Hiermit versichere ich, dass ich die eingereichte Bachelorarbeit selbstständig verfasst und ausschließlich die angegebenen Quellen und Hilfsmittel benutzt habe.

München, den 18.09.09

(Gisela Schmidberger)

Einflussgrößen für Artenzahlen von Brutvögeln in Bayern

Bachelorthesis

von Gisela Schmidberger

Institut für Statistik

Ludwig-Maximilians-Universität München

Betreuer:
Prof. Dr. T. Hothorn
Dipl. Stat. Nikolay Robinzonov

Inhaltsverzeichnis

Kapitel 1

Einführung

1.1 Einleitung

Auf Vögel kann man in Bayern jeden Tag treffen. Man findet sie sowohl in Wäldern, als auch an Seen, auf hohen Bergen oder im Flachland, im Garten und sogar in Städten. Einige Arten wie beispielsweise die Kohlmeise *(Parus [m.] major)*, Amsel *(Turdus merula)* oder Stockente *(Anas [p.] platyrhynchos)* sind einfach zu entdecken, andere leben im Verborgenen oder nur vereinzelt in ganz bestimmten Lebensräumen (z.B. Graugans *(Anser anser)*, Nachtigall *(Luscinia [luscinia] megarhynchos)* oder Wanderfalke *(Falco [p.] peregrinus)*).

Für die Beobachtung von Vögeln in Bayern interessieren sich die Menschen bereits seit vielen Jahren. So findet man eine erste Übersicht in der „Fauna boica“ von Franz v. Paula Schrank (1747 - 1835). Die erste umfassende Arbeit über die Brutvögel Bayerns erstellte jedoch der Pfarrer Andreas Johannes Jäckel (1822 - 1885). Sein Werk „Systematische Übersicht der Vögel Bayerns“ erschien 1891.

Die landschaftliche Vielfalt ist in Bayern sehr groß. Allein durch die Höhenamplitude reicht diese von Flussniederungen bis hin zum Hochgebirge und kann damit vielen unterschiedlichen Vogelarten optimale Brut und Lebensbedingungen bieten. (vgl. [Bezzel, 2005])

1.2 Problemstellung

Für die vorliegende Arbeit war es von Interesse Einflüsse auf die Artenzahlen von Brutvögeln in Bayern zu untersuchen. Die Hauptschwierigkeit lag dabei im Umgang mit dem dafür zur Verfügung stehenden hochdimensionalen Datensatz, der neben der Zielvariable „Artenanzahl“ noch 52 Umweltfaktoren, sowie räumliche Variablen in Form von Koordinaten enthält.

Mit Hilfe von passenden statistischen Methoden wurden aus der Vielzahl der zur Verfügung stehenden Kovariablen diejenigen herausgearbeitet, die gemeinsam den Einfluss auf die Zielgröße „Artenanzahl“ möglichst gut erklären können.

Eine interessante Frage war dabei auch, ob es einen räumlichen Einfluss gibt. Deshalb wurden als statistische Methodik Geoadditive Regressionsmodelle gewählt, die mit Hilfe eines Boosting-Algorithmus untersucht wurden.

Die Auswertung des Datensatzes wurde mit dem Statistikprogramm R durchgeführt ([R Development Core Team]) und erfolgte hauptsächlich mit Funktionen aus dem Paket **mboost** ([Hothorn, 2009]).

In Kapitel 1.3 und 1.4 werden im Folgenden der vorliegende Datensatz und die entsprechenden Variablen näher erläutert. Das zweite Kapitel beinhaltet die Methodik zu Geoadditiven Regressionsmodellen (vgl. [Fahrmeir, 2007]) und für den verwendeten Boosting-Algorithmus (vgl. [Bühlmann, 2007] und [Kneib, 2007]). Abschließend werden in Kapitel 3 die Ergebnisse zur Auswer-

tung des Datensatzes vorgestellt und näher erläutert. Kapitel 4 enthält einige abschließende Bemerkungen.

1.3 Datensatz

Die Daten des vorliegenden Datensatzes stammen aus den drei unterschiedlichen Quellen „CORINE“, „FRAGSTATS“ und „WORLDCLIM“.

CORINE:

CORINE (Coordinated Information on the European Environment) ist ein Programm der Europäischen Union. Die Daten werden mit Hilfe von Satellitenbildern gesammelt und beziehen sich hauptsächlich auf die Bodenbedeckung bzw. Landnutzung in Europa.

FRAGSTATS:

FRAGSTATS ist ein Programm für kategoriale Landkarten, mit dem die räumlichen Muster der Landschaften analysiert werden können. Die Analyse erfolgt dabei auf drei Stufen. Die erste davon sind die sogenannten „patches“, sie erfassen die Landbedeckung bzw. Landnutzung einer bestimmten Einheit. Solche Einheiten sind beispielsweise Laub-, Misch- oder Nadelwald. Die zweite Stufe fasst dann die einzelnen Patches zu bestimmten Klassen zusammen, z.B. Wald(W), Acker oder Gewässer. Die letzte Stufe wird schließlich mit „landscape“ bezeichnet und betrifft somit die gesamte Landschaft(L) des untersuchten Gebiets. (vgl. [McGarigal, 2002])

WORLDCLIM:

Im Programm WORLDCLIM findet man Daten zu den klimatischen Bedingungen eines bestimmten Gebietes, wie beispielsweise Durchschnittstemperaturen oder Niederschlagsmengen. (vgl. [Hijmans])

1.4 Deskription

Im Anschluss werden die Variablen des Datensatzes näher beschrieben. Untersucht wurde dabei vor allem die Art der Variable (stetig, kategorial), die Verteilung der Daten und ob es eventuell auffällige Beobachtungen gibt. Zusätzlich zu dieser univariaten Analyse wurde auch noch eine Zusammenhangsanalyse durchgeführt, bei der jeweils in einem Streudiagramm die interessierende Variable „Artenanzahl“ gegen je eine der Kovariablen abgetragen wurde. Mit Hilfe der eingezeichneten Streudiagrammglätter (siehe Kapitel 2) konnte so ein erster Überblick über die Art (beispielsweise linear vs. nichtlinear) und die Stärke des Zusammenhangs gewonnen werden. Dabei ist aber zu beachten, dass die eigentlich interessierende Fragestellung der gemeinsame Einfluss der Kovariablen auf die Zielvariable ist. Die Ergebnisse der durchgeführten bivariaten Analyse können davon stärker abweichen und dienen deshalb nur dazu, sich im Vorfeld mit den Daten näher vertraut zu machen.

UQuad

Die Variable „UQuad“ gibt die Nummer des Gebiets an, in dem die Variablen beobachtet wurden. Insgesamt liegen Daten für n=1918 Beobachtungsgebiete vor.

1.4.1 Zielvariablen

Arten

Die interessierende Zielvariable des Datensatzes wird mit „Arten“ bezeichnet. Sie gibt die erfasste Artenanzahl von Brutvögeln im jeweiligen Untersuchungsgebiet an. Durchschnittlich konnten 73 Arten pro Gebiet gezählt werden. Die niedrigste erfasste Anzahl beträgt 13 Arten, die höchste Anzahl 132 Arten.

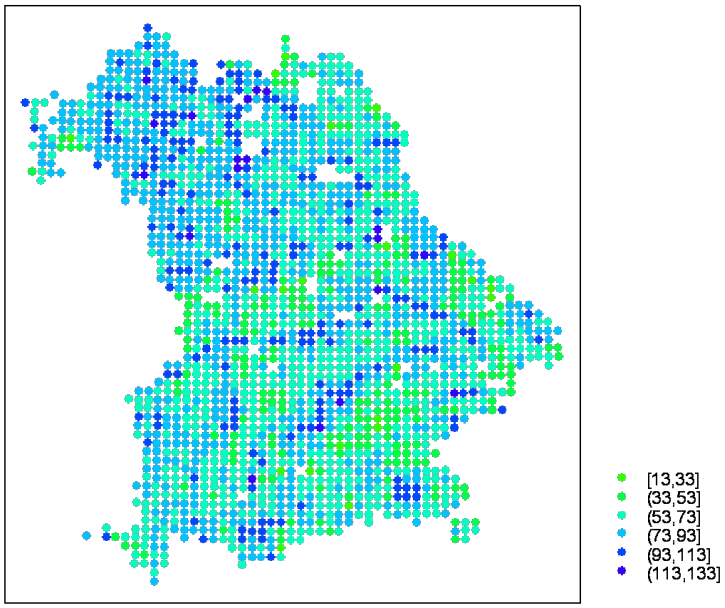

Abbildung 1.1: Artendichte der untersuchten Flächen

Wie man in Abb.(1.1) erkennen kann, können im Nordwesten Bayerns höhere Artenanzahlen in den Untersuchungsflächen beobachtet werden. Im Süden Bayerns sind deutlich die positiven Einflüsse von Donau, Isar und beispielsweise dem Chiemsee zu erkennen.

Die Verteilung der Variable „Arten“ folgt annähernd einer Normalverteilung

mit Mittelwert 73 und Standardabweichung 13.1, was auch mit Hilfe der Darstellung von Histogramm und QQ-Plot in (Abb. 1.2) gut zu sehen ist.

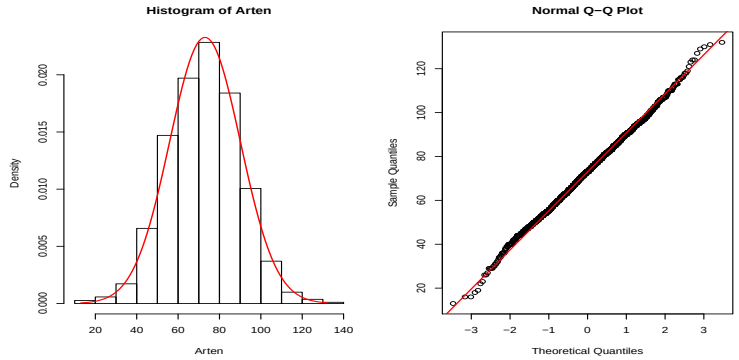

Abbildung 1.2: Histogramm und QQ-Plot für die Zielvariable Artenanzahl

RLArten, Waldvogelarten

Die Variablen „RLArten“ und „Waldvogelarten“ geben die Anzahl von Brutvogelarten für bestimmte Untergruppierungen wieder. Diese werden jedoch innerhalb dieser Arbeit nicht analysiert.

1.4.2 Räumliche Variablen

X, Y, XWorld, YWorld

Das Untersuchungsgebiet umfasst das gesamte Bundesland Bayern. Die räumlichen Daten X,Y bzw. XWorld, YWorld liegen in Form von Koordinaten des Gauß-Krüger-Koordinatensystems auf einem nahezu äquidistanten Raster vor. Da das Gauß-Krüger-Koordinatensystem auf die Längen- und Breitengrade zurückgreift, nehmen die Flächen in Nord-Südrichtung etwas zu. Dies wirkt sich allerdings innerhalb von Bayern nicht besonders stark aus.

1.4.3 Kovariablen aus CORINE

SAWald, Laubwald, Nadelwald, Mischwald, Waldrandgebiet

SAWald ist eine Variable zur Beschreibung der Waldabdeckung in Bayern. Sie fasst die Variablen Laubwald, Nadelwald, Mischwald und Waldrandgebiet zusammen. Die Verteilung der Daten ist linkssteil und liegt zwischen 0 und 100%. Das arithmetische Mittel konnte bei 0.34 beobachtet werden. Betrachtet man den Streudiagrammglätter scheint der Einfluss von SAWald auf die Artenanzahl fast linear und negativ zu sein. Je höher der Waldanteil also ist, desto weniger Arten werden im Durchschnitt beobachtet, wobei die Artenanzahlen für kleinere Waldanteile eher höher als der Mittelwert prognostiziert werden, für größere Anteile dagegen niedriger. Schaut man sich die Waldabdeckung aufgeteilt nach Laub-, Nadel- und Mischwald an, so sind die Werte entsprechend kleiner. Für den Laubwald liegt der Median sogar bei 0, Nadelwald erreicht einen Median von 19%. Ebenfalls ein leicht negativer Einfluss auf die interessierende Variable kann durch das Waldrandgebiet verzeichnet werden.

