wasseranteil“ (SAWasser), „Nadelwaldanteil“ (Nadelwald) und „Laubwald“. Bei der Variable „Wasseranteil“ wurde eine Kategorisierung durchgeführt. Ein geringer Anteil von Wasser im Quadrant wird hier unter „kein Anteil“ geführt. Mit diesem Wissen ist der Effekt, dass „kein Wasser“ die Chance auf das Vorkommen von Grauspechten um ca. 4 % erhöht und das „Vorhanden sein von Wasser“ einen negativen Effekt auf die Grauspechte hat, nicht ganz so missverständlich. Was man daraus interpretieren kann, ist, dass die Grauspechte nicht viele Gewässer in

ihrem Lebensraum benötigen.
Signifikanter Einfluss ist auch bei den Waldvariablen gegeben. Wenn der Anteil des Nadelwalds pro Quadrant zwischen 10 % und 40 % liegt, steigt die Chance auf Grauspechte leicht. Genauso verhält es sich für Werte über 90 % Nadelwaldbedeckung, wobei in diesem Fall die Beobachtungszahl sehr gering ist, sodass die Aussage nicht verallgemeinert werden kann. Für Flächen mit einem Nadelwaldanteil von 40 bis 90 % dagegen ist die Chance leicht verringert, ebenso für den Anteil unter 10 %.
Einen ähnlichen Effekt sieht man bei der kategorisierten Variable „Laubwald“: kommt im Quadrant kein Laubwald vor, steigt die Chance leicht (um ca. 8 %). Ist das Gebiet mit Laubwald bedeckt, sinkt die Chance um 15 %. Dies ist teilweise widersprüchlich zur Literatur, da man das Zuhause der Grauspechte im Laubwald der Mittelgebirge bis an den Alpenrand hin einordnet und Bäume wie Buche und Eiche als Lieblingsbrutbäume gelten (Bauer und Berthold, 1997). Eine mögliche Erklärung liegt, wie bei der Variable „Wasseranteil“, in der Kategorisierung der Variable Laubwald, denn was hier als „kein Laubwald vorhanden“ angenommen wird, kann ja durchaus ein kleiner prozentualer Anteil von Laubwald bedeuten. Was den Nadelwald betrifft, gilt als die Nordgrenze des Verbreitungsgebietes der Übergangsbereich zwischen geschlossenem Nadelwald und aufgelockertem Laubmischwald, was die steigende Chance für das Vorkommen von Grauspechten im Bereich mit Nadelwaldanteil von 10 - 40 % untermauert (Bauer und Berthold, 1997).

Die Effekte von Nadel- und Laubwald spiegeln sich auch im Effekt der Variable „Mischwald“ wider (Abbildung 13). Durch die Aussage, dass kein oder ein sehr geringer Laubwaldanteil und nur ca. 10 bis 40 % Nadelwaldanteil die Chance auf Grauspechte erhöht, erklärt sich der konstant negative Effekt bei wachsendem An-

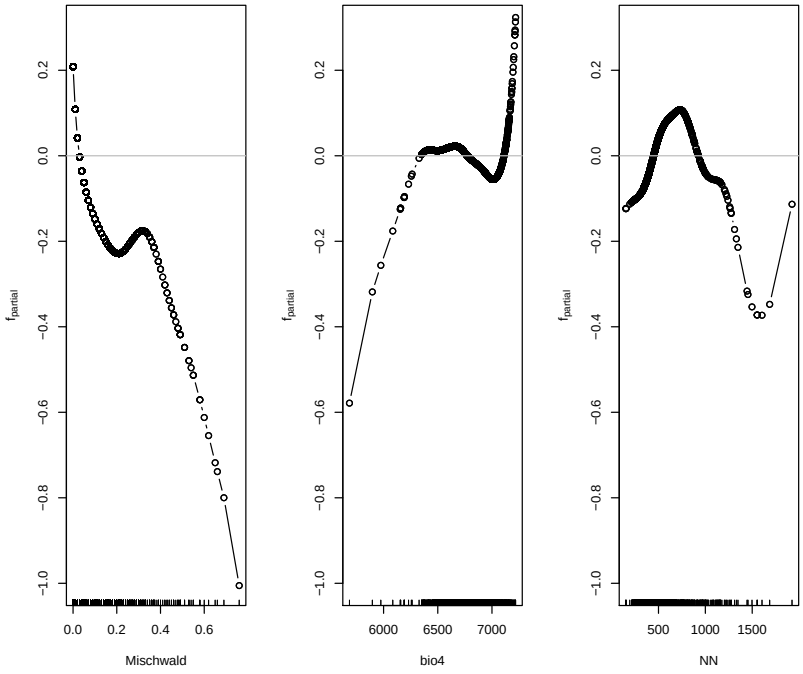


Abbildung 13: Partielle Effekte der Umweltvariablen „Mischwaldanteil“ (Mischwald), „Saisonabhängiger Temperaturunterschied“ (bio4), „Höhe“ (NN) - Modell Grauspecht.

teil von Mischwald pro Quadrant.

