LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Katharina Zink

Habitatmodelle für Heuschrecken in Bayern

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

27. August 2010

Danksagung

Diese Bachelorarbeit entstand am Institut für Statistik der Ludwig-Maximilians-Universität München in Zusammenarbeit mit Dr. Jörg Müller vom Nationalpark Bayerischer Wald, der mir die Daten für dieses interessante Thema zur Verfügung gestellt hat. An ihn und seinen Kollegen Dr. Claus Bäßler geht mein Dank für die kompetente Unterstützung in allen biologischen Fragen.
An dieser Stelle möchte ich mich weiterhin bei meinen Betreuern Prof. Dr. Torsten Hothorn, Nikolay Robinzonov und Esther Herberich bedanken für die freundliche und engagierte Betreuung und die vielen hilfreichen Gespräche und Anregungen.

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. Heuschrecken sind in Bayern traditionell eine gut dokumentierte und untersuchte Tierordnung, bereits seit dem 18. Jahrhundert gibt es Aufzeichnungen über die verschiedenen Arten und ihre bevorzugten Lebensräume. Aktuell sind in Bayern 71 Heuschrecken- und Grillenarten bekannt und nachgewiesen (Schlumprecht und Waeber, 2003), wovon 46 Arten auf der Roten Liste in Bayern stehen und damit vom Aussterben bedroht sind. Nicht nur deshalb ist es wichtig, möglichst gute Habitatmodelle zu erstellen, um die Lebensräume und damit die Heuschrecken besser schützen zu können. Es werden zwei Ordnungen mit den jeweiligen Familien unterschieden:

- Kurzfühlerschrecken (Dornschrecken und Feldheuschrecken)
- Langfühlerschrecken (Laubheuschrecken, Höhlenschrecken und Grillen)

Die häufigsten Arten zählen zu den Feld- und Laubheuschrecken:

- Feldheuschrecken: Gemeiner Grashüpfer (*Chorthippus parallelus*), Nachtigall-Grashüpfer (*Chorthippus biguttulus*), Wiesengrashüpfer (*Chorthippus dorsatus*), Brauner Grashüpfer (*Chorthippus brunneus*)
- Laubheuschrecken: Roesels Beißschrecke (*Metrioptera roeselii*), Gemeine Strauchschrecke (*Pholidoptera griseoaptera*), Grünes Heupferd (*Tettigonia viridissima*) .

Im speziellen Fokus in der Biologie steht die Artenvielfalt. Die durchschnittliche Artenzahl in Bayern liegt bei 15.8 Heuschreckenarten pro Quadrant (34 km^2). Die

maximale Artenvielfalt beträgt 41 Arten. An keinem Ort in Bayern kommen also alle existierenden Heuschreckenarten zugleich vor. Für jede dieser einzelnen Arten ist bekannt, in welchen Biotopen sie bevorzugt leben und durch welche Umweltvariablen sie besonders beeinflusst werden. So bevorzugen sie allgemein sonnige und extensiv bewirtschaftete Lebensräume. Einzelne heiße Sommertage oder Frost im Winter beeinflussen die Populationen weniger als insgesamt besonders warme oder kühle Jahre. Vor allem ausgeprägte Nass- oder Trockenjahre wirken sich stark negativ auf die Bestandsentwicklungen aus.

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 1. 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 (Temperatur, Niederschlag, Bodennutzung). Sie bietet verschiedene Möglichkeiten, komplexere Strukturen, wie z.B. 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 den angewendeten 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 die Artenzahl der Heuschrecken zu erstellen.

2 Datenbeschreibung

2.1 Herkunft der Daten und Bearbeitung

Der Datensatz wurde zur Verfügung gestellt vom Nationalpark Bayerischer Wald. Die Zielvariable „Anzahl von Heuschreckenarten“ stammt aus dem Heuschreckenatlas von Bayern (Schlumprecht und Waeber, 2003), der für die gesamte Fläche Bayerns aufgeteilt in durchschnittlich 33.9 km^2 große Quadranten erfasst, welche der 71 erfassten Heuschreckenarten jeweils vorkommen. Für jeden Quadranten wurde daraus die Artenzahl berechnet, die angibt, wie viele verschiedene Heuschreckenarten dort insgesamt leben. Außerdem wurde erfasst wie viele Exkursionen jeweils gemacht wurden, wobei die Beobachtungen mit 0 Exkursionen und keinen gefundenen Heuschrecken aus dem Datensatz entfernt wurden.

Die Kovariablen setzen sich zusammen aus den Klima- und Bodennutzungsfaktoren. Die Klimavariablen stammen aus dem Projekt WorldClim, das sich zum Ziel gesetzt hat, die wichtigsten Klimadaten für alle Regionen der Erde zu er-

fassen. 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 19 bioklimatische Variablen, die biologisch bedeutender sind, da man sie besser interpretieren kann. Sie beschreiben beispielsweise Jahrestrends, Saisonalität und Extremwerte sowie eventuelle limitierende Umweltfaktoren. Nur diese Bioclim-Variablen werden im Weiteren betrachtet. Diese sind in Tabelle 1 aufgeführt. Die Daten beruhen hauptsächlich auf Messungen der Jah-

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
Niederschlag im feuchtesten Quartal	bio16	metrisch
Niederschlag im trockensten Quartal	bio17	metrisch
Niederschlag im wärmsten Quartal	bio18	metrisch
Niederschlag im kältesten Quartal	bio19	metrisch

Tabelle 1: Bioklimatische Variablen von WorldClim

re 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 LandCover-Projekt 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 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 selten vorkamen. 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 kategorisiert mit den Ausprägungen $= 0$ und > 0. 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 $\frac{Höhe - min(Höhe)}{max(Höhe)}$ 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 jeweils Durchschnittswerte gebildet für die ca. 40 km^2 großen Quadranten, für die die

Variable	Messniveau
Wald (SAWald)	metrisch
Wasser (SAWasser)	kategorial: $= 0$, > 0
Abbauflächen	kategorial: $= 0$, > 0
Acker	metrisch
Deponien	nicht verwendet
Felsen	kategorial: $= 0$, > 0
Gletscher	nicht verwendet
Heiden und Moore (HeidenMoore)	kategorial: $= 0$, > 0
Industrie	kategorial: $= 0$, > 0
Komplex	metrisch
Laubwald	metrisch
Mischwald	metrisch
Moore	kategorial: $= 0$, > 0
Nadelwald	metrisch
Obst	kategorial: $= 0$, > 0
Stadt	kategorial: $= 0$, $0 < \mathrm{x} \leq 0.1$, > 0.1
Sumpf	kategorial: $= 0$, > 0
Verkehr	nicht verwendet
Waldrandgebiet (WaldrandGeb)	kategorial: $= 0$, > 0
Fließende Gewässer (WasserFl)	kategorial: $= 0$, > 0
Stehende Gewässer (WasserSteh)	kategorial: $= 0$, > 0
Weinbau	kategorial: $= 0$, > 0
Wiesen	metrisch

Tabelle 2: Bodennutzungsvariablen von CORINE.

Heuschreckendaten vorlagen.

2.2 Deskriptive Analyse

In Abbildung 1 sind die Verteilungen einiger bioklimatischer Variablen dargestellt: „Jahresdurchschnittstemperatur“ (bio1) in °C (multipliziert mit 10), „Jahresniederschlag“ (bio12) in mm und „Isothermalität“ (bio3) in %. Die übrigen Variablen befinden sich in Anhang A.1.