SAWasser, WasserSteh, WasserFl, Sumpf

Die Daten zur Wasserabdeckung im Untersuchungsgebiet finden sich in der Variable SAWasser wieder. In diese Variable gehen die Werte aller Gewässer mit Ufer ein, also sowohl fließende Gewässer, als auch stehende Gewässer, sowie Sümpfe. Es konnte hier zwar ein einzelner Wert von 0.92 festgestellt werden, allerdings finden sich die meisten Werte in einem Bereich nahe Null wieder, so dass nur höchstens ein Viertel der Daten einen Wert größer Null annimmt. Diese Werte finden sich größtenteils im Bereich bis 10%. Der Streudiagrammglätter steigt im Bereich zwischen 0 und 0.1 deutlich an und prognostiziert Artenanzahlen von über 90. Im weiteren Verlauf wird die Funktion

sehr unruhig, was an der kleinen Anzahl an vorliegenden Beobachtungen in diesem Wertebereich liegt. Somit ist eine Interpretation hier wohl weniger sinnvoll. Auch für die einzelnen Variablen „Sumpf“, „WasserSteh“ und „WasserFl“ werden die Einflüsse positiv geschätzt. Die vorhergesagten Artenzahlen erreichen sogar Werte über 100.

Abbauflaechen

Der Anteil an beobachteten Abbauflächen ist sehr klein. Das Maximum liegt hier nur bei 0.08. Der Effekt scheint positiv linear zu sein, allerdings kann der Scatterplot-Smoother hier eventuell nicht valide geschätzt werden, da nur 7 unterschiedliche Werte angegeben werden konnten.

Felsen

Ähnlich verhält es sich auch bei der Variable Felsen. Es werden zwar vereinzelt Beobachtungen bis zu 64% Felsenanteil gemacht, allerdings treten auch hier Werte größer 10% nur sehr vereinzelt auf und führen so zu einer niedrigen Schätzgenauigkeit. Lässt man sich die Glättungsfunktion trotzdem ausgeben wird der Effekt linear und negativ geschätzt, so dass für ca 10% mehr Felsenanteil ca. 5 Brutvögelarten weniger prognostiziert werden.

HeidenMoore, Moore

Für die ebenfalls stark linkssteil verteilte Variable „HeidenMoore“ wird, wie auch für „Moore“, der Effekt ebenfalls linear geschätzt. Allerdings scheinen sich Moore positiv auf die Artenanzahl auszuwirken. Das Problem der schwach besetzten rechten Seite des Wertebereichs tritt auch hier auf. Der größte Anteil an Mooren in einem Pixel des Rasters liegt bei 25%, für die zusammengesetzte Variable „HeidenMoore“ liegt der Wert etwas höher nämlich

bei 33%.

Komplex

Auch diese Daten sind linksteil verteilt, allerdings ist die Ausprägung Null nicht ganz so stark vertreten, so dass der Median bei 0.11 liegt und das Maximum bei fast 70%. Betrachtet man wieder den Scatterplot-Smoother, so scheinen bis zu Werten von ca. 20% überdurchschnittlich viele Arten vorhanden zu sein. Größere Werte von „Komplex“ wirken sich hingegen negativ auf die Artenanzahlen aus.

Obst, Weinbau

Ebenso haben die Variablen „Weinbau“ und „Obst“ nur einen relativ kleinen Anteil an Werten größer Null. Für den Obstplantagenanteil liegen eventuell ausreichend Beobachtungen zwischen 0 und 0.1 vor, so dass in diesem Bereich der Scatterplot-Smoother interpretiert werden kann. Dieser zeigt eine leichte Parabelform, die überhalb des Mittelwerts verläuft. Der größte positive Effekt kann also bei ca. 5% Obstplantagenanteil beobachtet werden. Auch hier kann es aber eventuell zu Fehlinterpretationen wegen der relativ geringen Anzahl an Beobachtungen kommen. Beim Weinbau scheint der Einfluss auf die Brutvogelarten ebenfalls positiv zu sein.

Stadt, Industrie, Verkehr

Die Variable „Stadt“ bezeichnet den Anteil an Städten und Dörfern im Untersuchungsgebiet. Der Mittelwert liegt hier bei einem Anteil von 0.05. Schaut man sich die Streudiagrammglätter für „Stadt“, sowie für die Variablen „Industrie“ und „Verkehr“ an, kann man erkennen, dass die bivariaten Schätzungen auch für diese Variable einen positiven Einfluss auf die Anzahl der

Vogelarten vorhersagen. Ist keine Stadt vorhanden, scheinen sogar nur unterdurchschnittlich viele Artenanzahlen erwartet werden.

Wiesen, Acker

Acker und Wiesen sind in den untersuchten Flächen jeweils relativ häufig zu finden, so dass es hier auch sehr viele Beobachtungen größer als Null gibt. Im Mittel verfügt jedes Untersuchungsgebiet über einen Anteil von 30% Acker und 17% Wiesen. Die Scatterplot-Smoother sollten demnach hier relativ gut den Einfluss schätzen. Allerdings scheinen Wiesen kaum einen Effekt auf die Anzahlen zu haben, so dass der Glätter hier als fast waagrechte Gerade beim Mittelwert von 73 verläuft. Bei der Variable „Acker“ ist der Verlauf ähnlich, jedoch wird hier eine leichte Parabelform geschätzt, so dass man von einem leicht positiven Effekt ausgehen kann.

1.4.4 Kovariablen aus FRAGSTATS

LNumberPatches, LPatchDensity

NumberPatches (NP) ist ein Maß, dass die Fragmentierung der untersuchten Landschaft misst.

$$NP = n_i, \qquad n_i \text{ bezeichnet die Anzahl der Patches der Klasse i}$$

Für die Variable „LNumberPatches“ konnten so in den einzelnen Pixeln des Untersuchungsrasters zwischen 2 und 57 Landbedeckungseinheiten erfasst werden. Die Variable ist annähernd normalverteilt mit einem Mittelwert von 26.

PatchDensity (PD) misst prinzipiell die gleiche Information mit dem Unterschied, dass sie den Wert proportional zur gesamten Landschaftsfläche

angibt.

$$PD = \frac{n_i}{A}(10000)(100), \qquad A \text{ ist die Gesamtfläche in } m^2$$

Die Einheit von „LPatchDensity ist damit Anzahl pro 100 Hektar. Handelt es sich bei den untersuchten Landschaften um äquidistante Rasterflächen, so wird also in beiden Variablen die gleiche Information gespeichert.

Die geschätzte Glättungsfunktion der zweidimensionalen Zusammenhangsanalyse liegt für niedrige Werte deutlich unterhalb des Mittelwertes, danach steigt sie an, flacht aber immer weiter ab.

LLargestPatchIndex

Der Largest Patch Index LPI wird berechnet mit Hilfe der Formel

$$LPI = \frac{max(a_{ij})}{A}, \qquad a_{ij} \text{ ist die Fläche } (m^2) \text{ des Patches ij, j=1,...,n}$$

Die Variable „LLargestPatchIndex“ gibt somit die prozentuale Fläche des größten Patches wieder. Sie ist also ein Maß für die Dominanz einzelner Gebiete in der Landschaft. Für die untersuchten Regionen in Bayern lag das Minimum bei 8.56%, das bedeutet, dass die untersuchte Fläche aus sehr vielen unterschiedlichen kleinen Teilstücken besteht. Das Maximum hingegen lag bei 98,5%. Diese Fläche wird demnach von einer einzelnen Landnutzungseinheit dominiert. Im Durchschnitt nimmt die größte zusammenhängende Nutzungsfläche ca. 40% der Gesamtfläche ein. Die geschätzte Funktion dieser Variable ist linear und negativ. Für kleine Werte liegen die Artenanzahlen bei fast 80, je größer die Werte werden, desto kleiner ist die Artenanzahl.

LTotalEdge, LEdgeDensity

Das Maß TE (Total Edge) erfasst die gesamte Länge der Randlinien von Patches einer bestimmten Klasse.

$$TE = \sum_{k=1}^{m} e_{ik},$$

e_{ik} ist die Länge der Grenzen eines einzelnen Patches in Metern

Für die Variable „LTotalEdge“, die dieses Maß auf die gesamte Landschaft anwendet, wurden somit die Grenzen aller vorhandenen Patches aufsummiert. Dabei ergaben sich Längen zwischen 15,2 km und 188,4 km. Betrachtet man die Ergebnisse der bivariaten Analyse scheinen die Artenanzahlen in der ersten Hälfte des Wertebereichs leicht anzusteigen, dann aber eher konstant auszulaufen.

Edge Density bezieht, wie auch schon vorher bei Patch Density beschrieben wurde, wieder die Gesamtfläche mit in die Analyse ein.

$$ED = \frac{\sum_{k=1}^{m} e_{ik}}{A}(10000)$$

LEdgeDensity misst demnach die Meter an Randlinien pro Hektar Untersuchungsgebiet. Auch hier bringt diese Variable wegen der fast gleich großen Untersuchungsgebiete kaum neue Erkenntnisse.

LLandscapeShapeIndex

Der Landscape Shape Index setzt den Umfang aller Patches der Klasse i ins Verhältnis zum kleinstmöglichen Umfang für eine Fläche derselben Größe.

$$LSI = \frac{e_i}{\min e_i}$$

Er gibt damit, ähnlich wie auch TotalEdge, einen Überblick über die Stärke der Zergliederung einer Landschaft. Der Index kann Werte ≥ 1 annehmen, wobei 1 bedeutet, dass keine Zergliederung vorliegt.

Der Index für die gesamte Landschaft nimmt Werte zwischen 1.16 und 8.01 an. Die Verteilung der Daten folgt annähernd einer Normalverteilung mit einem Mittelwert von 4.6.

CAWaldflache, WPLANDWaldanteil

„CAWaldflache“ summiert die Flächen aller Patches der Klasse Wald auf. CA misst also wie viel absolute Fläche (in Hektar) einer bestimmten Klasser (hier Wald) in der untersuchten Landschaft vorkommt.

$$CA = \sum_{j=1}^{n} a_{ij}(\frac{1}{10000})$$

Im vorliegenden Datensatz treten dabei Werte zwischen einem und 3386 Hektar Waldfläche auf. Das arithmetische Mittel liegt bei 1162 Hektar Wald pro untersuchtem Landschaftsabschnitt.

Auch hier existiert eine zweite ähnliche Variable, welche die gleichen Informationen proportional zur Untersuchungsfläche berechnet. Diese wurde als „WPLANDWaldanteil“ bezichnet.

$$PLAND = P_i = \frac{\sum_{j=1}^{n} a_{ij}}{A}(100)$$

Für den vorliegenden Datensatz bringt also auch diese Variable kaum neuen Erkenntnisse.

Des Weiteren wurden diese Informationen bereits über die Variable „SAWald“ aus dem CORINE-Datensatz aufgenommen.

WNumberPatches, WPatchDensity

Die Variablen „WNumberPatches“ bzw. „WPatchDensity“ werden genauso berechnet, wie „LNumberPatches“ bzw. „LPatchDensity“, beziehen sich jetzt allerdings nur noch auf die Klasse Wald.

Es liegen in den untersuchten Landschaftsgebieten zwischen einer und 20 nicht zusammenhängende Waldflächen vor. Der Mittelwert liegt bei 6.5. Der Einfluss auf die Responsevariable wird wieder nahezu linear und leicht negativ geschätzt. Für große Werte werden demnach nur noch etwas über 60 Arten erwartet.

WLargestPatchIndex

Die größten Waldflächen in den Untersuchungsgebieten nehmen einen Anteil zwischen 3 und 98 Prozent der Gesamtfläche ein. Die Verteilung der Variable ist linkssteil und der Mittelwert liegt bei etwa 25%. Auch hier scheint der Einfluss ab einem Wert von 40 negativ zu sein.