Bei der Variable „Saisonabhängiger Temperaturunterschied“ (bio4) handelt es sich um einen Variationskoeffizienten, der das Verhältnis Mittelwert zu Varianz bzw. Standardabweichung angibt. Ein negativer Effekt bei dieser Variable lässt sich unterhalb von 6400 nicht bestätigen, da die Beobachtungszahl hier sehr gering ist. Im Bereich von 6750 bis 7000 tritt jedoch ein wachsender negativer Effekt auf, der die Chance für Grauspechte bei dem Wert 7000 um ca 10 % senkt. Ab 7000 ist ein deutlich rasanter Trend nach oben ersichtlich, der beschreibt, dass die Chance für das Vorkommen steigt, wenn der saisonabhängige Temperaturunterschied um mehr als 7000 abweicht. Das heißt je größer die Streuung der Temperatur ist, de-

sto großer ist die Chance auf Grauspechte.

Die Variable „Höhe“ ist kein direkter physiologischer Faktor, sondern eine Proxy-Variable unter anderem für Klima und Fläche. Wenn man sich die Werte von 500 bis 1000 m ansieht, erkennt man den positiven Effekt auf das Vorkommen des Grauspechts. Der optimale Lebensraum scheint in einer Höhe von 800 m über NN zu liegen, was für eine Chancenerhöhung für den Grauspecht um 10 % sorgt. Außerhalb des Bereichs, das heißt unterhalb von 500 m und überhalb von 1000 m, erweist sich die Höhe über NN als negativer Einfluss. Überhalb der Baumgrenze bei 1200 m verstärkt sich dieser Effekt noch. Ab 1500 m schein sich der negative Effekt zu neutralisieren, allerdings ist die Anzahl der Beobachtungen zu gering, um diese Aussage zu verallgemeinern.

In Abbildung 14 erkennt man einen eindeutig positiven Einfluss der Variable „Acker“ auf das Vorkommen von Grauspechten. Bis zu einer Ackerfläche von ungefähr 20 % ist der Effekt negativ und verringert die Chancen. Je größer jedoch die prozentuale Bedeckung des Quadranten mit Ackerfläche ist, umso größer wird auch die Chancen für Grauspechte. Zum einen kann man dies durch Habitate in Feldrainen erklären, die nicht bewirtschaftet werden, weil sie nur schwer zugänglich sind. Somit bieten sie ideale Lebensbedingungen für Vögel und andere Tiere. Zum anderen liegen Ackerflächen zur Erhaltung der Bodenfruchtbarkeit regelmäßig brach und ermöglichen so den Grauspechten einen ungestörten Lebensraum. Zur Relativierung dieses Trends muss jedoch erwähnt werden, dass Ackerfläche auch ein Indikator für intensive Landwirtschaft sein kann. In diesen Flächen sind normalerweise wenig Vögel vorzufinden, weil sie eine starke Barriere zur Ausbreitung der Populationen darstellen (Bauer und Berthold, 1997).