In Abbildung 2 sind beispielhaft die Verteilungen einiger Bodennutzungsvariablen

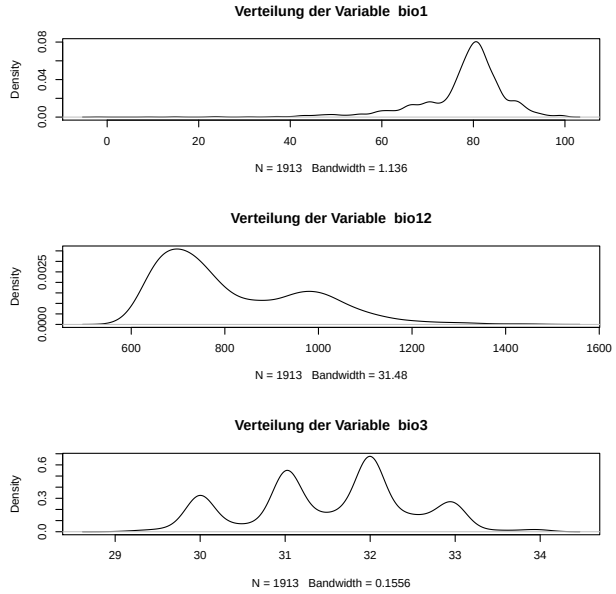


Abbildung 1: Verteilung ausgewählter bioklimatischer Variablen: „Jahresdurchschnittstemperatur“ (bio1), „Jahresniederschlag“ (bio12), „Isothermalität“ (bio3).

abgebildet: „Waldanteil“ (SAWald), „Ackeranteil“ (Acker), „Stadtanteil“ (Stadt) und „Höhe über NN“ (GewHoehe) in m. 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.

In Bayern sind zur Zeit 71 Heuschreckenarten bekannt und nachgewiesen. Zu den häufigsten Arten (in über 1200 Quadranten gefunden) zählen:

- Gemeiner Grashüpfer (*Chorthippus parallelus*) (1805)
- Roesels Beißschrecke (*Metrioptera roeselii*) (1719)

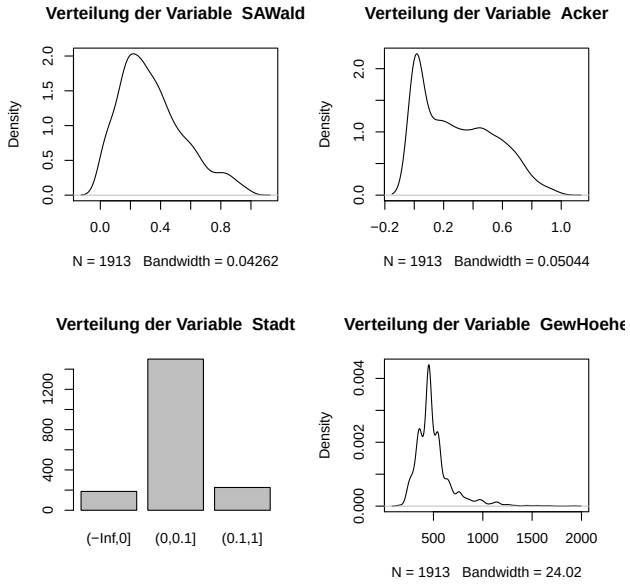


Abbildung 2: Verteilung ausgewählter Bodennutzungsvariablen: „Waldanteil“ (SAWald), „Ackeranteil“ (Acker), „Stadtanteil“ (Stadt), „Höhe über NN“ (GewHoehe).

- Nachtigall-Grashüpfer (*Chorthippus biguttulus*) (1667)
- Gemeine Strauchschrecke (*Pholidoptera griseoaptera*) (1604)
- Grünes Heupferd (*Tettigonia viridissima*) (1429)
- Wiesengrashüpfer (*Chorthippus dorsatus*) (1325)
- Brauner Grashüpfer (*Chorthippus brunneus*) (1320)

In Abbildung 3 erhält man einen groben Überblick über die Verteilung der Gesamtartenzahl in Bayern. Im bereinigten Datensatz ist die Mindestartenzahl 1, maximal wurden 41 Arten in einem Grid entdeckt, also 58% aller in Bayern existierenden Arten. Offensichtlich ist die Artenzahl nicht homogen in Bayern verteilt. Besonders wenig Arten gibt es beispielsweise im südlichen Niederbayern. Dagegen ist die Artenvielfalt im Raum Bayreuth (Oberfranken) sehr hoch.

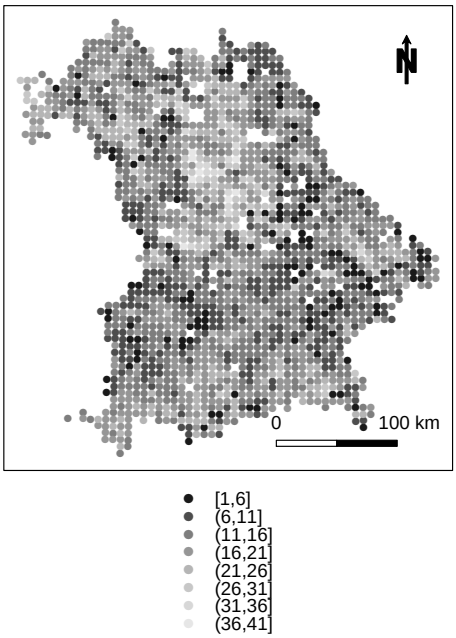


Abbildung 3: Artenzahl der Heuschrecken in Bayern.

2.3 Besonderheiten der Daten

Im Folgenden soll ein Habitatmodell zur Untersuchung der Heuschreckenartenzahl 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 die Entdeckung einer Heuschreckenart in einem Quadranten, die Wahrscheinlichkeit dafür, dass sich diese Art auch im Nachbarquadranten befindet, erhöhen, obwohl dies durch die beob-

achteten 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 in Hothorn *et al.* (2010b).

3 Methoden

3.1 Generalisiertes additives Modell

Bei den vorliegenden Daten handelt es sich um eine Zählvariable als Response, daher gilt die Annahme, dass $Y_i|\mathbf{x}_i \sim Po(\lambda_i)$, wobei der Parameter λ_i dem Erwartungswert und der Varianz entspricht. Im generalisierten additiven Modell geht man davon aus, dass sich der Prädiktor η_i additiv aus glatten eindimensionalen Funktionen der einzelnen Kovariablen zusammensetzt:

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

Über die Exponentialfunktion wird der Prädiktor mit der erwarteten Artenzahl λ_i verknüpft:

$$\lambda_i = \mathbb{E}(\text{Artenzahl}_i|\mathbf{x}_i, s_i) = \exp(f(\mathbf{x}_i, s_i)) \tag{1}$$

Das bedeutet, dass die mittlere erwartete Artenzahl an einem Punkt s_i, abhängig von den Umweltvariablen $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ dem Wert der Exponentialfunktion ausgewertet an der Stelle der Regressionsgleichung entspricht.

Es wird jedoch vermutet, dass der beobachtete Response auch durch die Zahl der

Exkursionen im Quadranten i ($\#\text{Exkursionen}_i$) beeinflusst wird, dass also mit einer erhöhten Anzahl an Exkursionen in einem Feld auch die erwartete Anzahl der dort gefundenen Arten steigt. Deshalb wird die Exkursionenzahl als Offset in den Prädiktor aufgenommen, deren Effekt auf 1 gezwungen wird. Die erwartete Artenzahl λ_i ist demnach $\lambda_i = \#\text{Exkursionen}_i \cdot \exp(f(\mathbf{x}_i, s_i))$ und es ergibt sich die strukturelle Komponente

$$\mathbb{E}(\text{Artenzahl}_i|\mathbf{x}_i, s_i) = \lambda_i = \#\text{Exkursionen}_i \cdot \exp(f(\mathbf{x}_i, s_i)) =$$

$$= \exp(\underbrace{\log(\#\text{Exkursionen}_i)}_{\text{Offset}} + f(\mathbf{x}_i, s_i)). \tag{2}$$

3.2 Die Methode des Spatial Boosting

Bei hochdimensionalen Datensätzen sind übliche Schätzverfahren, wie z.B. 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 Nonstationarität 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 Nonstationarität 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 unbeobach-

teten Heterogenität, die durch räumliche Autokorrelation oder nonstationäre 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 (2) einfließt, folgende Form:

$$f(\mathbf{x}, s) = \underbrace{f_{env}(\mathbf{x})}_{global} + \underbrace{f_{ns}(\mathbf{x}, s) + f_s(s)}_{lokal} \tag{3}$$

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 nonstationäre Umwelteffekte, d.h. 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 An-

satz 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 z.B. 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 z.B. 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. Die Artenzahl folgt einer $Po(\lambda_i)$ Poissonverteilung mit $\lambda_i = \mathbb{E}(y_i|\mathbf{x}_i, s_i)$ und $\lambda_i(f) = \#\text{Exkursionen}_i \cdot \exp(f(\mathbf{x}_i, s_i))$. Damit ist die negative Log-Likelihood-Funktion

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

mit

$$\rho(y_i, \lambda_i(f)) = \lambda_i - y_i \log \lambda_i$$

als Beitrag einer Beobachtung zur Gesamt-Log-Likelihood.