WEdgeDensity

Die Länge der Randlinien der Waldflächen proportional zur Fläche des Untersuchungsgebiets, wird in der Variable „WEdgeDensity“ gespeichert. Das berechnete Minimum liegt bei 0.12, das Maximum bei 38.39 Metern pro Hektar. Etwa die Hälfte der Beobachtungen nimmt Werte zwischen 10 und 20 Metern pro Hektar an.

WLandscapeShapeIndex

Der Landscape Shape Index (LSI) für die Klasse Wald erreicht Werte zwischen 1 und 9. 1 bedeutet dabei eine bestmögliche Aggregation der Waldflä-

chenanteile im Untersuchungsgebiet.

1.4.5 Kovariablen aus WORLDCLIM

bio1

Mit „bio1“ wird die Variable für die Daten zu den Jahresdurchschnittstemperaturen bezeichnet. Die Einheit ist dabei °C·10. Die erfassten Werte gehen von -1.97 bis 99.85. Dies bedeutet also, dass die wärmste untersuchte Region eine Jahresdurchschnittstemperatur von knapp 10°C aufweist. Im Mittel hat es in Bayern ca. 7.6°C. Die Verteilung ist rechtssteil, so dass sich die meisten Werte zwischen 6°C und 9°C befinden.

bio2

Für „bio2“ wird die durchschnittliche Tages-Spannweite berechnet, nämlich durch Subtraktion der durchschnittlichen minimalen Temperatur eines Monats von der maximalen Temperatur. Solche Abweichungen konnten für den vorliegenden Datensatz zwischen 7°C und 13°C beobachtet werden. Höhere Spannweiten scheinen dabei zu einer größeren Artenanzahl zu führen.

bio3

$$Isothermality = \frac{bio2}{bio7} \cdot 100$$

Die Werte für diese Einflussgröße liegen zwischen 29.08 und 34, während der Mittelwert bei 31.53 liegt. Die geraden Werte 30, 31, 32, 33 kommen dabei mit einer größeren Häufigkeit vor, was eventuell an Rundungsfehlern liegen könnte.

bio4

Bei dieser Größe wird die Saisonabhängigkeit der Temperatur betrachtet. Die Berechnung erfolgt mit Hilfe der Standardabweichung (Standardabweichung · 100). Die Werte liegen zwischen 5685 und 7218, wobei nur vereinzelte Werte kleiner als 6400 sind, so dass in diesem Bereich keine valide Schätzung des Einflusses vorgenommen werden kann. Betrachtet man den Glätter für höhere Werte, schwankt die Funktion relativ unruhig, um den Mittelwert.

bio5, bio6

„bio5“ gibt die maximale Temperatur des wärmsten Monats an. Die Beobachtungen liegen hier zwischen 12,6°C und 24,7°C. Entsprechend steht „bio6“ für die minimale Temperatur des kältesten Monats, wobei hier Werte von -12°C bis -2°C gemessen wurden. Beide Variablen sind rechtssteil verteilt, so dass also im linken Wertebereich deutlich weniger Beobachtungen auftreten.

bio7

Die Variable „bio7“ berechnet sich aus den soeben erläuterten Variablen „bio5“ und „bio6“ durch

$$bio7 = bio5 - bio6$$

und wird als Jahres-Temperaturspannweite bezeichnet. Die resultierenden Abstände nehmen Werte zwischen 227,9 und 302,5 an, wobei wieder auf die Einheit mit °C·10 geachtet werden muss.

bio8, bio9, bio10, bio11

Diese Einflussgrößen stehen für die Durchschnittstemperatur des jeweils nassesten (bio8), trockensten (bio9), wärmsten (bio10) und kältesten (bio11)

Quartals. Ein Quartal bezeichnet dabei eine Periode von 3 Monaten. Alle vier Variablen sind rechtssteil verteilt. Ebenfalls wurden für alle vier glatte Funktionen geschätzt, die den Effekt auf die Zielgröße beschreiben. Diese verläuft für „bio11“ sehr unruhig, schwanken allerdings ansonsten alle nur leicht um den Mittelwert von 73 Arten.

„bio8“ hat einen Wertebereich von 72.9 bis 182.6 und einen Mittelwert von 162.3. Im trockensten Quartal liegen die Temperaturen zwischen -7.3°C und 4.9°C („bio9“). Die Ausprägungen für „bio10“, also der Mittelwert für das wärmste Quartal erreicht ein Minimum von 7.3°C und ein Maximum von 18.3°C, während ein Minimum von -7.8°C und ein Maximum von 1.6°C den Wertebereich des kältesten Quartals beschreiben.

bio12

Die Jahresniederschlagsmenge in Bayern beträgt durchschnittlich 819,5 Milimeter. Dies entspricht 819,5 Liter pro Quadratmeter. Die erfassten Mengen in den untersuchten Rastergebieten gehen von 590,4 mm bis 1464,1 mm. Zwischen 700 und 900 steigt die geschätzte Funktion etwas an und verläuft dann konstant bei knapp unter 80 Arten.

bio13, bio14

Schaut man sich die Niederschlagsmengen für die extremen Variablen „nassester Monat“ (bio13) und „trockenster Monat“ (bio14) an, so treten hier Durchschnittswerte zwischen 70.8 und 170.7 bzw. für „bio14“ 31.4 und 85.8 auf. Die Angabe erfolgt auch hier in Litern pro Quadratmeter.

bio15

Die Saisonabhängigkeit der Niederschlagsmenge wird mit Hilfe des Variationskoeffizienten berechnet. Dieser setzt die Standardabweichung ins Verhältnis zum Mittelwert

$$v = \frac{s}{\bar{x}}$$

Die Werte liegen hier im Bereich von 14 bis 38, wobei sich zwischen ca. 25 und 35 ein positiver Effekt auf die Artenanzahl zeigt.

bio16

Ähnlich wie „bio13“ beschreibt die Variable „bio16“ die Niederschlagsmengen für das nasseste Quartal, bei einem Minimum von 190mm, einem Maximum von 500mm und einem arithmetischen Mittel von 288mm.

NN

Die letzte Variable des Datensatzes „NN“ liefert die Höhe über dem Meeresspiegel. Der niedrigste untersuchte Bereich in Bayern liegt dabei auf 152 m Höhe, das höchste Gebiet auf 1925 Metern. Im Mittel findet man sich auf ca. 500 Metern wieder. Dies entspricht beispielsweise in etwa dem Wert für München. Der geschätzte Streudiagrammglätter zeigt für kleine Werte von NN sehr hohe Artenzahlen von bis zu 100. Zwischen 500 und 1000 Metern Höhe schwanken die Werte leicht um den Mittelwert, während für Höhenlagen von über 1000 Metern deutlich weniger Artenzahlen vorhergesagt werden.

Kapitel 2

Methodik

2.1 Geoadditive Regressionsmodelle

Sollen die Eigenschaften einer Zielgröße y in Abhängigkeit einer oder mehrerer Kovariablen beschrieben werden, so handelt es sich allgemein um Problemstellungen der Regression. Je nachdem welche Art von Zielgröße vorliegt (z.B. stetig, binär oder Zählvariablen) und welche Typen von Kovariablen ins Modell aufgenommen werden sollen, stehen unterschiedliche Modellklassen zur Modellierung zur Verfügung.

Werden die Effekte linear ins Modell aufgenommen, findet man sich beispielsweise im Rahmen der parametrischen Regression in der Klasse der linearen Modelle, oder deren Erweiterung den generalisierten linearen Modellen wieder. Hierbei ist zu beachten, dass sich die Linearität aber nur auf die Modellparameter, also die Regressionskoeffizienten β_j bezieht, die Kovariablen jedoch auch nichtlinear ins Modell eingehen können. Somit ist es möglich, durch Transformation einer Kovariable oder Verwendung eines Polynoms, auch in parametrischen Regressionsmodellen nichtlineare Effekte zu model-

lieren.

Da dieses Vorgehen allerdings vor allem für eine größere Zahl von Einflussgrößen relativ aufwendig ist und man sich oft auch im Voraus nicht festlegen möchte, in welcher funktionalen Form die Kovariablen in das Modell aufgenommen werden sollen, gibt es im Zuge der nichtparametrischen Regression flexiblere Modellierungsmöglichkeiten, die zudem auch relativ automatisiert anwendbar sind.

2.1.1 Nichtparametrische Regression

Für die lineare Regression liegen die Daten in Form von Beobachtungen $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$ einer metrischen Responsevariable y und von k Kovariablen $x_1, \ldots, x_k$ vor.

Das klassische lineare Regressionsmodell kann dann dargestellt werden in der Form

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

Die Fehler $\varepsilon_1, \ldots, \varepsilon_n$ werden dabei als unabhängig und identisch verteilt (i.i.d.) mit

$$\mathbb{E}(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2$$

vorausgesetzt.

Da bei der Regression der Zusammenhang zwischen der Zielgröße y und den erklärenden Variablen nicht exakt als Funktion von $x_1, \ldots, x_k$ gegeben ist, sondern durch zufällige Störungen ε überlagert wird, kann y als Zufallsvariable interpretiert werden, deren Verteilung von den Kovariablen abhängt. Der geschätzte lineare Prädiktor

$$\eta_i^{lin} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

kann dann als Schätzung für den bedingten Erwartungswert von y bei gegebenen Kovariablen $x_1, \ldots, x_k$ angesehen werden und damit zur Prognose von y verwendet werden.

$$\mathbb{E}(y_i|x_{i1}, \ldots, x_{ik}) = \eta_i^{lin}$$

Oft liegen allerdings neben den Kovariablen $x_1, \ldots, x_k$, deren Einfluss auf die Zielvariable y durch einen linearen Prädiktor modelliert werden kann, weitere Daten $(z_{i1}, \ldots, z_{iq})$, $i = 1, \ldots, n$ vor, für deren Kovariablen $z_1, \ldots, z_q$ eine flexiblere nichtparametrische Modellierung gewünscht wird. Dazu kann man den linearen Prädiktor η^{lin} zu einem additiven Prädiktor erweitern

$$\eta_i = \eta_i^{add} = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin}$$

wobei $f(\cdot)$ eine glatte Funktion bezeichnet. Unterschiedliche Annahmen über die Funktion f führen dann zu verschiedenen Modellierungsmöglichkeiten. Ein solches Modell bezeichnet man als Additives Modell (AM).

Additive Modelle können zudem erweitert werden, z.B. durch Interaktionen oder räumliche Effekte. Für letztere liegen Beobachtungen s_i einer Lokationsvariable s vor, z.B. in Form von Koordinaten oder Regionen. Der Prädiktor im sogenannten Geoadditiven Modell wird dann erweitert zu

$$\eta_i = \eta_i^{add} + f_{geo}(s_i)$$

Für alle diese Modellansätze wird die Zielvariable als metrisch und (approximativ) normalverteilt vorausgesetzt. Ist diese Annahme nicht unbedingt gegeben, können die Modelle zu generalisierten Modellen verallgemeinert werden, indem der Prädiktor mit einer Linkfunktion h verknüpft wird.

$$\mathbb{E}(y_i|x_{i1}, \ldots, x_{ik}, z_{i1}, \ldots, z_{iq}) = \mu_i = h(\eta_i^{add})$$

Bezeichnet werden diese Modelle dann als generalisierte additive Modelle (GAM)

2.1.2 Univariate Glättung

Polynomsplines

Wie bereits erwähnt, können mit Hilfe von Polynomen auch in linearen Modellen nichtlineare Einflüsse modelliert werden. Allerdings ist die Anpassung oft trotz eines hohen Polynom-Grades nicht besonders gut. In der nichtparametrischen Regression wird dieser Ansatz erweitert, indem man nicht ein Polynom über den gesamten Definitionsbereich einer Kovariable modelliert, sondern mehrere kleine Polynomstücke zusammensetzt. Wird dabei ein bestimmter Grad an Glattheit vorausgesetzt, so erhält man ein Polynom-Spline vom Grad $l \geq 0$, das an den Grenzen $l-1$ mal stetig differenzierbar sein soll. Die Anzahl an Unterteilungen des Definitionsbereichs bezeichnet man auch als „innere Knoten“ $\kappa_1, \ldots, \kappa_m$.