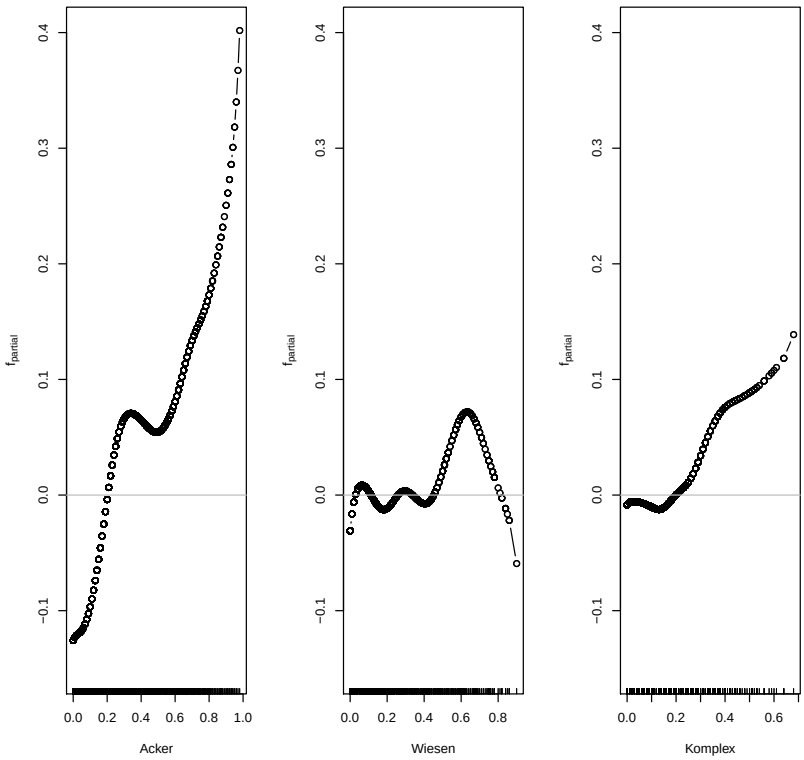


Abbildung 14: Partielle Effekte der Umweltvariablen „Ackeranteil“ (Acker), „Wiesenanteil“ (Wiesen), „Komplexe Parzellenstruktur“ (Komplex) - Modell Grauspecht.

Der Einfluss des Wiesenanteils pro Quadrant ist bis zu einem 50%-igen Anteil weder positiv noch negativ, da die Intervalle so klein sind, dass geringe Ausschläge nach oben und unten vernachlässigt werden können. Bei einem Anteil von 50 bis 80 % liegt der positive Effekt auf Grauspechte bei ca. 8 %. Über 80 % Wiese hat widerum einen negativen Einfluss auf das Vorkommen von Grauspechten.
Dies könnte die Einflussvariable „Komplexe Parzellenstruktur“ (Komplex) erklären, die für ein Nebeneinander kleiner Parzellen unterschiedlicher Prägung steht. Diese verringert bis zu einem Anteil von ungefähr 15 % die Chance auf Grauspechte, doch je größer der prozentuale Abdeckung von solchen komple-

xen Parzellstrukturen ist, desto größer ist die Chance. Das wiederum heißt, je abwechslungsreicher die Landschaft ist, desto wahrscheinlicher leben dort Grauspechte.

Durch die vorgestellten Kovariablen wird allerdings nicht die gesamte Variabilität des Modells für Grauspechte erklärt. Es besteht immer noch eine große unbeobachtete Heterogenität, die in der Modellkomponente $f_s(\mathbf{x}, s)$ dargestellt wird. Man kann nicht davon ausgehen, dass alle wirklich einflussreichen Kovariablen erfasst wurden. Diese unbeobachteten Kovariablen werden im räumlichen Effekt zusammengefasst. Abbildung 15 stellt diese grafisch dar.

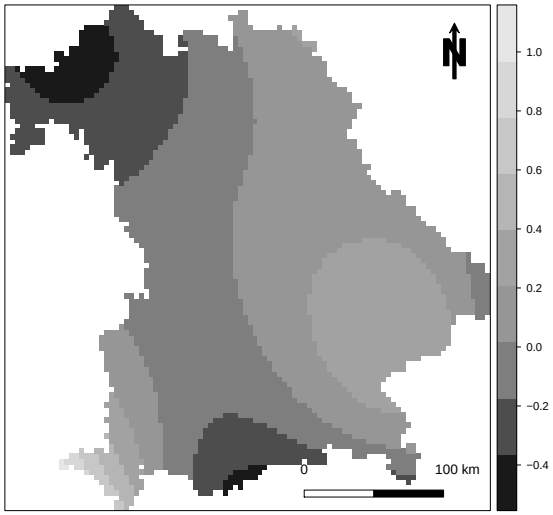


Abbildung 15: Geschätzter räumlicher Effekt im Modell Add/Spatial Grauspecht.

Aus der räumlichen Komponente des Add/Spatial-Modells für den Grauspecht (Abbildung 15) liest sich (wie beim räumlichen Effekt für den Wespenbussard in Abbildung 9 auch) eine erhöhte Auftrittswahrscheinlichkeit in Unterfranken und im Raum Garmisch-Partenkirchen. Zusätzlich können Grauspechte vermehrt in

Mittelfranken und im südwestlichen Oberbayern auftreten. Dies deckt sich auf den ersten Blick mit dem beobachteten Vorkommen von Grauspechten in Abbildung 3. Zum Vergleich finden sich in Abbildung 16 die gefitteten Werte für das (gesamte) Modell Add/Spatial für die Grauspechte.