Die Funktion $\hat{f}$, die die Verlustfunktion minimiert, wird mit einem Component-

wise Functional Gradient Descent Boosting-Algorithmus geschätzt. Für Modelle der Form (3) 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, d.h. $f_s(s) \equiv 0$ und genauso $f_{env}(\mathbf{x}) \equiv 0$, falls keine 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 die Artenzahl von Heuschrecken 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.

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

Nun wird diejenige Basisprozedur g_{j^*} ($f_j(x_j)$, f_{ns} oder f_s) ausgewählt, welche die Residuen am besten beschreibt, d.h. 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 z.B. 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:

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

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 z.B. 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 λ 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 Umwelt-

variablen 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 Nonstationarität oder Interaktionen zu berücksichtigen. Mit Regressionsbäumen als Baselearner für f_{env} können Interaktionen besser modelliert werden, ansonsten ist das Modell Tree/Spatial gleich wie das vorherige. 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 überhaupt keine Restriktionen an die Modellkomponenten gestellt werden. Aus diesen sechs Grundmodellen wird für die vorliegenden Daten das beste Modell ausgewählt (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 4 - 7).

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 jedoch heraus, dass ein weiteres Verringern der Schrittgröße die Ergebnisse für die vorliegenden Daten weiter 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 verschiedenen Schrittgrößen $\nu = 0.1, 0.05, 0.03$ und 0.01 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 Komplizierung des Algorithmus zu vermeiden. Bereits in der Praxis bekannt ist jedoch die Tatsache, dass $\nu = 0.1$ in einem Poissonmodell auf jeden Fall zu groß ist. Die nachfolgenden Boxplots (Abbildungen 4 bis 7) zeigen daher für alle Modelle mit dem optimalen m_{stop} die Out-of-Bootstrap negative Log-Likelihood für mehrere Bootstrapstichproben und für verschiedene Schrittgrößen ν.

Je kleiner die Schrittgröße gewählt wird, umso mehr nähern sich die Modellgüten einander an. Besonders für die Modelle Tree/Spatial und Tree/Vary verbessert sich der Modellfit, je kleiner ν gewählt wird. Bei allen Werten von ν hat immer das Modell Spatial die kleinste negative Log-Likelihood und dementsprechend die beste Modellanpassung. Dies ist ein Hinweis darauf, dass ein großer räumlicher Effekt besteht und f_s eine dominierende Modellkomponente ist. Nicht viel schlechter schneiden die Modelle Add/Spatial und Add/Vary ab. Da beide ungefähr die gleiche Modellgüte 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.

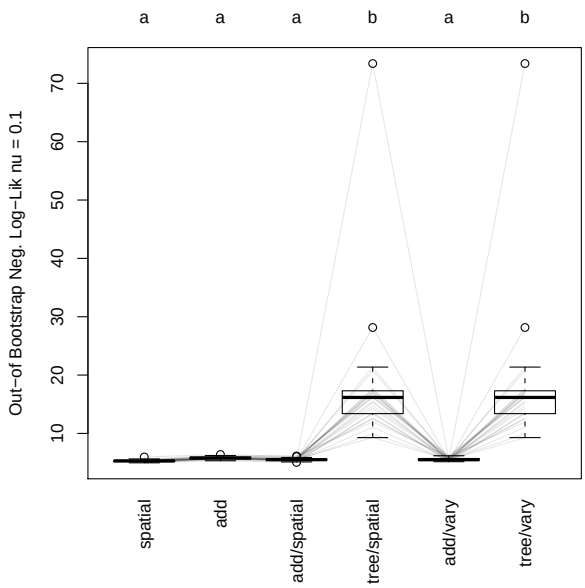


Abbildung 4: Out-of-Bootstrap Negative Log-Likelihoods $\nu = 0.1$.

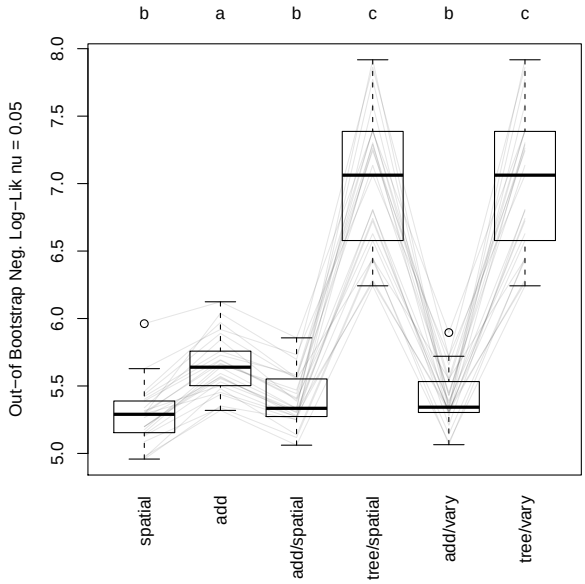


Abbildung 5: Out-of-Bootstrap Negative Log-Likelihoods $\nu = 0.05$.

Als Vergleichsmethode für die Fragestellung, welche der Modelle sich signifikant im Mittelwert der negativen Log-Likelihood unterscheiden, wurde ein mul-

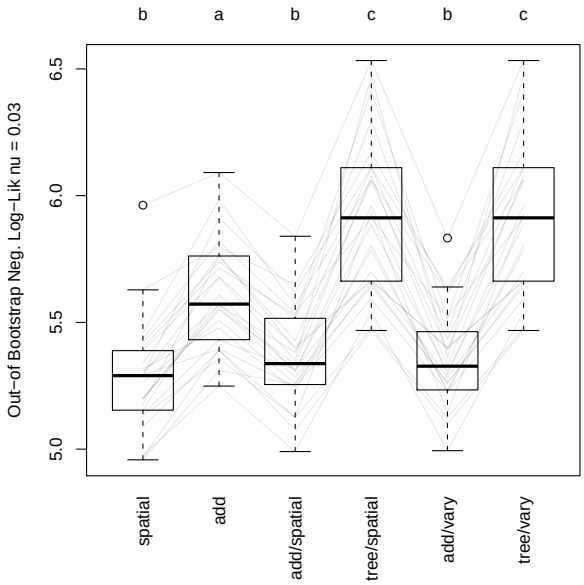


Abbildung 6: Out-of-Bootstrap Negative Log-Likelihoods $\nu = 0.03$.