Grundsätzlich gilt, dass die Glattheit des Polynom-Splines aus dem gewählten Grad l resultiert, wobei als Standard hier oft kubische Splines, also $l = 3$, gewählt werden, während man durch eine höhere Knotenanzahl eine flexiblere Schätzung erreicht.

B-Splines

Ein Polynom-Spline kann beispielsweise auch durch eine Linearkombination von $d = m + l - 1$ B-Spline Basisfunktionen dargestellt werden

$$f(z_i) = \sum_{j=1}^{d} \gamma_j B_j(z_i).$$

Die B_j, $j = 1, \ldots, d$ bezeichnen die Basisfunktionen, während die γ_j die einzelnen Parameterschätzer sind.

Die Datenanpassung mit Hilfe von B-Splines wird in Abb. 2.1 dargestellt. Als erstes wird eine vollständige B-Spline Basis zu einer vorgegebenen Kno-

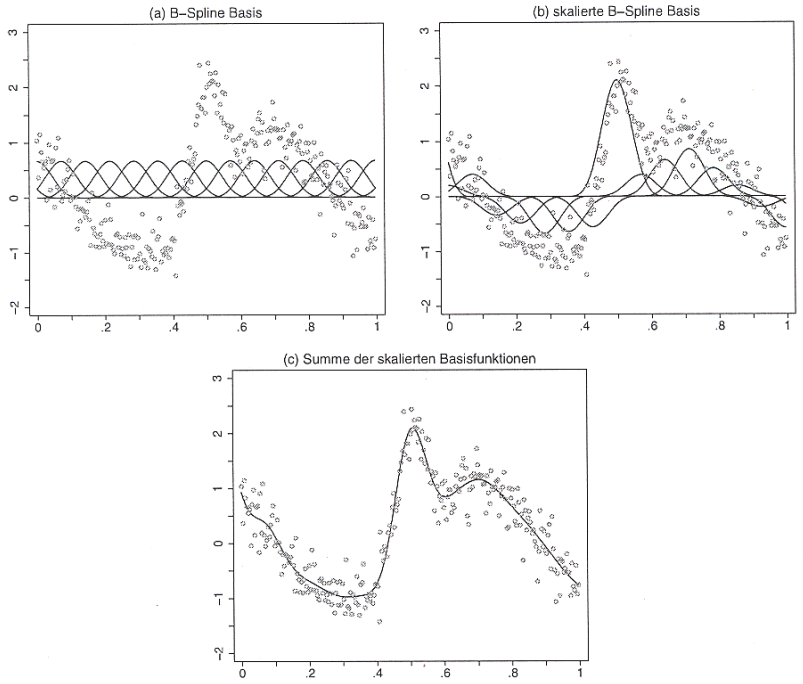

Abbildung 2.1: Univariate Glättung mit B-Splines (Quelle: Abb. 7.13 [Fahrmeir, 2007])

tenzahl berechnet und auf dem Definitionsbereich abgetragen (Abb. a)). Im zweiten Schritt wird dann mit Hilfe der KQ-Methode der Koeffizientenvektor $\hat{\gamma}$ geschätzt. Dieser liefert für jede Basisfunktion B_j eine zugehörige Amplitude $\hat{\gamma}_j$. Durch Addition der, mit den $\hat{\gamma}_j$ skalierten Basisfunktionen erhält man abschließend die Schätzung für den nichtparametrischen Effekt.

P-Splines

Da die Güte einer solchen Funktionsschätzung sehr stark von der Anzahl der verwendeten Knoten abhängt, ist es wichtig, diese beispielsweise mit Modellwahlstrategien bestmöglich zu bestimmen. Diese Methodik ist jedoch, vor

allem wenn viele Einflussgrößen vorliegen relativ aufwendig. Deshalb ist es oft günstig auf eine einfacher zu handhabende Methodik zurückzugreifen, nämlich auf penalisierte Splines. Dieser Ansatz wurde von Eilers und Marx beschrieben [Eilers, 1996].

Für solche P-Splines wählt man eine relativ große Anzahl an Knoten (üblicherweise 20 bis 40 Knoten), so dass die zu schätzende Funktion $f(z)$ auf jeden Fall ausreichend flexibel ist. Zusätzlich wird jedoch für die Schätzung noch ein Strafterm eingeführt, der eine zu große Variabilität bestraft. Es wird dann anstelle des üblichen KQ-Kriteriums ein penalisiertes KQ-Kriterium minimiert, um den Effekt zu schätzen.

$$PKQ(\lambda) = \underbrace{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{d}\gamma_j B_j(z_i)\right)^2}_{KQ(\lambda)} + \underbrace{\lambda\ pen(\gamma)}_{Strafterm}$$

Das als Glättungsparameter bezeichnete λ beeinflusst die Stärke der Bestrafung. Für den Strafterm wird häufig auch das Integral über die quadrierte zweite Ableitung von z verwendet $\lambda \int (f''(z))^2 dz$, beziehungsweise speziell für B-Splines Strafterme der Form $\sum_{j=k+1}^{d}(\Delta^k \gamma_j)^2$ wobei Δ^k die Differenzen k-ter Ordnung bezeichnen.

Ganz allgemein kann die Darstellung polynomialer Splines mit Hilfe von Basisfunktionen ($f(z) = \sum_{j=1}^{d} \hat{\gamma}_j B_j(z)$) als großes lineares Modell betrachtet werden

$$y = Z\gamma + \varepsilon$$

wobei die Designmatrix Z durch die Basisfunktionen definiert wird

$$Z = \begin{pmatrix} B_1^l(z_1) & \dots & B_d^l(z_1) \\ \vdots & & \vdots \\ B_1^l(z_n) & \dots & B_d^l(z_n) \end{pmatrix}$$

Dies ist der Grund, warum auch hier zur Schätzung der Regressionskoeffizienten die Methode der kleinsten Quadrate angewendet werden kann.

Wahl des Glättungsparameters

Zur Wahl eines geeigneten Glättungsparameters gibt es unterschiedliche Ansätze. Verwendet man die Bayesianische Betrachtungsweise, kann die Wahl von λ zum Beispiel auf MCMC-Verfahren (Markov-Chain-Monte-Carlo Simulationsverfahren) basieren. Auch mit Hilfe von ML und REML (Restricted Maximum Likelihood) lässt sich der Parameter, bei geeigneter Betrachtung der Penalisierungsansätze als gemischte Modelle, schätzen. In [Wood, 2006] wird für generalisierte additive Modelle die Minimierung des UBRE (Un-Biased Risk Estimator) beschrieben, wobei hier allerdings der Dispersionsparameter als bekannt vorausgesetzt wird.

Eine andere Möglichkeit, die nun näher beschrieben werden soll, ist die Wahl basierend auf Optimalitätskriterien, wie z.B. AIC oder Kreuzvalidierung. Da λ ja eine möglichst optimale Beziehung zwischen Varianz und Verzerrung sicherstellen soll, scheint die Betrachtung des MSE (mean squared error) gemittelt über die beobachteten Kovariablenausprägungen als angebracht, da er sich additiv aus quadrierter Verzerrung und Varianz zusammensetzt.

$$\frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}\left(\hat{f}(z_i) - f(z_i)\right)^2\right)$$

Nachdem sich die einfache Approximationsmöglichkeit durch die Residuenquadratsumme $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(z_i))^2$ als ungeeignet erweist, wird λ mit Hilfe des Vorhersagefehlers für neue Beobachtungen bestimmt. Stehen keine neuen Daten zur Verfügung, was üblicherweise der Fall ist, so kann man den Vorhersagefehler durch Kreuzvalidierung approximieren.

Man erhält das Kreuzvalidierungskriterium (CV), indem man jeweils zur Schätzung von λ eine Beobachtung (z_i, y_i) nicht berücksichtigt und dann mit Hilfe dieser Schätzung den Funktionswert $f(z_i)$ für die weggelassene Be-

obachtung prognostiziert.

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{(-i)}(z_i)\right)^2$$

wobei $\hat{f}^{(-i)}(z_i)$ die Funktionsschätzung bezeichnet, die sich durch Entfernen der Beobachtung (z_i, y_i) ergibt.

Der optimale Glättungsparameter kann dann durch n separate Modellanpassungen bestimmt werden. Für Penalisierungsansätze lässt sich allerdings zeigen, dass das Kreuzvalidierungskriterium auch in der Form

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(z_i)}{1 - s_{ii}}\right)^2$$

dargestellt werden kann. Dabei bezeichnet s_{ii} die n Diagonalelemente der Glättungsmatrix $\mathbf{S}$. Da die Berechnung der Diagonalelemente aber vor allem für große Datensätze mit einem hohen numerischen Aufwand verbunden ist, werden diese häufig durch ihren Mittelwert $\frac{1}{n}\sum_{i=1}^{n} s_{ii}$ ersetzt. Nachdem die Summe der Diagonalelemente gleich der Spur der Glättungsmatrix ist, erhält man also folgendes generalisiertes Kreuzvalidierungskriterium

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(z_i)}{1 - tr(S)/n}\right)^2$$

zur Bestimmung des Glättungsparameters λ.

2.1.3 Bivariate Glättung

Wie bereits in Kapitel 2.1.1 erwähnt, liegen in Geoadditiven Modellen zusätzliche Lokationsvariablen s_i vor. Für die Schätzung solcher räumlichen Effekte gibt es grundsätzlich die Unterscheidung zwischen stetigen und diskreten Informationen.

Räumliche Effekte diskreter Variablen

Liegen die Daten in Form von Regionenvariablen vor, so handelt es sich um diskrete Lokationsvariablen. Die Schätzung der glatten räumlichen Effekte kann dann beispielsweise mit Hilfe von Markov-Zufallsfeldern erfolgen. Koeffizienten benachbarter Regionen sollten sich dabei nicht zu stark voneinander unterscheiden. Demnach werden über das PKQ-Kriterium zu große Abweichungen zwischen Effekten in Nachbarschaften bestraft.

Räumliche Effekte stetiger Variablen

Basiert die Schätzung der räumlichen Effekte auf stetigen Lokationsvariablen, wie z.B. Koordinaten (z_1, z_2), so können sogenannte Tensorprodukt-P-Splines verwendet werden.

Zur Schätzung der zweidimensionalen Oberfläche $f(z_1, z_2)$ werden zuerst die eindimensionalen Basen gebildet

$$B_j^{(1)}(z_1),\ j = 1, \ldots, d_1 \quad \text{bzw.} \quad B_k^{(2)}(z_2),\ k = 1, \ldots, d_2.$$

Die Tensorprodukt Basen entstehen dann durch die Bildung aller paarweisen Produkte der univariaten Basisfunktionen

$$B_{jk}(z_1, z_2) = B_j^{(1)}(z_1) \cdot B_k^{(2)}(z_2), \quad j = 1, \ldots, d_1,\ k = 1, \ldots, d_2.$$

Die Funktion $f(z_1, z_2)$ kann damit folgendermaßen dargestellt werden

$$f(z_1, z_2) = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \gamma_{jk} B_{jk}(z_1, z_2)$$

Auch hier werden oft die B-Spline Basen verwendet, da sie numerische Vorteile besitzen.

Die Schätzung der bivariaten Glättungsansätze kann auch hier wieder im Rahmen von linearen Modellen erfolgen, da sich die Tensorprodukt-Ansätze ebenfalls in der Form eines großen linearen Modells darstellen lassen.

$$y = Z\gamma + \varepsilon$$

$z_i^\top = (B_{11}(z_{i1}, z_{i2}), \ldots, B_{d_1 1}(z_{i1}, z_{i2}), \ldots, B_{1d_2}(z_{i1}, z_{i2}), \ldots, B_{d_1 d_2}(z_{i1}, z_{i2}))$
$\gamma = (\gamma_{11}, \ldots, \gamma_{d_1 1}, \ldots, \gamma_{1d_2}, \ldots, \gamma_{d_1 d_2})^\top$

Ebenfalls wie im eindimensionalen Fall geht man bei der Bestimmung der optimalen Knotenanzahl vor. Um den entsprechenden Strafterm für den Penalisierungsansatz zu konstruieren, nutzt man die räumliche Anordnung der Basisfunktionen aus. Geeignete räumliche Nachbarschaften sind in Abbildung 2.2 dargestellt. Die schwarz eingezeichneten Punkte stellen dabei die Nachbarschaften 1. bzw. 2. Ordnung dar.