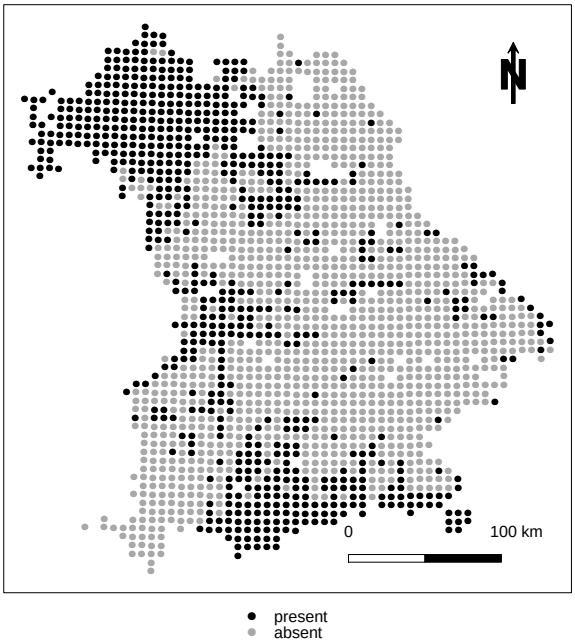


Abbildung 16: Gefittetes Vorkommen der Grauspechte im Modell Add/Spatial.

Auch hier bestätigt sich das erhöhte Vorkommen in Unterfranken, im südwestlichen Oberbayern und im Grenzbereich zwischen Mittelfranken und Schwaben. Doch für das vereinzelte Vorkommen (im Vergleich zu den Ballungsgebieten bei Garmisch-Partenkirchen und in Unterfranken) werden in der östlichen Hälfte Bayerns nur sehr wenige Grauspechte prognostiziert bzw. geschätzt.
Sowohl für das Modell Add/Spatial für den Grauspecht, als auch für das Modell Spatial für den Wespenbussard gilt somit, dass sehr oft zu wenige Vorkommen vorhergesagt werden (vgl. mit Abbildungen 3 und 4).

5 Zusammenfassung und Diskussion

Das Ziel dieser Arbeit war, ein Habitatmodell für das Brutverhalten von Grauspecht und Wespenbussard in Bayern zu erstellen. Dazu wurde ein generalisiertes additives Modell mit Binomial-verteiltem Response geschätzt. Der Prädiktor wurde in eine globale und eine lokale Komponente aufgeteilt und die Effekte mit der Methode „Spatial Boosting“ geschätzt.

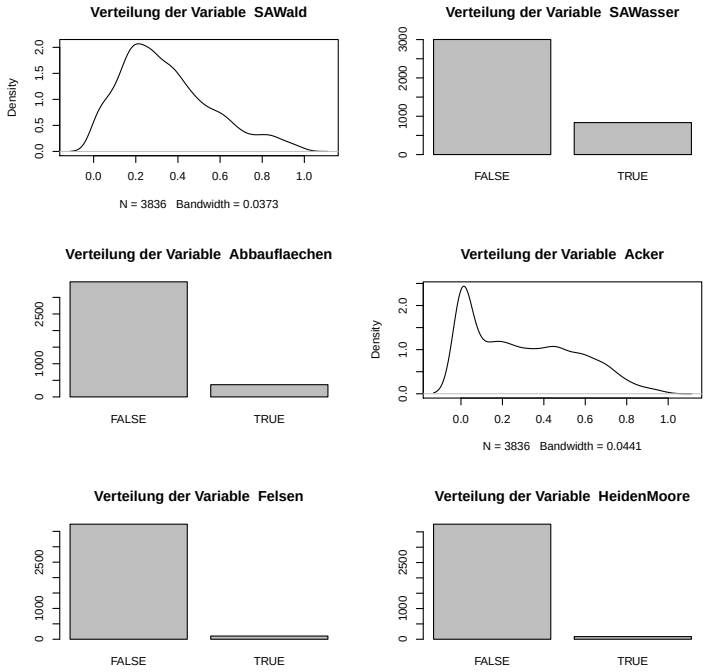
Das angepasste Modell Spatial für den Wespenbussard besteht nur aus der räumlichen Komponente, das heißt die gesamte Heterogenität wird anhand der räumlichen Verteilung erklärt.