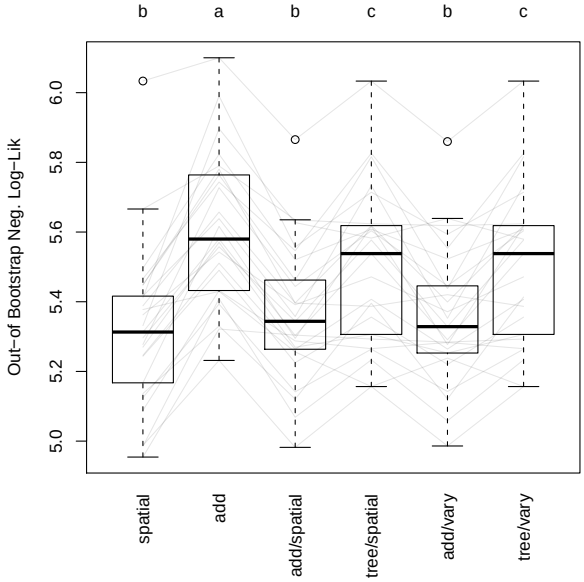


Abbildung 7: Out-of-Bootstrap Negative Log-Likelihoods $\nu = 0.01$.

tipler 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.

Tabelle 4 zeigt, welche Variablen für die verschiedenen Schrittgrößen im Modell Add/Spatial ausgewählt wurden: Es fällt auf, dass immer die Variablen Stadt

Schrittgröße ν	Ausgewählte Variablen
0.10	Stadt, Höhe, bspatial
0.05	bio3, bio13, Stadt, Acker, Höhe, bspatial
0.03	bio3, bio4, bio13, Stadt, Acker, SAWald, Höhe, bspatial
0.01	bio3, bio13, Stadt, Acker, SAWald, Höhe, bspatial

Tabelle 4: Selektierte Variablen für verschiedene Schrittgrößen.

und Höhe und die räumliche Komponente („bspatial“) unter den finalen Variablen sind. Wenn die Schrittgröße verkleinert wird, steigt normalerweise die Zahl der Iterationen und damit die der Modellkomponenten, die im Boosting selektiert werden. Erwartungsgemäß werden damit bei einer Verkleinerung der Schrittgröße auf 0.05 drei Variablen mehr (bio3 [Isothermalität], bio13 [Niederschlag im feuchtesten Monat], Acker) ausgewählt. Bei einer weiteren Reduzierung von ν auf 0.03 werden zusätzlich noch die zwei Variablen bio4 (Saisonalität der Temperatur) und SAWald (Waldanteil) ausgewählt. Wenn nun ν auf 0.01 gesetzt wird, kommt schließlich keine Variable mehr hinzu, im Gegenteil, bio4 fällt weg. Der Effekt dieser Variable war allerdings fast konstant bei 0, die Variable war also nicht sehr einflussreich.

Man kann nun davon ausgehen, mit diesem letzten Modell die bestmögliche Anpassung an die Daten gefunden zu haben. Die weitere Interpretation beschränkt sich auf die Modelle Spatial und Add/Spatial für die Schrittgröße $\nu = 0.01$, die anderen Modellanpassungen finden sich im elektronischen Anhang.

4.2 Interpretation

Im Modell Spatial wird die gesamte Heterogenität nur anhand der räumlichen Verteilung erklärt. In Abbildung 8 sind die relativen Unterschiede in der Artenzahl für den zentrierten räumlichen Effekt dieses Modells gezeichnet. Man sieht, dass besonders im Raum München und im Raum Nürnberg/Fürth/Erlangen die Anzahl der Heuschreckenarten geringer ist als im restlichen Bayern, wenn man den Offset, also den Einfluss der Exkursionenzahl unberücksichtigt lässt. Dies ist auf den ersten Blick widersprüchlich zur beobachteten Verteilung der Artenzahl (Abbildung 3) und zu den gefitteten Werten (Abbildung 9), verdeutlicht aber nochmals den Einfluss der Zahl der Untersuchungen auf den beobachteten Response. Hier macht sich vermutlich bemerkbar, dass die Biotope in der Stadt und in unmittelbarer Nähe dazu leichter zugänglich sind und deswegen öfter besucht werden. Ansonsten fällt die extrem verringerte Artenzahl (relativ gesehen) im südwestlichen Raum Oberallgäu/Lindau sowie im nordwestlichen Raum Aschaffenburg auf. Eine erhöhte Artenzahl findet man in den Regionen Unterfranken (mit Ausnahme Aschaffenburg), zentrale Oberpfalz und westliches Mittelfranken sowie in Westschwaben und im östlichen Oberbayern.

Das Modell Add/Spatial, in das die Effekte der Umweltvariablen als additive glatte Funktionen aufgenommen wurden, hat eine vergleichbar gute Modellanpassung. Als einflussreiche Kovariablen ergeben sich durch Stability Selection die sechs Kovariablen Isothermalität (bio3), Niederschlag im feuchtesten Monat (bio13) und die Höhe über Normalnull sowie der prozentuale Anteil an Waldgebiet, Ackergebiet und Stadtgebiet und die räumliche Komponente. Zuerst überprüft man, wieviel Variabilität überhaupt durch die einzelnen Modellkomponenten erklärt wird. In Abbildung 10 wird deutlich, dass der Hauptteil der Variabilität durch den Offset (*log(Anzahl der Exkursionen)*) erklärt wird. Das bedeutet, dass

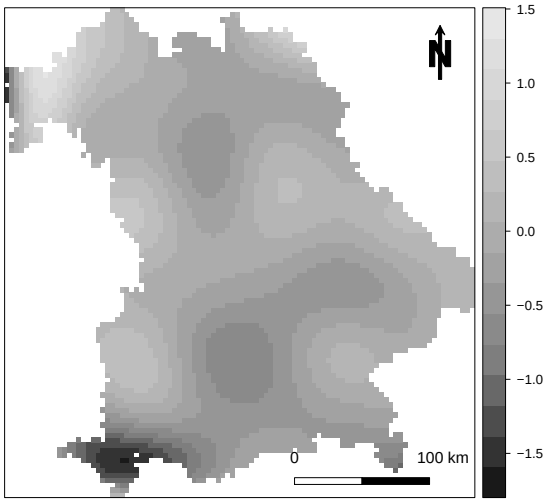


Abbildung 8: Geschätzter räumlicher Effekt im Modell Spatial.

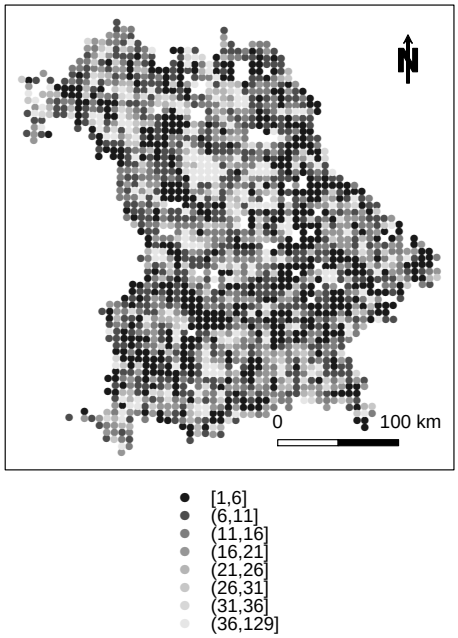


Abbildung 9: Gefittete Artenzahl im Modell Spatial.

die beobachtete Artenzahl hauptsächlich davon abhängt, wie oft ein Quadrant untersucht wird. Nichtsdestotrotz ist klar, dass die tatsächliche Artenzahl nicht von

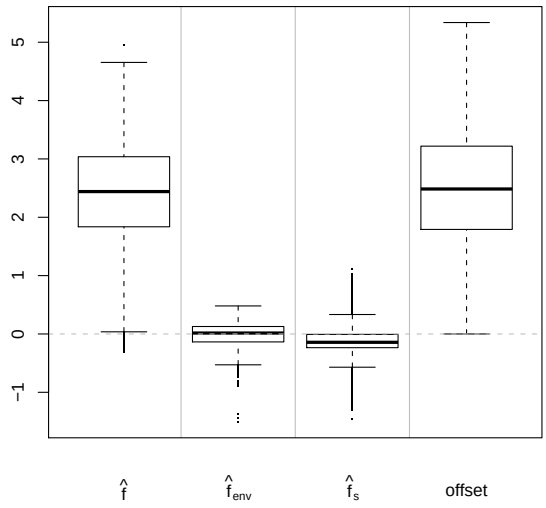


Abbildung 10: Zerlegung der erklärten Variabilität für die einzelnen Modellkomponenten (gefittete Werte auf der Log-Skala).

der Anzahl der Exkursionen abhängen kann. Deswegen wird im Poissonmodell der Parameter λ_i so modifiziert, dass der Effekt der Exkursionenzahl auf 1 gezwungen wird (siehe Kapitel 3.2.2). Dadurch möchte man den Effekt der Exkursionen bereinigen und erhält als weitere erklärende Größe die Umweltvariablen und die räumliche Komponente, deren Einfluss im Vergleich zum Offset jedoch viel geringer ist.