Abbildung 2.2: Räumliche Nachbarschaften 1. und 2. Ordnung (Quelle: Abb. 7.39 [Fahrmeir, 2007])

Die quadrierten Differenzen zu Koeffizienten benachbarter Basisfunktionen werden nun sowohl in z_1- als auch in z_2-Richtung berechnet und die univariaten Differenzenmatrizen als D_1 bzw. D_2 bezeichnet.

Die zeilenweisen Differenzen erster Ordnung lassen sich dann durch Anwendung von $I_{d_2} \otimes D_1$ auf den Koeffizientenvektor γ bestimmen. Somit ergeben

sich die zeilenweisen quadrierten Differenzen zu

$$\gamma^\top (I_{d_2} \otimes D_1)^\top (I_{d_2} \otimes D_1)\gamma = \sum_{k=1}^{d_2} \sum_{j=1}^{d_1} (\gamma_{jk} - \gamma_{j-1,k})^2$$

Analog dazu werden auch die spaltenweisen quadrierten Differenzen gebildet

$$\gamma^\top (D_2 \otimes I_{d_1})^\top (D_2 \otimes I_{d_1})\gamma = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} (\gamma_{jk} - \gamma_{j,k-1})^2$$

Der Strafterm $\lambda\gamma^\top K\gamma$ lässt sich damit berechnen zu

$$\lambda\gamma^\top K\gamma = \lambda\gamma^\top [I_{d_2} \otimes K_1 + K_2 \otimes I_{d_1}]\gamma$$

mit den univariaten Strafmatrizen $K_1 = D_1^\top D1$ und $K_2 = D_2^\top D_2$

Dieses Prinzip kann auch auf Matrizen höherer Ordnung angewendet werden

$$\lambda\gamma^\top K\gamma = \lambda\gamma^\top [I_{d_2} \otimes K_1^{(k_1)} + K_2^{(k_2)} \otimes I_{d_1}]\gamma$$

mit den univariaten Strafmatrizen $K_1^{(k_1)}$ und $K_2^{(k_2)}$ der Ordnungen k_1 und k_2.

So ergibt sich beispielsweise für $k_1 = k_2 = 2$ ein Strafterm basierend auf zweiten Differenzen, aus den acht nächsten Nachbarn entlang den Koordinatenachsen, wie es auf der rechten Seite der Abb. 2.2 dargestellt ist.

Anstatt der eben erläuterten B-Spline Basisfunktionen können beispielsweise auch radiale Basisfunktionen zur Schätzung einer glatten Oberfläche verwendet werden. Diese bestehen aus kreisförmigen Höhenlinien und sind jeweils genau einem Knoten zugeordnet. Eine dritte Möglichkeit, die aus der Geostatistik stammt, ist das sogenannte Kriging. Die Basisfunktionen basieren bei diesem Verfahren auf Korrelationsfunktionen $B_j(s) = \rho(s, s_{(j)})$.

2.2 Boosting-Algorithmus

2.2.1 Ziel

Ziel des Boosting-Algorithmus ist es Modelle für evtl. hochdimensionale Daten zu fitten, die eine möglichst gute Vorhersage des wahren Zusammenhangs liefern. Wobei hochdimensional hier bedeutet, dass eine große Anzahl an Kovariablen zur Beschreibung des Responses y vorliegen. Dabei ist der Boosting-Algorithmus eine sehr flexible Methodik zur Modellanpassung, da unterschiedlichste Arten von Effekten einbezogen werden können, beispielsweise auch räumliche Effekte, nichtparametrische Effekte von stetigen Variablen oder Interaktionen.

2.2.2 FGD-Algorithmus

Versucht man in Regressionsproblemen einen Schätzer zu finden, minimiert man üblicherweise den erwarteten Verlust. Die Verlustfunktion $\rho(y, f(X))$ ist ein Maß für die Güte der Anpassung des Modells an die Daten. In der Datensituation geht man vom erwarteten Verlust über zum empirischen Risiko $\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(X_i))$. Die Minimierung kann zum Beispiel mit Hilfe der KQ-Methode in der linearen Regression oder Minimierung der negativen Log-Likelihood für die generalisierte Regression erfolgen, oder allgemein mit Hilfe eines iterativen Verfahrens, wie z.B. der Gradienten-Abstiegs-Methode.

Auch Boosting kann als funktionale Gradienten-Abstiegs-Methode (functional gradient descent (FGD)) interpretiert werden, welche die Lösung für folgendes Optimierungsproblem sucht:

$$f^*(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}(\rho(y, f(X))$$

wobei $\rho(\cdot, \cdot)$ eine Verlustfunktion ist.

Der allgemeine FGD-Algorithmus lautet dann wie folgt (vgl. [Bühlmann, 2007]):

1. Initialisiere $\hat{f}^{[0]}(\cdot)$ mit einem Startwert. z.B.

   $$\hat{f}^{[0]}(\cdot) \equiv \arg\min_c \frac{1}{n}\sum_{i=1}^{n} \rho(y_i, c)$$

   oder

   $$\hat{f}^{[0]}(\cdot) \equiv 0.$$

   Setze m=0.

2. Erhöhe m um 1.
   Berechne den negativen Gradientenvektor an der Stelle $f = \hat{f}^{[m-1]}(X_i)$:

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

   Der negative Gradient entspricht der Schrittrichtung.

3. Passe $U_1, ..., U_n$ und $X_1, ..., X_n$ mit Hilfe der Basisprozedur (z.B. Regression) an. Dies führt dann zum Funktionsschätzer $\hat{g}^{[m]}(\cdot)$

   $$(X_i, U_i)_{i=1}^n \overset{BasisProzedur}{\rightarrow} \hat{g}^{[m]}(\cdot)$$

   $\hat{g}^{[m]}(\cdot)$ kann damit als Approximation an den negativen Gradientenvektor angesehen werden.

4. Aktualisiere $\hat{f}^{[m]}(\cdot) = \hat{f}^{[m-1]}(\cdot) + \nu \cdot \hat{g}^{[m]}(\cdot)$,
   wobei ν $(0 < \nu \leq 1)$ die Schrittlänge bezeichnet. Man geht also schrittweise in Richtung des Minimums.

5. Wiederhole die Schritte 2 bis 4 bis $m = m_{stop}$

Der Boosting-Algorithmus kann vom Anwender sehr flexibel gestaltet werden, da verschiedene Verlustfunktionen $\rho(\cdot, \cdot)$ definiert werden können. Im Paket `mboost` steht dafür die Funktion `boost_family()` bereit.

Speziell für die Regression wird dabei oft der quadrierte L_2-Verlust verwendet:

$$\rho_{L_2}(y, f) = \frac{1}{2}|y - f|^2$$

Diese Verlustfunktion führt zu dem bekannten Schätzer $f^*_{L_2}(x) = \mathbb{E}(Y|X = x)$ Die Skalierung mit $\frac{1}{2}$ bewirkt, dass der negative Gradientenvektor gleich den Residuen ist. Der entsprechende Boosting-Algorithmus wird dann L_2-Boosting genannt.

Alternative Verlustfunktionen für die Regression wären z.B. der L_1-Verlust

$$\rho_{L_1}(y, f) = |y - f|$$

oder die Huber-Verlustfunktion

$$\rho_{Huber}(y, f) = \begin{cases} |y - f|^2/2 & \text{falls } |y - f| \leq \delta \\ \delta|y - f| - \delta/2 & \text{falls } |y - f| > \delta \end{cases}$$

Wendet man also den L_2-Verlust auf den FGD-Algorithmus an, so erhält man den L_2-Boosting Algorithmus. Dieser wird in den folgenden Punkten angepasst:
Als Startwert wird im L_2-Boosting oft der Mittelwert verwendet

$$\hat{f}^{[0]}(\cdot) \equiv \bar{y}.$$

Des Weiteren wird der negative Gradientenvektor hier zum Residuenvektor. Es werden also im 2. Schritt jeder Iteration die Residuen

$$U_i = y_i - \hat{f}^{[m-1]}(X_i), \quad i = 1, \ldots, n$$

berechnet.

Baselearner

Jeder Boosting-Algorithmus erfordert die Wahl einer Basisprozedur. Diese konstruiert einen Funktionsschätzer $\hat{g}(\cdot)$ der auf den Daten $(X_1, U_1), \ldots, (X_n, U_n)$ basiert,

$$(X_i, U_i)_{i=1}^n \overset{BasisProzedur}{\rightarrow} \hat{g}(\cdot)$$

wobei $(U_1, \ldots, U_n)$ den aktuellen negativen Gradienten und $\hat{g}$ einen Funktionsschätzer bezeichnet, der in jedem Schritt den aktuellen negativen Gradienten schätzt. Durch die mehrmalige Anpassung von $\hat{g}$ wird der Vorhersagefehler verringert.

Als Basisprozedur kann z.B. ein Regressionsmodell (lineares Modell oder additives Modell) gewählt werden. Da sich der allgemeine Boosting Schätzer aus einer Summe von Basisprozedur Schätzungen zusammensetzt

$$\hat{f}^{[m]}(\cdot) = \nu \sum_{k=1}^{m} \hat{g}^{[k]}(\cdot)$$

werden die strukturellen Eigenschaften des Schätzers $\hat{f}^{[m]}(\cdot)$ durch die Eigenschaften der Basisprozedur erzeugt. Das bedeutet, falls die Basisprozedur z.B. ein lineares Modell ist, so stammt auch der entsprechende Modellschätzer $\hat{f}^{[m]}(\cdot)$ wieder aus dieser Modellklasse. Die konkrete Darstellung von zwei ausgewählten Basisfunktionen findet sich in Kapitel 2.2.4.

Eine der größten Schwierigkeiten ist es, Baselearner zu finden, die in ihrer Komplexität vergleichbar sind um z.B. Verzerrungen durch flexiblere Effekte zu vermeiden.

Bestimmung von ν und $\mathbf{m_{stop}}$

Zwei wichtige Parameter im Boosting Algorithmus sind ν und m_{stop}. Für sie müssen angemessene Werte definiert werden.

Die Schrittweite ν liegt im Intervall (0,1] und wird typischerweise möglichst klein gewählt. Als Standardwert wird oft $\nu = 0.1$ verwendet. Je kleiner ν ist, desto größer ist die Anzahl an Boosting Iterationen. Demnach erhöht sich

mit kleinerem ν die Genauigkeit der Schätzungen aber auch die benötigte Rechenzeit.

Der Boosting-Algorithmus wird nach einer vorgegebenen Anzahl an Iterationen gestoppt. Dies ist nötig, um eine Überanpassung an die Daten zu verhindern. Die Wahl von m_{stop} kann festgelegt werden durch Informationskriterien wie z.B. AIC, korrigiertes AIC, BIC, durch Kreuzvalidierung oder durch Bootstrap-Verfahren. Vor allem für große Datensätze ist Bootstrapping eine geeignete Alternative zum AIC-Kriterium, da die Berechnung der Hat-Matrix hier sehr aufwändig ist.

2.2.3 Variablenselektion und Modellwahl

Da nicht unbedingt immer alle Kovariablen einen Einfluss auf die Zielgröße y haben, möchte man mit Hilfe von Verfahren der Variablenselektion bestimmte Kovariablen auswählen, so dass nur die relevanten Kovariablen ins Modell aufgenommen werden. Ebenso muss man auch zwischen verschiedenen Modellvarianten für eine Kovariable wählen und damit z.B. entscheiden ob man eine Variable nur als linearen Effekt ins Modell aufnimmt oder als flexibleren glatten Effekt.