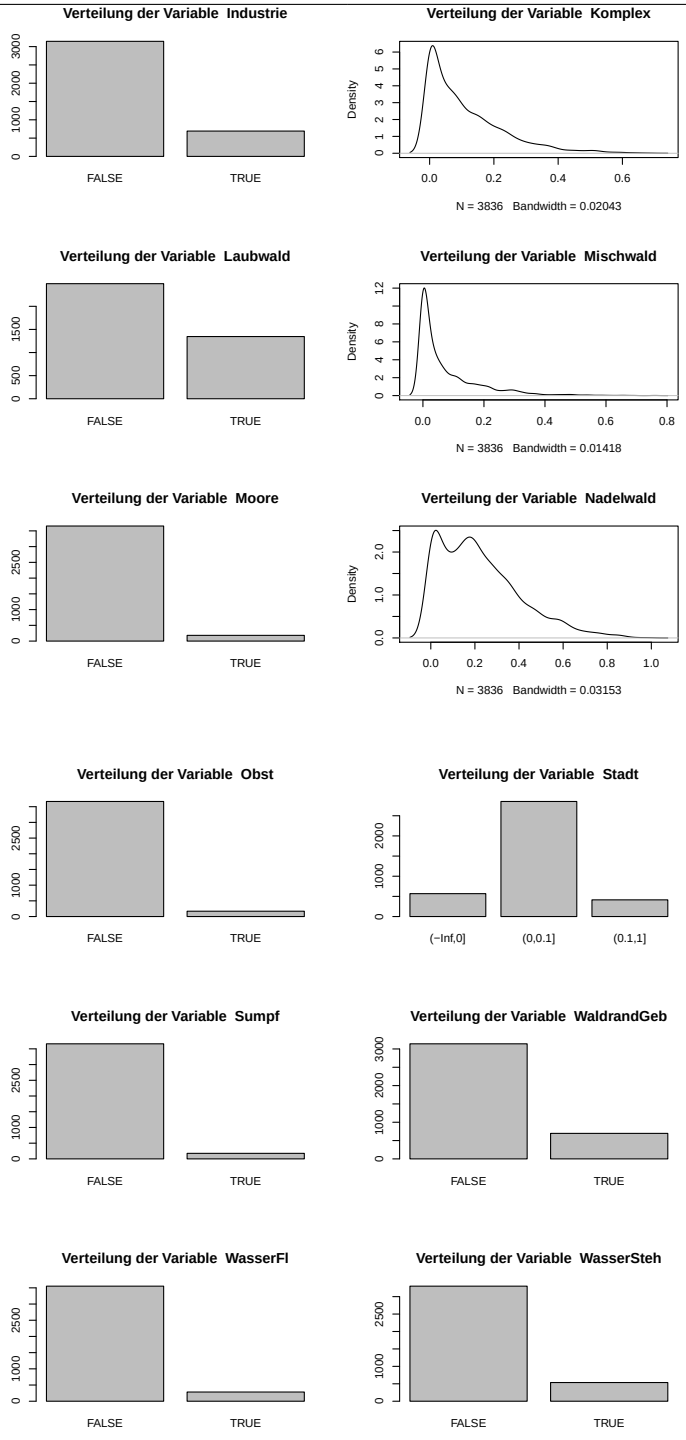
Im angepassten Modell Add/Spatial für den Grauspecht stellt man fest, dass der Hauptteil der Variabilität durch die Umweltvariablen und wenig durch die räumliche Komponente erklärt wird, die im Vergleich dazu nur einen kleinen Teil der Variabilität ausmachen. Bei den Klima- und Bodenfaktoren ist der Effekt der Klimavariable „Saisonabhängiger Temperaturunterschied“ am differenziertesten. Der positive Trend in der Variable „Acker“ kann zwar durch Habitate in Feldrainen und Bracheflächen erklärt werden, muss aber durch den Effekt der intensiven Landwirtschaft relativiert werden. Ähnlich verhält es sich mit dem negativen Effekt der Variable „Mischwald“, die widersprüchliche Aussagen zur Literatur liefert. Jedoch lässt sich durch den positiven Einfluss der komplexen Parzellenstruktur behaupten, dass je abwechslungsreicher die Landschaft ist, desto größer ist die Wahrscheinlichkeit, dass dort Grauspechte leben.