Die geschätzten Effekte der einzelnen Umweltvariablen f_{partial} lassen sich so interpretieren, dass sich die mittlere erwartete Artenzahl 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.

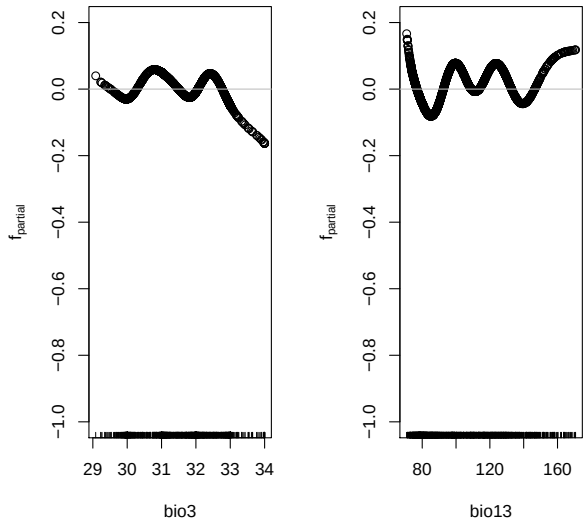


Abbildung 11: Geschätzte partielle Effekte der Umweltvariablen „Isothermalität“ (bio3) und „Niederschlag im feuchtesten Monat“ (bio13).

In Abbildung 11 sieht man die geschätzten Effekte für die Variablen „Isothermalität“ (bio3) und „Niederschlag im feuchtesten Monat“ (bio13). Beide sind nicht eindeutig zu interpretieren, weil die Funktionen insgesamt stark schwanken. Die Isothermalität beschreibt die prozentuale Tagestemperaturschwankung im Vergleich zur Jahresschwankung und ist damit ein starker ökologischer Filter, der das Vorkommen von Tieren und Pflanzen beeinflusst. In Abbildung 1 war bereits erkennbar, dass die Isothermalität im Verlauf sehr schwankend ist, daher ist es nicht verwunderlich, wenn auch der geschätzte Effekt dieser Variable starken Schwankungen unterworfen ist. Allerdings ist ein Trend ersichtlich, der beschreibt, dass die Artenzahl sinkt, wenn die Isothermalität über 33 % steigt. Das bedeutet, je größer die Tagesschwankung ist, umso schwieriger sind die Überlebensbedingungen. Für Werte unter 33 % ist die Artenzahl leicht erhöht oder gleichbleibend.

Die Niederschlagsmenge kann die Produktivität eines Ökosystems widerspiegeln;

je produktiver ein System dabei ist, umso mehr Arten kann es theoretisch beherbergen. Hier sieht man, dass der Niederschlag im feuchtesten Monat positiven Einfluss für eine Menge kleiner als 80 mm und größer als 150 mm sowie zwischen 90 mm und 130 mm hat. Bei einer durchschnittlichen Niederschlagsmenge zwischen 80 mm und 90 mm als auch zwischen 130 mm und 150 mm ist der Einfluss negativ, allerdings sind die Intervalle so klein, dass die Ausschläge nach unten eher vernachlässigt werden können. Der grobe Trend geht leicht nach oben, das heißt mit höherer Niederschlagsmenge im feuchtesten Monat steigt die Produktivität des Ökosystems und damit die Zahl der Heuschreckenarten. Dabei ist zu beachten, dass „Niederschlag im feuchtesten Monat“ und „Jahresniederschlag“ (bio12) natürlich sehr stark miteinander korrelieren (cor = 0.97). Daher kann man verallgemeinern, dass mit steigendem Niederschlag auch die Artenzahl leicht steigt.

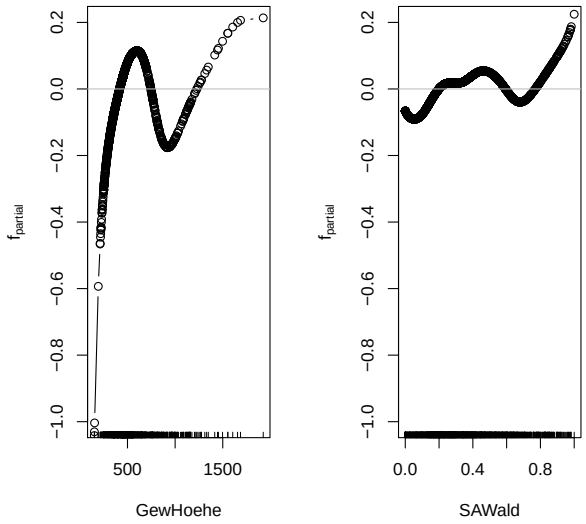


Abbildung 12: Geschätzte partielle Effekte der Umweltvariablen „Höhe über NN“ (GewHoehe) und „Anteil Waldgebiet“ (SAWald).

Die Variable „Höhe“ ist kein direkter physiologischer Faktor, sondern eine Proxy-Variable u.a. für Klima und Fläche. Wenn man sich nur die Werte bis 1200 m ansieht, erkennt man den in der Biologie typischen „Mid-Domain Effect“: Unter 400 m sowie zwischen 750 m und 1200 m erwartet man eine niedrigere mittlere Artenzahl, hingegen für Höhen zwischen 400 m und 750 m eine höhere (Abbildung 12). Dieser parabelförmige Verlauf erklärt sich dadurch, dass sich in den mittleren Höhen viele Ausbreitungsgebiete verschiedener Arten überschneiden und somit zu einem Maximum an Artenvielfalt führen (Colwell und Lees, 2000). Der steigende Trend ab 1200 m lässt sich dahingehend interpretieren, dass es oberhalb der Baumgrenze viele offene Habitatflächen, wie z.B. Almen oder Schotterflächen gibt, in denen Heuschrecken bevorzugt leben (Schlumprecht und Waeber, 2003). Allerdings ist ab 1500 m die Beobachtungszahl sehr gering, sodass diese Aussagen nicht verallgemeinert werden können.

Wenn der Anteil des „Waldgebietes“ pro Quadrant zwischen 20 % und 60 % liegt, steigt die erwartete Artenzahl, genauso für Werte über 80 % Waldbedeckung. Für Flächen mit geringem Waldanteil (unter 20 %) dagegen ist die Zahl der Heuschrecken im Mittel leicht verringert. Die kleine Schwankung ins Negative bei 70 % kann vernachlässigt werden. Insgesamt kann man den Waldanteil als einen Naturnäheindikator für ursprüngliche, naturbelassene Räume interpretieren, in denen die Artenzahl höher liegt als in naturfernen Räumen.