An dieser Stelle steht man oft vor dem Problem eines sogenannten Bias-Varianz-Trade Offs (vgl. [Fahrmeir, 2007]). Dies bedeutet, dass je komplexer das Modell gewählt wird, desto geringer ist die Verzerrung der Schätzung und desto größer ist gleichzeitig die Varianz. Umgekehrt weisen einfachere Modelle (z.B. die Wahl eines glatteren Schätzverfahrens mit einer kleineren Zahl effektiver Freiheitsgrade in der nichtparametrischen Regression) eine weniger variable aber stärker verzerrte Schätzung auf. Im Boosting wählt man die Basisprozedur typischerweise mit geringer Varianz und großem Bias. Der Bias kann dann durch zusätzliche Boosting-Iterationen noch verringert

werden.

Die Aufgaben der Variablenselektion und Modellwahl sind vor allem für die komplexen Geoadditiven Regressionsmodelle relativ schwierig zu bewältigen. Ein möglicher Lösungsansatz, der beide Aufgaben vereint ist das sogenannte komponentenweise Boosting, das im Folgenden näher erläutert wird.

2.2.4 Komponentenweises Boosting

Für lineare Modelle bzw. generalisierte lineare Modelle kann die Modellanpassung beispielsweise mit L_2-Boosting mit Hilfe von komponentenweisen linearen kleinsten Quadraten erfolgen. Es wird also folgende Basisprozedur angenommen:

$$\begin{aligned} \hat{g}(x) &= \hat{\beta}^{(\hat{\vartheta})} x^{(\hat{\vartheta})} \\ \hat{\beta}^{(j)} &= \sum_{i=1}^{n} X_i^{(j)} U_i / \sum_{i=1}^{n} \left(X_i^{(j)} \right)^2 \\ \hat{\vartheta} &= \arg\min_{1 \le j \le p} \sum_{i=1}^{n} \left(U_i - \hat{\beta}^{(j)} X_i^{(j)} \right)^2, \end{aligned}$$

wobei $c^{(j)}$ die j-te Komponente eines Vektors c bezeichnet. Damit wird die beste Variable in einem einfachen linearen Modell ausgewählt. D.h. also diejenige mit der kleinsten Residuenquadratsumme.

Wendet man diese Basisprozedur auf den L_2-Boosting-Algorithmus an, so wird also in jeder Iteration eine der Kovariablen ausgewählt. Die Funktion wird linear angepasst:

$$\hat{f}^{[m]}(x) = \hat{f}^{[m-1]}(x) + \nu \hat{\beta}^{(\hat{\vartheta}_m)} x^{(\hat{\vartheta}_m)}$$

$\hat{\vartheta}_m$ bezeichnet dabei den Index der ausgewählten Prädiktorvariable im m-ten Iterationsschritt.

Da nicht unbedingt in jeder Iteration eine andere Kovariable ausgewählt wird, und man den Algorithmus mit Hilfe von m_{stop} frühzeitig abbricht, erreicht man dadurch gleichzeitig eine Variablenselektion.

Analog dazu kann man für additive Modelle eine nichtparametrische Basisprozedur wählen. Für komponentenweise Glättungssplines nimmt man dazu folgende Basisprozedur an, die man durch Minimierung des PKQ-Kriteriums aus Kap. 2.1.2 erhält.

$$\begin{aligned}
\hat{g}(x) &= \hat{\gamma}^{(\hat{\vartheta})} z^{(\hat{\vartheta})} \\
\hat{\gamma}^{(j)} &= \arg\min_{\gamma(\cdot)} \sum_{i=1}^{n} \left(U_i - \gamma(Z_i^{(j)}) \right)^2 + \lambda \sum_{j=k+1}^{d} \left(\Delta^k \gamma_j \right)^2 \\
\hat{\vartheta} &= \arg\min_{1 \le j \le p} \sum_{i=1}^{n} \left(U_i - \hat{\gamma}^{(j)}(Z_i^{(j)}) \right)^2
\end{aligned}$$

Auch hier wird in jeder Iteration der beste Glättungsspline ausgewählt. Die Anzahl der Freiheitsgrade für die Glättungssplines wird im Voraus festgelegt. Man wählt hier oft df=4, da sie grundsätzlich „klein“ gewählt werden soll, was zu einer niedrigen Varianz aber großen Bias führt.

Möchte man nicht nur entscheiden ob eine Kovariable ins Modell aufgenommen werden soll oder nicht, sondern auch wie sie aufgenommen werden soll, so kann man für jede Modellierungsalternative einen separaten Baselearner spezifizieren. Eine Kovariable kann also beispielsweise sowohl als linearer Effekt, als auch als flexiblerer glatter Effekt zu Beginn ins Modell aufgenommen werden. Mit Hilfe von komponentenweisem Boosting wird dann zwischen den beiden Möglichkeiten selektiert. Somit kann man gleichzeitig Variablenselektion und Modellwahl betreiben.

Kapitel 3

Anwendung

3.1 Aufbereitung des Datensatzes

Bevor nun die statistische Analyse mit der Methodik aus Kapitel 2 durchgeführt wird, soll der Datensatz in geeigneter Weise aufbereitet werden. Dazu werden einige der Kovariablen aus dem Datensatz entfernt. Als erstes bietet sich dafür die Variable „SAWald“ an, da sie wie bereits in Kapitel 1 beschrieben wurde, die selbe Information wie „WPLANDWaldanteil“ enthält. Zusätzlich werden die drei Variablen „LNumberPatches“, „LTotalEdge“ und „CAWaldflaeche“ nicht in die Analyse mit einbezogen, weil sie aufgrund der nahezu gleichgroßen Untersuchungsgebiete kaum neue Erkenntnisse liefern können. Wegen hoher Korrelationen und aufgrund der schwierigen Interpretierbarkeit werden aus den FRAGSTATS-Daten auch noch die Variablen „LLargestPatchIndex“ bzw. „WLargestPatchIndex“ und die Variable „LLandscapeShapeIndex“ entfernt, da schon ähnliche Informationen einer Flächenanalyse über die Variablen „LPatchDensity“ bzw. „WPatchDensity“ eingehen, und eine Randlinienanalyse in den Variablen „LEdgeDensity“ bzw. „WEdgeDensity“ enthalten ist.

Aus dem Worldclim Datensatz stehen elf Variablen mit Temperaturangaben zur Verfügung. Nachdem zur Berechnung von „bio3“ die Variablen „bio2“, “bio5“, “bio6“ und „bio7“ verwendet werden, werden die Kovariablen zur Temperatur für weitere Analysen auf „bio1“, „bio3“ und „bio4“ beschränkt. Auch die fünf Variablen zu den Niederschlagsmengen werden auf zwei reduziert, nämlich auf „bio12“ und „bio15“. Somit stehen statt der ursprünglichen 52 Kovariablen nun noch 32 zur Auswertung zur Verfügung.

3.2 Modellwahl

In einem ersten Modell werden dem L_2-Boosting-Algorithmus also diese 32 Kovariablen übergeben. Nachdem alle Zeilen mit fehlenden Werten aus dem Datensatz entfernt wurden, stehen nun Daten für n=1869 Untersuchungsgebiete zur Verfügung. Als Baselearner wird die Funktion `bbs()` aus dem R-Paket `mboost` gewählt, also penalisierte Regressionsplines wie sie in Kapitel 2.1.2 beschrieben wurden. Da für eine valide Schätzung von glatten Effekten eine bestimmte Anzahl an unterschiedlichen Beobachtung notwendig ist, werden vorerst alle Kovariablen ausgeschlossen, die weniger als 60 verschiedene Werte haben.

Somit gehen in das erste Modell („Mod1_covar“) folgende Kovariablen ein: Acker, Komplex, Laubwald, Mischwald, Nadelwald, Wiesen, LPatchDensity, LEdgeDensity, WPLandWaldanteil, WEdgeDensity, bio1, bio3, bio4, bio12, bio15, NN.

Die univariaten glatten Effekte werden modelliert durch kubische Regressionssplines (d.h. Grad $l = 3$) und mit 20 inneren äquidistanten Knoten, um eine ausreichende Flexibilität der Splines zu gewährleisten. Für die Bestrafung einer eventuell zu rauhen Funktion werden Differenzen zweiter Ordnung eingesetzt. Die Anzahl der Freiheitsgrade wird mit $df = 4$ festgelegt. Dieser

Wert definiert die Komplexität der Basisprozedur und entspricht der Spur der Hat-Matrix. Prinzipiell sollte er relativ klein gewählt werden, muss aber bei Ansätzen mit P-Splines größer als die Ordnung der verwendeten Differenzen sein ([Hothorn, 2009]).

Anschließend wird zu diesen 16 Kovariablen noch ein räumlicher Effekt ins Modell mit einbezogen. Dieses Modell wurde mit „Mod1_covarspatial“ bezeichnet.

Für die Modellierung der bivariaten glatten Oberflächen, also für die räumlichen Effekte, wird die R-Funktion `bspatial()` verwendet. Auch hier wird die Anzahl an Freiheitsgraden mit $df = 4$ festgelegt um eine Vergleichbarkeit mit den univariaten Glättungssplines zu gewährleisten. Zusätzlich wird dieser Baselearner mit jeweils 6 Knoten in x- und y-Richtung definiert. Die Bestrafung beruht wieder auf Differenzen zweiter Ordnung, jedoch nun in beide Koordinatenrichtungen. Analog zu den univariaten Splines beträgt der Grad der Glättung $l = 3$.

Zum Vergleich wird noch ein drittes Modell gefittet, das nur den räumlichen Effekt als erklärende Variable beinhaltet. Dieses wurde „Mod_spatial“ genannt.

Mit Hilfe der so definierten Baselearner kann nun ein additives Modell angepasst werden. Die Funktion `gamboost()` verwendet dazu einen Boosting-Algorithmus der auf den komponentenweisen Glättungssplines beruht. Als Verlustfunktion wird der quadrierte L_2-Verlust gewählt.

Für den Boosting-Algorithmus müssen zusätzlich zu den Baselearnern noch Werte für ν und m_{stop} bestimmt werden. Diese werden hier vorerst mit $\nu = 0.1$ und $m_{stop} = 500$ für die Modelle „Mod1_covar“ und „Mod1_covarspatial“ festgelegt. Für das Modell „Mod_spatial“ wählt man m_{stop} etwas höher nämlich $m_{stop} = 2000$.

Ein optimaler Wert für m_{stop} wird anschließend mit Hilfe von Bootstrapping, und zwar mit 25 Bootstrap Stichproben, bestimmt. Die Funktion `cvrisk()` berechnet dafür das geschätzte empirische Risiko für die verschiedenen Werte des Parameters m_{stop}. Anschließend wird die Anzahl an Boosting Iterationen zur Modellanpassung verwendet, die den kleinsten Wert für das empirische Risiko liefert. Für das Modell „Mod1_covar“ liegt dieses Minimum bei 490 Iterationen, für das Modell „Mod1_covarspatial“ bei 498. Da der Wert für „Mod_spatial“ der zu Beginn festgelegten Anzahl von 2000 entspricht, könnte hier eventuell durch eine Erhöhung der Boosting-Iterationen noch eine geringfügige Verbesserung des Vorhersagefehlers erreicht werden (siehe Abb. 3.1).

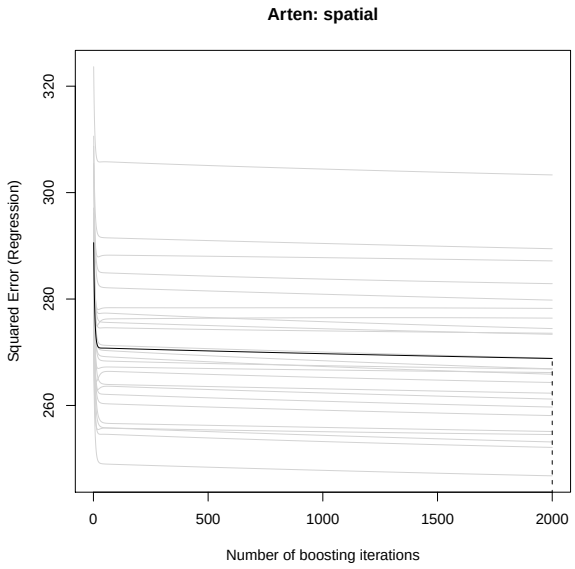

Abbildung 3.1: Bestimmung eines optimalen Wertes für den Parameter m_{stop}

Vergleicht man diese drei Modelle miteinander, so liegt das empirische Risiko für „Mod1_covarspatial“ im Mittel am niedrigsten. Wird die räumliche Komponente weggelassen so ist die Vorhersage für neue Daten etwas schlech-

ter. Es gibt also einen geringen räumlichen Effekt, der nicht durch die Kovariablen erklärt werden kann. Deutlich weniger gut schneidet das Modell „Mod_spatial“ ab. Der räumliche Effekt alleine kann also die Anzahl der Arten im Untersuchungsgebiet nicht gut genug wiedergeben. Der Vergleich der drei Modelle ist in Abb. 3.2 dargestellt.