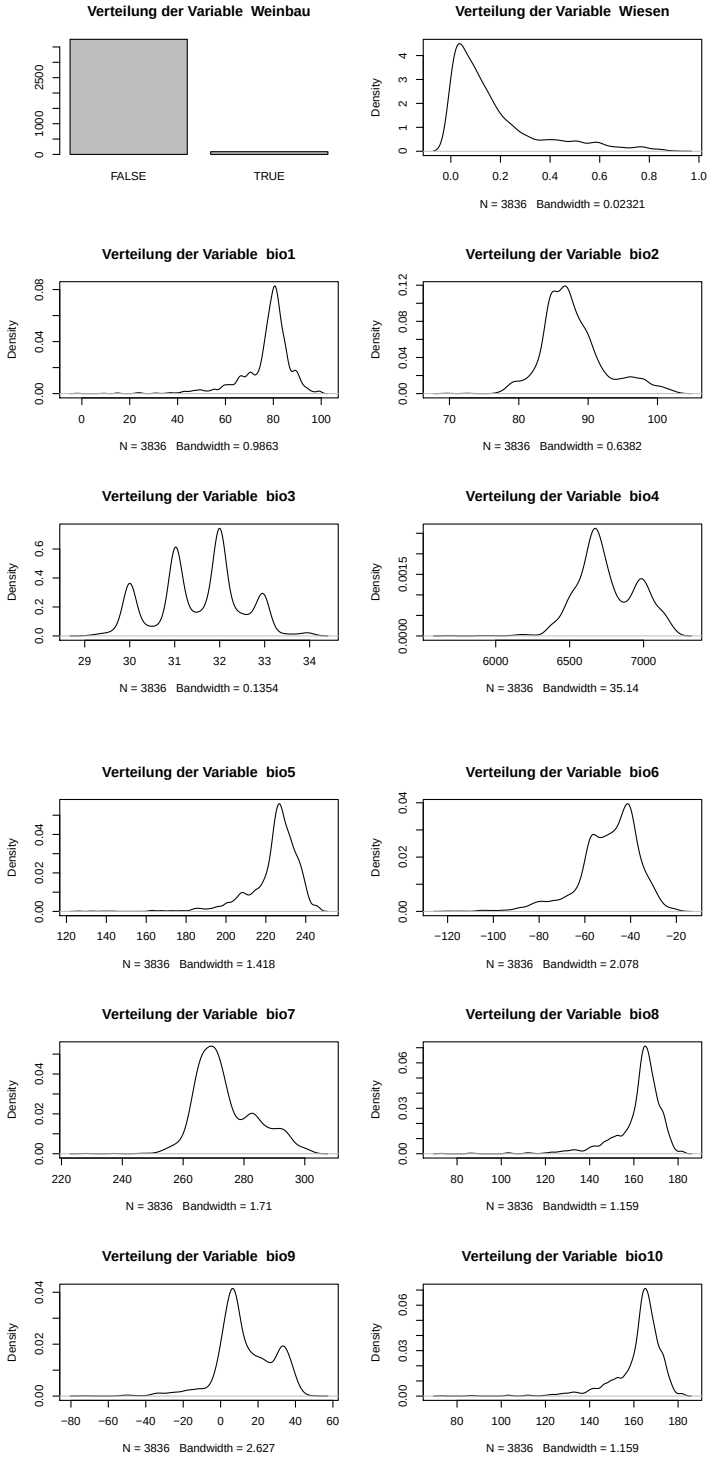
Die hier angewandte Methode des Spatial Boostings bietet eine sehr große Flexibilität zur Modellanpassung durch die Aufspaltung der Einflussfaktoren in globale und lokale Komponenten. So können alle häufig bei Habitatmodellen auftretenden Schwierigkeiten wie Interaktionen zwischen Variablen, nicht-lineare Effekte, nicht-stationäre Einflüsse und räumliche Autokorrelationen beachtet und ins Modell aufgenommen werden. In anderen Anwendungen kann sogar zusätzlich eine räumlich-zeitliche Autokorrelation modelliert werden. Die Neuerung dabei ist, dass dies alles nicht einzeln im Modell beachtet und die anderen Effekte ignoriert werden müssen, sondern, dass gleichzeitig auf alle diese Probleme eingegangen werden kann. Die Zerlegung der Modellkomponenten macht es auch einfacher, Vorhersagen für andere Gebiete und Zeiträume zu treffen als die erhobenen, ohne stark verzerrte Schätzer zu erhalten. Schließlich erhalten wir sehr sparsame Modelle mit nur wenigen einflussreichen Variablen. Dies geschieht durch die effektive Variablenselektion im Boosting-Verfahren und die Vermeidung der Aufnahme nicht-informativer Parameter ins Modell mit der Stability Selection.