In Abbildung 13 erkennt man einen eindeutig positiven Einfluss der Variable „Ackergebiet“ auf die Artenzahl. Bis zu einer Ackerfläche von ungefähr 20 % ist die Artenzahl verringert, je größer jedoch die prozentuale Bedeckung des Quadranten mit Ackerfläche ist, umso größer wird auch die mittlere erwartete Heuschreckenartenzahl. Zum einen kann man dies durch Habitate in Feldrainen erklä-

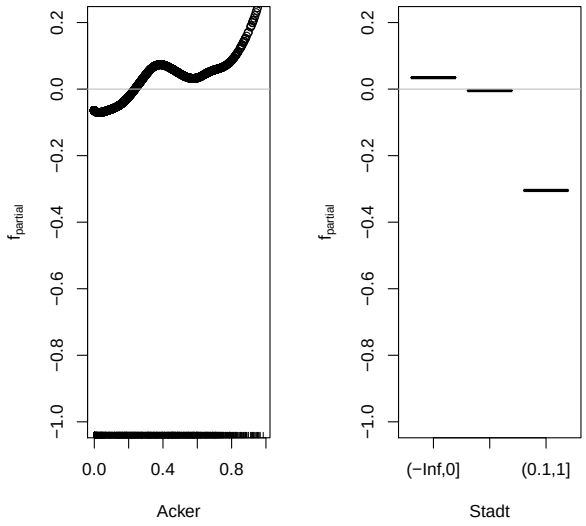


Abbildung 13: Geschätzte partielle Effekte der Umweltvariablen „Anteil Ackergebiet“ (Acker) und „Anteil Stadtgebiet“ (Stadt).

ren, die nicht bewirtschaftet werden, weil sie nur schwer zugänglich sind. Somit bieten sie ideale Lebensbedingungen für Tiere. Zum anderen liegen Ackerflächen zur Erhaltung der Bodenfruchtbarkeit regelmäßig brach und ermöglichen so den Heuschrecken 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 Heuschrecken vorzufinden, weil sie eine starke Barriere zur Ausbreitung der Populationen darstellen (Schlumprecht und Waeber, 2003).

Den umgekehrten Effekt sieht man bei der kategorisierten Variable „Stadt“: je höher der Stadtanteil ist, umso geringer ist die Artenzahl. Bei einer prozentualen Stadtfläche zwischen 10 % und 100 % ist die mittlere erwartete Artenzahl sogar um 25 % geringer als in Gebieten mit kleinerem Stadtanteil. Dies lässt sich durch die fehlenden Habitate in städtischen Gebieten erklären.

Durch die vorgestellten Kovariablen wird allerdings nicht die gesamte Variabilität erklärt. Die räumliche Komponente dominiert, was man nicht allein dadurch sieht, dass das Modell Spatial im Gesamtmodellvergleich am besten abgeschnitten hat. Es besteht immer noch eine sehr 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 14 stellt diese grafisch dar. Die stark verringerte Artenzahl im Oberallgäu wird auch

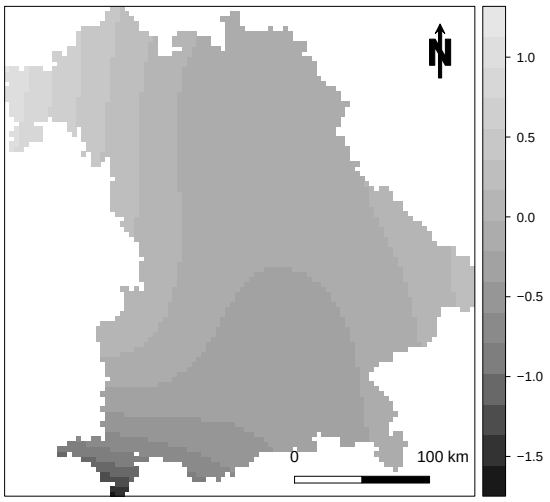


Abbildung 14: Geschätzter räumlicher Effekt im Modell Add/Spatial.

auf dieser Karte wieder sichtbar. Die Ballungsräume München sowie Nürnberg/-Fürth/Erlangen dagegen stechen optisch nicht mehr hervor, da der Stadteffekt bereits durch die Landnutzungsvariable modelliert wurde. Im südlichen Schwaben sowie im Großteil Oberbayerns ist die erwartete Artenzahl eher kleiner als im Rest Bayerns. Je nördlicher die Lage, desto mehr Heuschreckenarten werden erwartet, wenn die oben erwähnten Kovariablen bereits miteinberechnet sind und der Off-

set nicht beachtet wird. Besonders auffallend sind hier das östliche Niederbayern sowie Unterfranken.

Zum Vergleich finden sich in Abbildung 15 die gefitteten Werte für das Modell Add/Spatial. Sie unterscheiden sich kaum von denen des Modells Spatial. Für den Modellfit ist es also unerheblich, ob die Umweltvariablen in das Modell aufgenommen werden oder ob alles als räumlicher Effekt zusammengefasst wird. Für beide Modelle gilt jedoch, dass sehr oft zu niedrige Artenzahlen vorhergesagt werden (vgl. mit Abbildung 3).

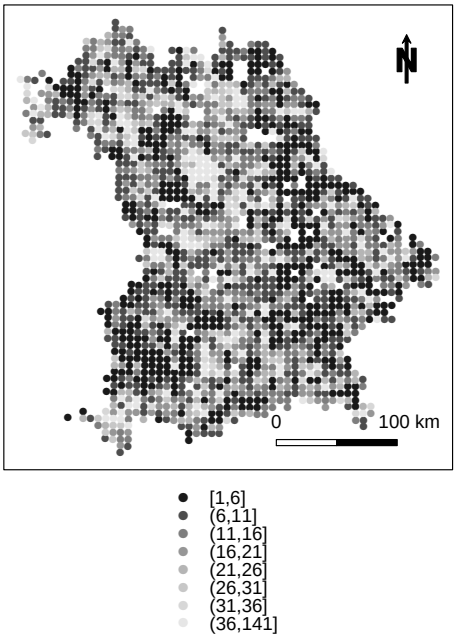


Abbildung 15: Gefittete Artenzahl im Modell Add/Spatial.

5 Zusammenfassung und Diskussion

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

Im angepassten Modell stellte man fest, dass der Hauptteil der Variabilität durch die Anzahl der durchgeführten Exkursionen erklärt wird. Mit großem Abstand folgen die Umweltvariablen und die räumliche Komponente, die im Vergleich dazu nur einen kleinen Teil der Variabilität ausmachen. Bei den Klima- und Bodenfaktoren ist der Effekt der Höhe am differenziertesten, welcher sich durch den für Flora und Fauna typischen „Mid-Domain Effect“ erklärt. 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.

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,

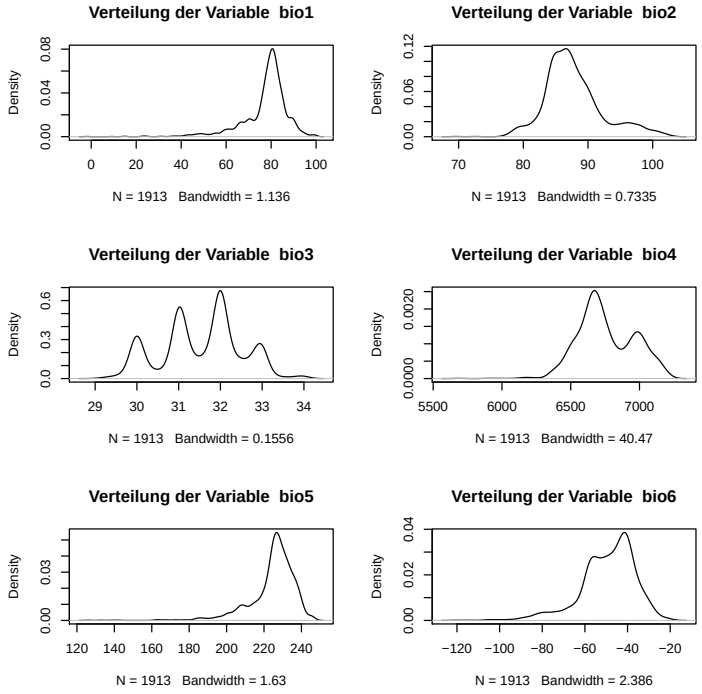
Vorhersagen für andere Gebiete und Zeiträume zu treffen als die erhobenen, ohne stark verzerrte Schätzer zu erhalten. Schließlich erhalten wir sehr sparsame Modelle mit nur wenigen einflussreichen Variablen. Dies geschieht durch die effektive Variablenselektion im Boosting-Verfahren und die Vermeidung der Aufnahme nicht-informativer Parameter ins Modell mit der Stability Selection.