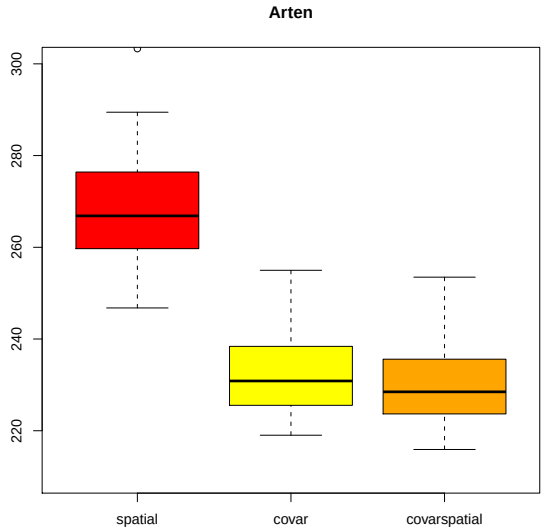

Abbildung 3.2: Vergleich der drei Modelle „Mod_spatial“, „Mod1_covar“ und „Mod1_covarspatial“ bezüglich des empirischen Risikos

Ergebnisse

Im Folgenden sollen die Ergebnisse des vorerst „besten“ Modells, also von „Mod1_covarspatial“, dargestellt und mit den Ergebnissen der beiden anderen Modelle verglichen werden. Durch die Variablenselektion konnte eine der Kovariablen („bio3“) identifiziert werden, keinen Effekt auf die Artenanzahlen zu haben. Drei weitere Variablen, nämlich „bio4“, „bio12“ und „bio15“ wurden zwar durch den Boosting-Algorithmus ausgewählt, zeigen aber nur

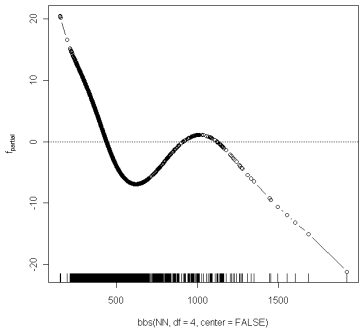

Abbildung 3.3: Effekt der Kovariable NN auf die Artenanzahlen

sehr geringe Effekte, und diese vor allem an Stellen für die nur sehr wenige Beobachtungen vorliegen.

Die eindeutig stärksten Effekte konnten bei der Kovariable „NN“ also der Höhe über dem Meeresspiegel beobachtet werden. Dieser Effekt ist in Abb. 3.3 dargestellt. Die Höhenlage des Untersuchungsgebiets wirkt sich also relativ stark auf die Anzahl an brütenden Vögeln aus, was sich auch schon in der bivariaten Zusammenhangsanalyse von Kapitel 1 gezeigt hat. Die Effekte reichen von ca. +15 bis -10, für sehr niedrige bzw. sehr hohe Lagen sogar bis +20 bzw. -20. Prinzipiell scheinen bei Höhen bis 400 m und zwischen 900 m und 1200 m eher mehr Brutvögel beobachtet zu werden als im Durchschnitt, ansonsten eher weniger. Dies kann eventuell dadurch erklärt werden, dass die verschiedenen Höhenlagen sehr unterschiedliche Naturräume bieten, die nicht für alle Vogelarten gleichermaßen geeignet sind. So findet man beispielsweise das Alpenschneehuhn (*Lagopus muta*) zur Brutzeit in steinigen alpinen Rasen bis zu einer Höhe von 2350 m, während sich z.B. der Raubwürger (*Lanius [e.] excubitor*) in Gebieten unter 1000 m zum Brüten aufhält [Bezzel, 2005].

Auch die Jahresdurchschnittstemperatur zeigt einen Effekt auf die Zahl der Vogelarten. So wirkt sich eine Durchschnittstemperatur von ca. 5°C bis 8°C

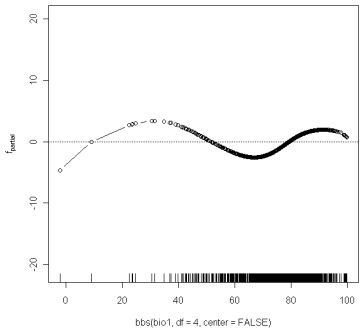

Abbildung 3.4: Effekt der Kovariablen bio1 (Jahresdurchschnittstemperatur)

eher negativ aus (Abb. 3.4) Für niedrigere sowie höhere Werte werden eher überdurchschnittlich viele Brutvögelarten erwartet.

Die geschätzten Effekte der Kovariablen Acker, Komplex, Mischwald und Nadelwald sind in Abb. 3.5 dargestellt. Für alle fünf Einflussgrößen liegen die Funktionen für kleine Werte im positiven Bereich und fallen dann mehr oder weniger stark ab. Eine mögliche Interpretation dafür ist, dass je größer der Bereich einer dieser Kovariablen im Untersuchungsgebiet ist, desto weniger andere Naturräume sind verfügbar. Demnach könnte die Anzahl an brütenden Vogelarten hier niedriger ausfallen, weil mit steigenden Anteilen einzelner Variablen die biologische Vielfalt kleiner wird. Besonders starke negative Effekte für hohe Werte der Variable treten bei „Acker“ und „Komplex“ auf.

Hingegen hat ein relativ großer Anteil an Laubwald oder Wiesen (bis ca. 50% bzw. 60% der untersuchten Fläche) eine eher positve Wirkung auf die Vogelzahlen (Abb. 3.6). Diese Flächen scheinen vielen Vogelarten also gute Nistmöglichkeiten, Schutz oder vielleicht auch eine ausreichende Nahrungsversorgung zu bieten. Zu den bodenbrütenden Wiesenvögeln gehören beispielsweise der Kiebitz (*Vanellus vanellus*) und der Wiesenpieper (*Anthus*

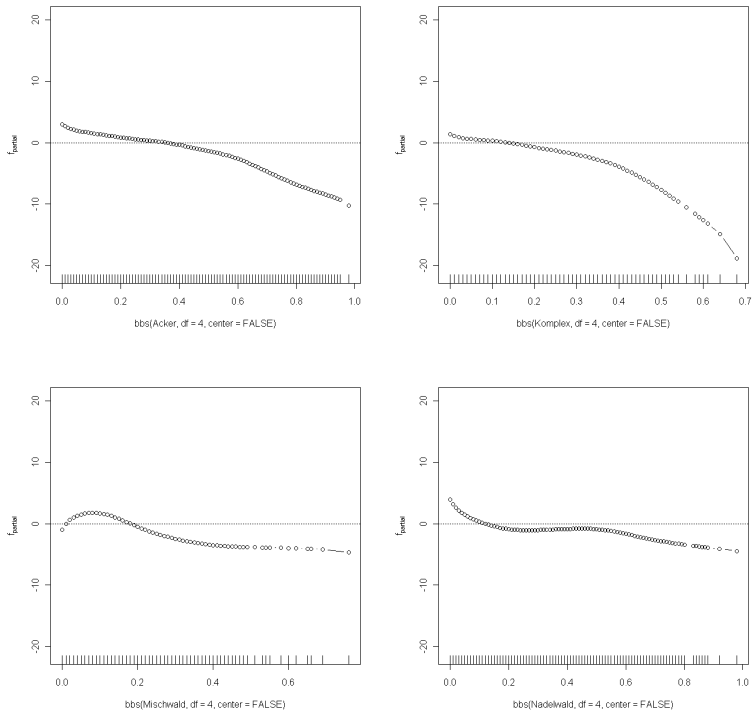

Abbildung 3.5: Effekt der Kovariablen Acker, Komplex, Mischwald und Nadelwald auf die Artenanzahlen

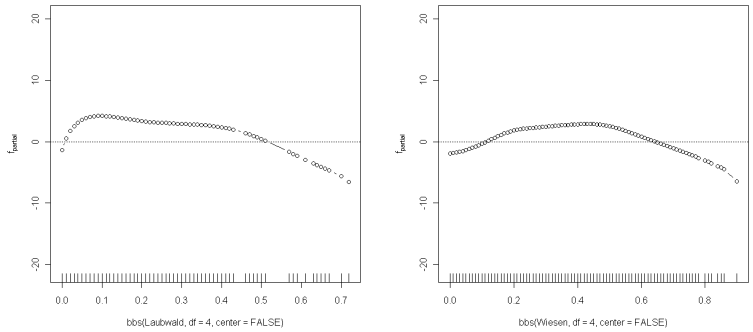

Abbildung 3.6: Effekt der Kovariablen Laubwald und Wiesen auf die Artenanzahlen

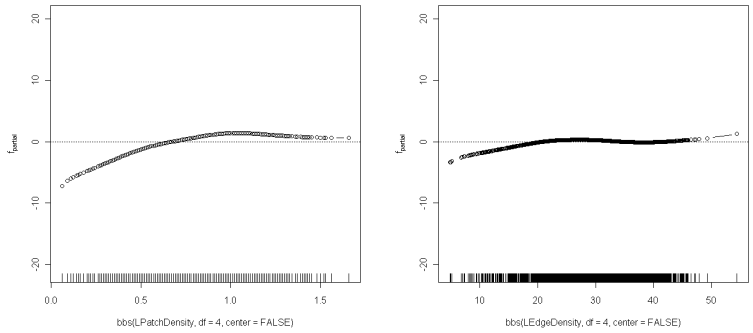

Abbildung 3.7: Effekt der Kovariablen LPatchDensity und LEdgeDensity auf die Artenanzahlen

pratensis). Im Laubwald sind hingegen die Sumpfmeise (*Parus palustris*) oder der Kleinspecht (*Dryobates minor*) zu Hause [Bezzel, 2005].

Auch wenn die Funktionen in Abb. 3.7 im Gegensatz zu den Funktionen in Abb. 3.5 steigend sind, könnten sie eventuell ähnlich interpretiert werden. Ein niedriger Wert bedeutet hier eine kleine Anzahl einzelner Landbedeckungseinheiten, während ein hohe Werte mehrere unterschiedliche Lebensräume zur Folge haben. Somit werden also hier für kleine Werte wenige Vogelarten vorhergesagt, für große Werte durchschnittlich bzw. sogar leicht überdurchschnittlich viele.

Schaut man sich im Gegensatz dazu nur die Klasse Wald an (Abb. 3.8), ist der Effekt wieder umgekehrt. Bis zu einer gewissen Anzahl an Waldflächen ist der Effekt positiv. Ist der Anteil von Wald im Untersuchungsgebiet jedoch sehr groß, so wirkt sich das eher negativ auf die Artenanzahlen aus.