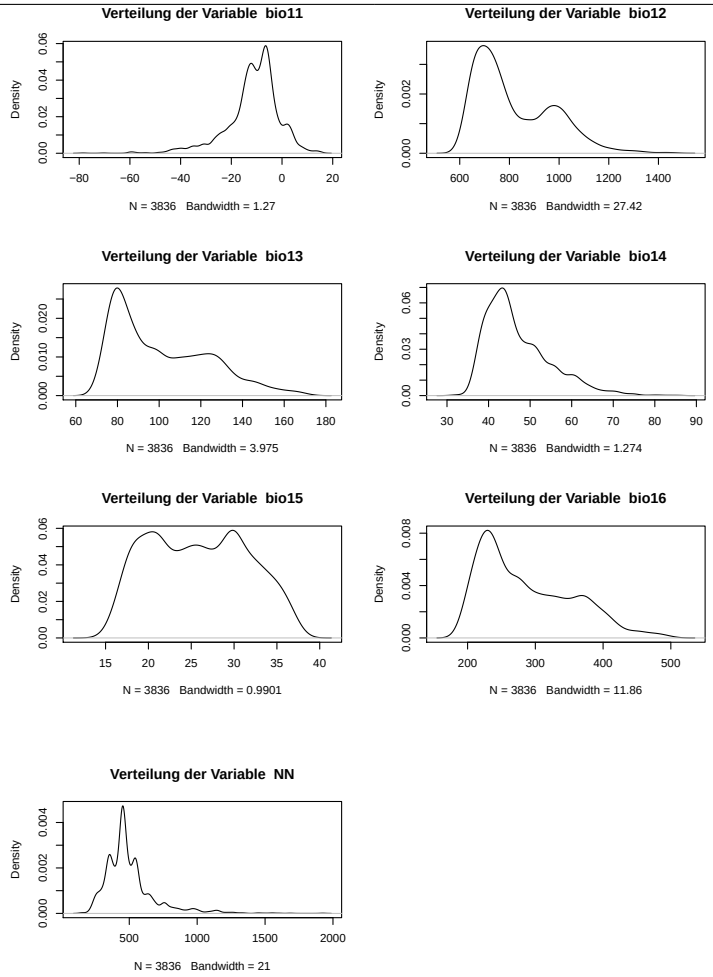
A Anhang

A.1 Verteilung der Bodennutzungs- und Umweltvariablen




Verteilung der Variable Industrie
FALSE
TRUE
Verteilung der Variable Komplex
Density
N = 3836 Bandwidth = 0.02043
Verteilung der Variable Laubwald
FALSE
TRUE
Verteilung der Variable Mischwald
Density
N = 3836 Bandwidth = 0.01418
Verteilung der Variable Moore
FALSE
TRUE
Verteilung der Variable Nadelwald
Density
N = 3836 Bandwidth = 0.03153
Verteilung der Variable Obst
FALSE
TRUE
Verteilung der Variable Stadt
(−Inf,0]
(0,0.1]
(0.1,1]
Verteilung der Variable Sumpf
FALSE
TRUE
Verteilung der Variable WaldrandGeb
FALSE
TRUE
Verteilung der Variable WasserFl
FALSE
TRUE
Verteilung der Variable WasserSteh
FALSE
TRUE


Verteilung der Variable Weinbau
FALSE
TRUE
Verteilung der Variable Wiesen
Density
N = 3836 Bandwidth = 0.02321
Verteilung der Variable bio1
Density
N = 3836 Bandwidth = 0.9863
Verteilung der Variable bio2
Density
N = 3836 Bandwidth = 0.6382
Verteilung der Variable bio3
Density
N = 3836 Bandwidth = 0.1354
Verteilung der Variable bio4
Density
N = 3836 Bandwidth = 35.14
Verteilung der Variable bio5
Density
N = 3836 Bandwidth = 1.418
Verteilung der Variable bio6
Density
N = 3836 Bandwidth = 2.078
Verteilung der Variable bio7
Density
N = 3836 Bandwidth = 1.71
Verteilung der Variable bio8
Density
N = 3836 Bandwidth = 1.159
Verteilung der Variable bio9
Density
N = 3836 Bandwidth = 2.627
Verteilung der Variable bio10
Density
N = 3836 Bandwidth = 1.159


Verteilung der Variable bio11
Density
N = 3836 Bandwidth = 1.27
Verteilung der Variable bio12
Density
N = 3836 Bandwidth = 27.42
Verteilung der Variable bio13
Density
N = 3836 Bandwidth = 3.975
Verteilung der Variable bio14
Density
N = 3836 Bandwidth = 1.274
Verteilung der Variable bio15
Density
N = 3836 Bandwidth = 0.9901
Verteilung der Variable bio16
Density
N = 3836 Bandwidth = 11.86
Verteilung der Variable NN
Density
N = 3836 Bandwidth = 21

A.2 Inhalt der CD

Alle Berechnungen und Modellanpassungen für diese Bachelorarbeit wurden durchgeführt mit der R-Version 2.11.0 (R Development Core Team, 2009) und dem Paket „**mboost**“ (R package version 2.0-3) (Bühlmann und Hothorn, 2007).
Die beiliegende CD enthält neben der digitalen Ausgabe der vorliegenden Arbeit den gesamten R-Code sowie den vollständigen Datensatz, mit dem alle Berechnungen reproduziert werden können.