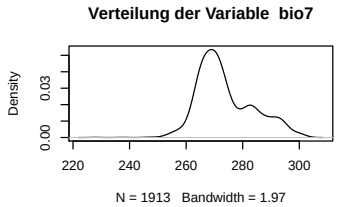
Ein großes Problem der vorliegenden Daten ist die Diskrepanz zwischen tatsächlichem und beobachtetem Response. Die beobachtete Artenzahl ist stark davon abhängig, wie oft ein Quadrant untersucht wurde und steigt natürlich mit der Zahl der Exkursionen. Die tatsächliche Artenvielfalt kann dagegen nicht genau erhoben werden. Dieses Problem liegt jedoch im Datensatz und im Studiendesign selbst. Nur wenn überall gleich viele Untersuchungen vorgenommen werden, kann eine verlässlichere Modellschätzung vorgenommen und damit bessere Prognosen gemacht werden.

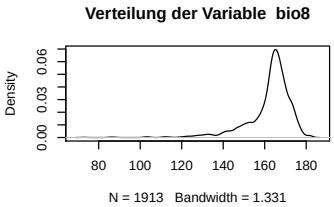
Ebenso problematisch ist die Wahl des Hyperparameters ν, der die Schrittgröße im Boosting-Algorithmus bestimmt. Momentan ist es nicht möglich den Parameter ν ebenso wie die Glättungsparameter der einzelnen Variablen λ und die optimale Anzahl an Iterationen m_{stop} über Kreuzvalidierung oder ähnliche Verfahren zu schätzen. Das würde den Algorithmus zu aufwändig und rechenintensiv machen. Stattdessen muss dieser Parameter per Hand festgelegt werden, wobei man sich an Erfahrungswerten orientieren kann. Allerdings war es auch ersichtlich, dass ν im Vergleich zu m_{stop} nur einen geringen Einfluss auf die Variablenauswahl hat, wenn es klein genug gewählt wird.

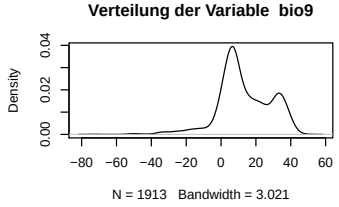
A Anhang

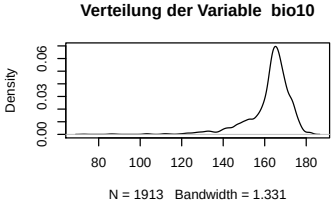
A.1 Verteilung der Umwelt- und Bodennutzungsvariablen

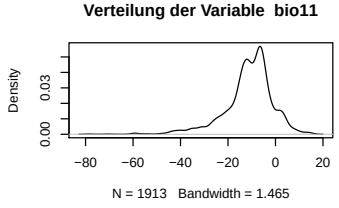


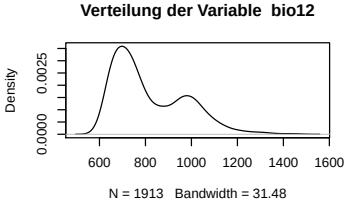

Verteilung der Variable bio7
Density
0.00
0.03
220
240
260
280
300
N = 1913 Bandwidth = 1.97

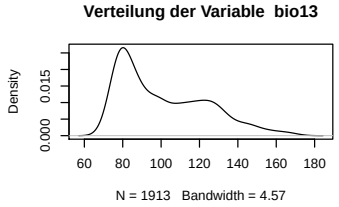

Verteilung der Variable bio8
Density
0.00
0.03
0.06
80
100
120
140
160
180
N = 1913 Bandwidth = 1.331

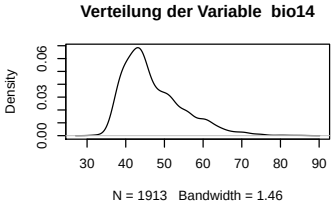

Verteilung der Variable bio9
Density
0.00
0.02
0.04
−80
−60
−40
−20
0
20
40
60
N = 1913 Bandwidth = 3.021

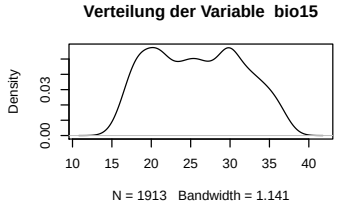

Verteilung der Variable bio10
Density
0.00
0.03
0.06
80
100
120
140
160
180
N = 1913 Bandwidth = 1.331

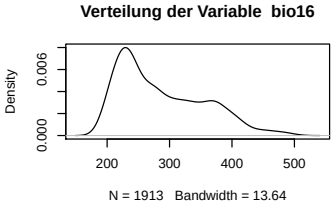

Verteilung der Variable bio11
Density
0.00
0.03
−80
−60
−40
−20
0
20
N = 1913 Bandwidth = 1.465

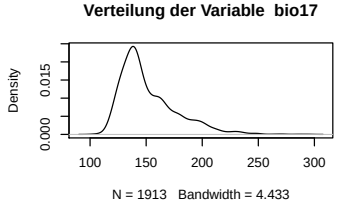

Verteilung der Variable bio12
Density
0.0000
0.0025
600
800
1000
1200
1400
1600
N = 1913 Bandwidth = 31.48

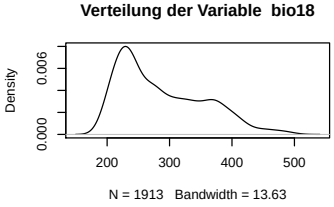

Verteilung der Variable bio13
Density
0.000
0.015
60
80
100
120
140
160
180
N = 1913 Bandwidth = 4.57