Der geschätzte räumliche Effekt des Modells „Mod1_covarspatial“ ist auf der linken Seite der Abb. 3.9 dargestellt. Die Effekte sind relativ niedrig und schwanken größtenteils um Werte von zwei bis drei um Null. Nur in den „Ecken“ Bayerns treten etwas stärkere Effekte auf, so finden sich weiter

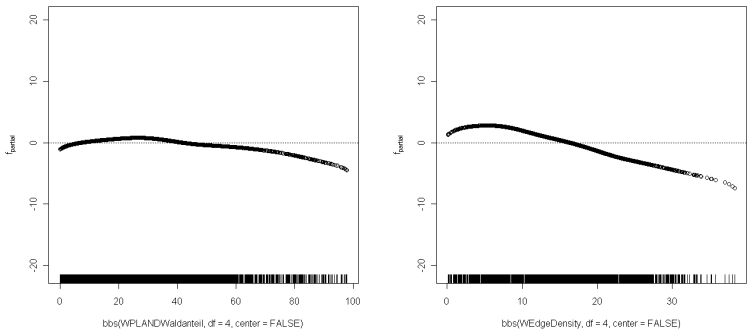

Abbildung 3.8: Effekt der Kovariablen LPatchDensity und LEdgeDensity auf die Artenanzahlen

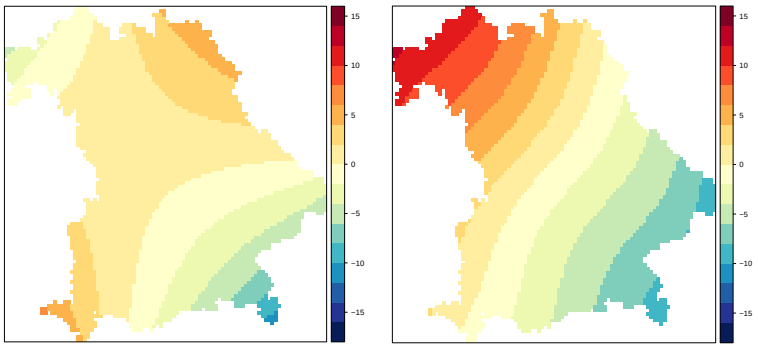

Abbildung 3.9: Vergleich der räumlichen Effekte (Mod1_covarspatial, Mod_spatial)

im Nordwesten sowie im Südosten deutlich weniger Vögel, im Nordosten und Südwesten hingegen sind die Effekte positiv. Zum Vergleich ist auf der rechten Seite der Abb. 3.9 der Effekt des Modells „Mod_spatial“ dargestellt. Hier kann man ein deutliches Gefälle des Effekts vom Nordwesten hin zum Südosten erkennen. Ein starker räumlicher Effekt kann ein Hinweis auf eine oder mehrere fehlende Kovariablen im Modell sein.

Vergleicht man die Effekte der einzelnen Kovariablen zwischen den beiden Modellen „Mod1_covarspatial“ und „Mod1_covar“, so kann man kaum Unterschiede erkennen. Die Aufnahme des räumlichen Effekts wirkt sich also auf keine der Kovariablen besonders stark aus. Nur bei einzelnen Funktionen, wie z.B. bei WEdgeDensity oder WPLANDWaldanteil ist die Ausprägung des Effekts im Modell „Mod1_covar“ etwas stärker. Die Form der geschätzten Funktionen bleibt aber im Prinzip für alle Kovariablen erhalten. Im Unterschied zu „Mod1_covarspatial“ wird die Variable „bio3“ allerdings durch die Variablenselektion nicht ausgeschlossen. Der gezeigte Effekt ist jedoch minimal.

Vergleich mit Modell2

Da in den deskriptiven Analysen bereits deutlich wurde, dass die Variable „SAWasser“ einen relativ starken Einfluss auf die Artenzahlen hat, soll in einem zweiten Modell, zusätzlich zu den Variablen aus den bisherigen Modellen, eine Dummyvariable „SAWasserD“ aufgenommen werden. Nachdem die Werte von „SAWasser“ relativ niedrig sind, (vgl. Kapitel 1.4.3) soll nicht auf die Höhe des vorhandenen Gewässeranteils eingegangen werden, sondern nur darauf, ob ein See, Fluss oder Sumpf vorhanden ist, oder nicht. Die Variable wurde also folgendermaßen definiert

$$SAWasserD = \begin{cases} 0 & \text{kein Gewässer mit Ufer vorhanden} \\ 1 & \text{Gewässer mit Ufer vorhanden} \end{cases}$$

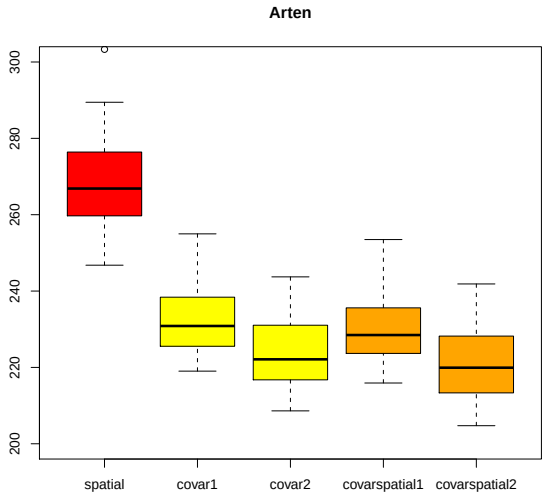

Abbildung 3.10: Vergleich aller Modelle bzgl. des empirischen Risikos

Zur Modellierung wird eine lineare Basisprozedur gewählt, die mit Hilfe der Funktion `bols()` in das Modell aufgenommen wird.

Das Modell ohne räumliche Komponente wird nun als „Mod2_covar“ bezeichnet, das Modell mit räumlicher Komponente als „Mod2_covarspatial“.

Vergleicht man die beiden Modelle wieder anhand des empirischen Risikos, so kann man erkennen, dass sich, analog zum ersten Modell, die Aufnahme des räumlichen Effekts positiv auf die Vorhersagegüte auswirkt. Betrachtet man die Modelle „Mod1_covarspatial“ und „Mod2_covarspatial“, so lässt sich eine deutliche Verbesserung des Vorhersagefehlers durch die Aufnahme der Variable „SAWasserD“ erkennen. Der Vergleich aller fünf Modelle ist in Abb. 3.10 dargestellt.

Die Ergebnisse des Modells „Mod2_covarspatial“ ähneln prinzipiell den Ergebnissen aus „Mod1_covarspatial“. Die Aufnahme der Dummyvariable hat also kaum Auswirkungen auf die Effekte der anderen Kovariablen. Nur der

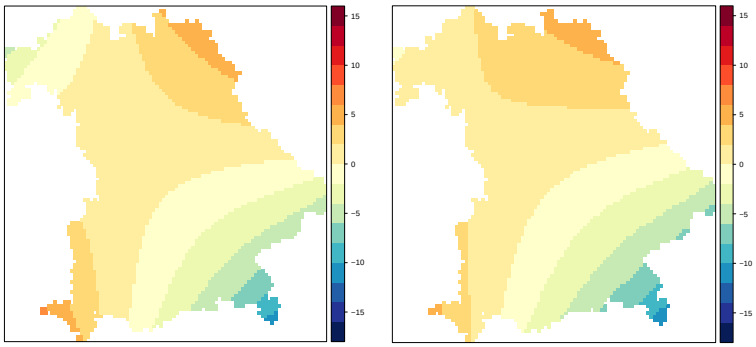

Abbildung 3.11: Vergleich der räumlichen Effekte (Mod1_covarspatial, Mod2_covarspatial)

räumliche Effekt verändert sich leicht, so dass der negative Einfluss des Nordwestens etwas zurückgeht. Den direkten Vergleich kann man in Abb. 3.11 sehen.

Das Vorhandensein von Gewässern wirkt sich positiv auf die Artenzahlen aus. So werden ca. 6 Arten mehr als der Durchschnitt prognostiziert, wenn es im Untersuchungsgebiet mindestens ein Gewässer gibt. Ist weder Fluss noch See oder Sumpf vorhanden, so liegen die Artenzahlen um 1.5 unterhalb des Durchschnitts. Für einige Vogelarten spielen also auch Gewässer eine wichtige Rolle für die Brut bzw. zur Aufzucht ihrer Jungen oder zur Nahrungssuche. Zu diesen Schwimmvögelarten gehören beispielsweise der Höckerschwan (*Cygnus olor*) oder der Kormoran (*Phalacrocorax [c.] carbo*) (vgl. [Bezzel, 2005]).

Kapitel 4

Zusammenfassung

In der vorliegenden Arbeit wurde der Einfluss verschiedener Umweltfaktoren auf die Brutvogelanzahl in Bayern untersucht. Nach einer inhaltlichen Vorauswahl geeigneter Kovariablen wurden dazu mit Hilfe des L_2-Boosting Algorithmus additive bzw. geoadditive Regressionsmodelle angepasst und dadurch automatisch geeignete Variablen selektiert.

Als Ergebnis kann festgehalten werden, dass der räumliche Effekt alleine für eine gute Modellierung nicht ausreicht. Wird er allerdings zusätzlich zu den restlichen Kovariablen ins Modell aufgenommen, so kann er zu einer geringfügigen Verringerung des Vorhersagefehlers beitragen.

Den stärksten Effekt auf die Artenzahlen scheint die Höhe über dem Meeresspiegel zu haben. Allgemein lassen sich mehr Brutvogelarten in den Gebieten vermuten, in denen eine große landschaftliche Vielfalt zur Verfügung steht.

Auch durch die Aufnahme der Dummyvariable „SAWasserD“ konnte eine Verbesserung der Vorhersage der Daten erreicht werden. Wobei auch hier in Gebieten mit Gewässer eine durchschnittlich größere Artenanzahl beobachtet werden konnte.

Literaturverzeichnis

[Bezzel, 2005] E.Bezzel, I.Geiersberger, G.v.Lossow und R.Pfeifer, 2005, *Brutvögel in Bayern: Verbreitung 1996 bis 1999* Stuttgart: Verlag Eugen Ulmer.

[Bühlmann, 2007] P.Bühlmann, T.Hothorn, 2007, *Boosting Algorithms: Regularization, Prediction and Model Fitting*: Statistical Science, 2007, Vol.22, No. 4, 477-505

[Eilers, 1996] P.H.C.Eilers, B.D.Marx, 1996, *Flexible smoothing with B-splines and penalties* Statistical Science, 11(2), 89-121

[Fahrmeir, 2007] L.Fahrmeir, T.Kneib, S.Lang, 2007, *Regression: Modelle, Methoden und Anwendungen*: 1. Auflage, Springer-Verlag Berlin Heidelberg

[Hijmans] R.Hijmans, S.Cameron, J.Parra, *WORLDCLIM*, University of California, Berkeley. www.worldclim.org

[Hothorn, 2009] T.Hothorn, P.Bühlmann, T.Kneib, M.Schmid, B.Hofner, *mboost-Package, Version 1.1-1*: http://cran.r-project.org

[Kneib, 2007] T.Kneib, T.Hothorn, G.Tutz, 2007, *Variable Selection and Model Choice in Geoadditive Regression Models*: Techni-

cal Report Number 003, 2007, Department of Statistics, University of Munich

[McGarigal, 2002] K. McGarigal, S.A.Cushman, M.C.Neel, und E.Ene, 2002, *FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps.* Computer software program produced by the authors at the University of Massachusetts, Amherst. www.umass.edu/landeco/research/fragstats/fragstats.html

[R Development Core Team] R Development Core Team, 2008, *R: A Language and Environment for Statistical Computing*: Vienna, Austria: R Foundation for Statistical Computing, 2008, http://www.r-project.org

[Wood, 2006] S.Wood, 2006, *Generalized Additive Models: An Introduction with R*: Chapman & Hall/CRC, Boca Raton

Anhang A

Elektronischer Anhang

Der elektronische Anhang enthält die zwei Ordner „analysis“ und „data“.

Im Ordner „data“ befinden sich die zur Verfügung gestellten Datensätze „bva.Rda“ und „XY.Rda“.

Der zweite Ordner „analysis“ enthält die R-Datei „deskriptive Analysen“. Dieser Programmcode wurde zur Auswertung der Deskriptiven Analysen von Kapitel 1 verwendet.

Die Auswertungsergebnisse von Kapitel 3 sind in der Datei „Modellwahl.R“ dokumentiert. Ein gespeicherter Workspace dieses R-Codes ist ebenfalls im Anhang abgelegt.

Zur Erstellung der Graphiken wurde die Funktion „Plot_Funktion.R“ verwendet, die mit Hilfe von „Modell_summary“ angewendet werden kann.