Literatur

Bates D, Maechler M (2010). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 0.999375-33, URL `http://CRAN.R-project.org/package=lme4`.

Bauer HG, Berthold P (1997). *Die Brutvögel Mitteleuropas: Bestand und Gefährdung*. 2. Wiesbaden.

Bivand RS, Pebesma EJ, Gomez-Rubio V (2008). *Applied Spatial Data Analysis with R*. Springer, NY. URL `http://www.asdar-book.org/`.

Bühlmann P, Hothorn T (2007). "Boosting Algorithms: Regularization, Prediction and Model Fitting (with Discussion)." *Statistical Science*, **22**(4), 477–505.

Colwell RK, Lees DC (2000). "The Mid-Domain Effect: Geometric Constraints on the Geography of Species Richness." *Trends in Ecology and Evolution*, **15**, 70–76.

Deutsches Zentrum für Luft-und Raumfahrt eV DF (ed.) (2005). *CORINE Land Cover 2000 – Europaweit harmonisierte Aktualisierung der Landnutzungsdaten für Deutschland*, volume UBA-FB000826. URL `http://www.corine.dfd.dlr.de/`.

Dormann CF, McPherson JM, Araujo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Jetz W, Kissling WD, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM, Wilson R (2007). "Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review." *Ecography*, **30**, 609–628.

Fahrmeir L, Kneib T, Lang S (2004). "Penalized Structured Additive Regression for Space-Time Data: A Bayesian Perspective." *Statistica Sinica*, **14**, 715–745.

fÃ¼r Wald und Forstwirtschaft BL (2002). "Die zweite Bundeswaldinventur 2002: Ergebnisse fÃ¼r Bayern - LWF-Wissen 49 - Waldfläche und Waldstruktur." URL http://www.lwf.bayern.de/veroeffentlichungen/lwf-wissen/49/lwf-wissen-49_02.pdf.

Graves S, with help from Sundar Dorai-Raj HPP (2006). *multcompView: Visualizations of Paired Comparisons*. R package version 0.1-0.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005). "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *International Journal of Climatology*, **25**, 1965–1978. URL http://www.worldclim.org.

Hothorn T, Bretz F, Westfall P (2008). "Simultaneous Inference in General Parametric Models." *Biometrical Journal*, **50**(3), 346–363.

Hothorn T, Bühlmann P, Kneib T, Schmid M, Hofner B (2010a). *Model-Based Boosting*. R package version 2.0-3, URL http://CRAN.R-project.org/package=mboost.

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.

Hothorn T, Müller J, Schröder B, Kneib T, Brandl R (2010b). "Decomposing Environmental, Spatial, and Spatiotemporal Components of Species Distributions." *Ecological Monographs*. Accepted 2010-07-15.

Kneib T, Hothorn T, Tutz G (2007). "Variable Selection and Model Choice in Geoadditive Regression Models." URL http://epub.ub.uni-muenchen.de/2063/.

Kneib T, Müller J, Hothorn T (2008). "Spatial Smoothing Techniques for the Assessment of Habitat Suitability." *Environment and Ecological Statistics*, **15**, 343–364.

Legendre P (1993). "Spatial Autocorrelation: Trouble or new Paradigm." *Ecology*, **74**(6), 1659–1673.

Meinshausen N, Bühlmann P (2010). "Stability Selection." *Journal of the Royal Statistical Society, Series B*, **72**(4), 1–32.

Neuwirth E (2007). *RColorBrewer: ColorBrewer palettes*. R package version 1.0-2.

Pebesma EJ, Bivand RS (2005). "Classes and Methods for Spatial Data in R." *R News*, **5**(2), 9–13. URL `http://CRAN.R-project.org/doc/Rnews/`.

Pfeifer R, Jörg Müller JSuRB (2009). "Welchen Einfluss haben urbane Lebensräume auf die Artenvielfalt? Eine quantitative Analyse am Beispiel der Vogelwelt Bayerns."

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Sarkar D (2009). *lattice: Lattice Graphics*. R package version 0.17-26, URL `http://CRAN.R-project.org/package=lattice`.

von Blotzheim UG, Bauer K, Bezzel E (1989). *Handbuch der Vögel Mitteleuropas Band 4*. 2. AULA-Verlag, Wiesbaden.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelorarbeit selbstständig verfasst und dabei ausschließlich die angegebenen Quellen und Hilfsmittel verwendet habe. Ich habe diese Arbeit noch nicht einer anderen Prüfungsbehörde vorgelegt und noch nicht veröffentlicht.

München, den 24. Oktober 2010

(Julia Schiele)