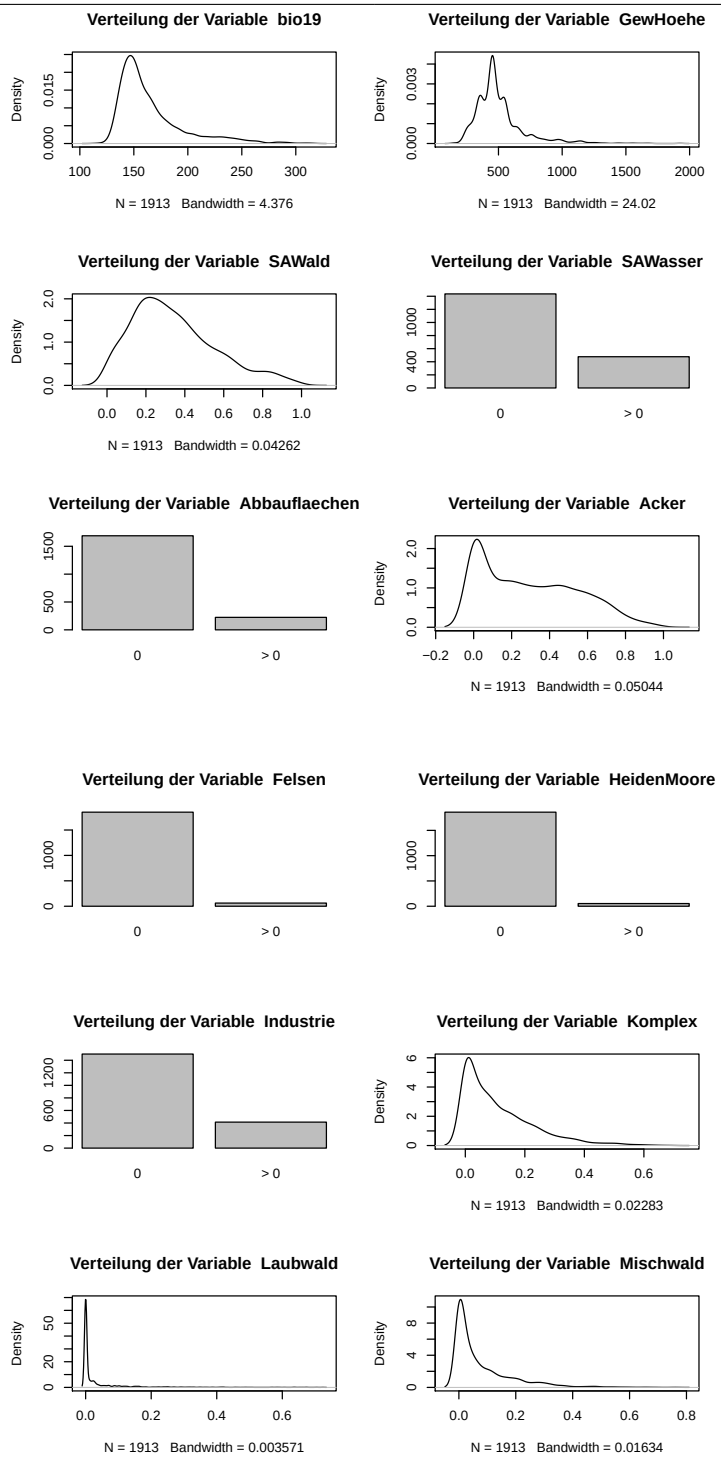

Verteilung der Variable bio14
Density
0.00
0.03
0.06
30
40
50
60
70
80
90
N = 1913 Bandwidth = 1.46

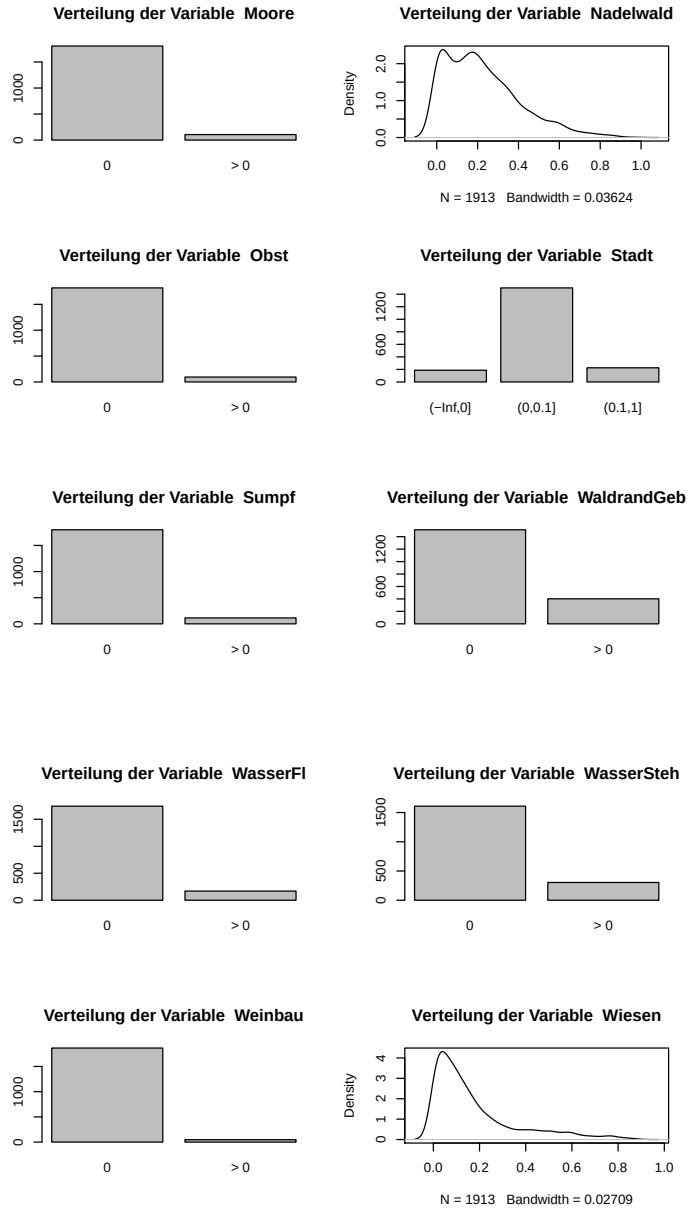

Verteilung der Variable bio15
Density
0.00
0.03
10
15
20
25
30
35
40
N = 1913 Bandwidth = 1.141


Verteilung der Variable bio16
Density
0.000
0.006
200
300
400
500
N = 1913 Bandwidth = 13.64


Verteilung der Variable bio17
Density
0.000
0.015
100
150
200
250
300
N = 1913 Bandwidth = 4.433


Verteilung der Variable bio18
Density
0.000
0.006
200
300
400
500
N = 1913 Bandwidth = 13.63


Verteilung der Variable bio19
Density
0.000
0.015
100
150
200
250
300
N = 1913 Bandwidth = 4.376
Verteilung der Variable GewHoehe
Density
0.000
0.003
500
1000
1500
2000
N = 1913 Bandwidth = 24.02
Verteilung der Variable SAWald
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
N = 1913 Bandwidth = 0.04262
Verteilung der Variable SAWasser
0
400
1000
0
> 0
Verteilung der Variable Abbauflaechen
0
500
1500
0
> 0
Verteilung der Variable Acker
Density
0.0
1.0
2.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
N = 1913 Bandwidth = 0.05044
Verteilung der Variable Felsen
0
1000
0
> 0
Verteilung der Variable HeidenMoore
0
1000
0
> 0
Verteilung der Variable Industrie
0
600
1200
0
> 0
Verteilung der Variable Komplex
Density
0
2
4
6
0.0
0.2
0.4
0.6
N = 1913 Bandwidth = 0.02283
Verteilung der Variable Laubwald
Density
0
20
50
0.0
0.2
0.4
0.6
N = 1913 Bandwidth = 0.003571
Verteilung der Variable Mischwald
Density
0
4
8
0.0
0.2
0.4
0.6
0.8
N = 1913 Bandwidth = 0.01634


Verteilung der Variable Moore
0
1000
0
> 0
Verteilung der Variable Nadelwald
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
N = 1913 Bandwidth = 0.03624
Verteilung der Variable Obst
0
1000
0
> 0
Verteilung der Variable Stadt
0
600
1200
(−Inf,0]
(0,0.1]
(0.1,1]
Verteilung der Variable Sumpf
0
1000
0
> 0
Verteilung der Variable WaldrandGeb
0
600
1200
0
> 0
Verteilung der Variable WasserFl
0
500
1500
0
> 0
Verteilung der Variable WasserSteh
0
500
1500
0
> 0
Verteilung der Variable Weinbau
0
1000
0
> 0
Verteilung der Variable Wiesen
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
N = 1913 Bandwidth = 0.02709

A.2 Inhalt der CD

Alle Berechnungen und Modellanpassungen für diese Bachelorarbeit wurden durchgeführt mit der R-Version 2.10.1 (R Development Core Team, 2009) und dem Packet „**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`.

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

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

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, Araújo 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.

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

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/`.

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`.

Schlumprecht H, Waeber G (2003). *Heuschrecken in Bayern*. Stuttgart (Hohenheim): Ulmer.

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 27. August 2010

(Katharina Zink)