4

Variablenselektion

Björn Reineking[1,2] & Boris Schröder[3]

[1] UFZ Umweltforschungszentrum Leipzig-Halle GmbH, Sektion Ökosystemanalyse, Postfach 500136, 04301 Leipzig, Email: bjoern.reineking@ufz.de
[2] ETH Zürich, UNS, Haldenbachstr. 44, ETH-Zentrum HAD, CH-8092 Zürich
[3] Universität Potsdam, Institut für Geoökologie, Postfach 601553, 14415 Potsdam

4.1 Einleitung

Habitatmodelle quantifizieren die Beziehung von Vorkommen eines Organismus und Habitateigenschaften. Durch Fernerkundungstechnologie und geographische Informationssysteme (GIS) liegen heute oftmals viele Informationen über Habitateigenschaften vor. Angesichts der Fülle möglicher Prädiktorvariablen muss eine Auswahl getroffen werden. Dafür gibt es im wesentlichen zwei Gründe. Zum einen besteht die Gefahr der Überanpassung (*Overfitting*) des Modells an die Daten, wenn der großen Zahl von Prädiktorvariablen nicht hinreichend viele Vorkommens- und Nichtvorkommensaufnahmen gegenüber stehen (s. auch Schröder & Reineking 2004a,b, in diesem Band). Ein überangepasstes Modell macht schlechte Vorhersagen, und Überanpassung ist der häufigste Grund für unzuverlässige Modelle (Harrell 2001). Der zweite Grund, Variablen auszuwählen, ist, dass Modelle mit sehr vielen Variablen schwierig zu interpretieren sind. Ein Modell mit 20 oder mehr Variablen ist letztlich kaum mehr nachzuvollziehen, und Variablenselektion dient in diesem Fall dem Ziel, sich auf das Wesentliche zu konzentrieren.

Diese beiden Gründe dominieren in der Regel die Vorteile eines Modells, das alle verfügbaren Variablen enthält – die hohe Klassifikationsleistung auf den Trainingsdaten, die unverzerrte Abschätzung der Effektgrößen für alle verfügbaren Variablen und das Vermeiden der „Qual der Wahl".

Dieser Beitrag ist folgendermaßen aufgebaut: Zunächst diskutieren wir den *Bias-Variance-Tradeoff*, der dem Problem der Überanpassung zugrunde liegt. Ausgehend von dieser theoretischen Grundlage stellen wir die Grundidee und einige der wichtigsten Ansätze zur Variablenselektion vor. Im Anschluss daran werden einige Probleme der Variablenselektion aufgeführt sowie Alternativen und Ergänzungen aufgezeigt. Abschließend geben wir einige Empfehlungen für sichere und effektive Variablenselektion.

4.2 Bias-Variance-Tradeoff

Der Vorhersagefehler eines Modells kann in drei Anteile zerlegt werden: den irreduziblen Fehler, die Verzerrung (*bias*) und die Varianz (Hastie et al. 2001).

Der irreduzible Fehler ist durch die unvermeidbare Stochastizität des Systems bedingt. Die Umweltbedingungen legen nicht vollständig fest, ob ein Organismus an einer bestimmten Stelle vorkommt oder nicht, sondern beeinflussen lediglich seine Vorkommenswahrscheinlichkeit. Selbst wenn die Umweltbedingungen ideal sind, werden wir in der Regel in einem Teil der Fälle den Organismus nicht antreffen. Wenn die Vorkommenswahrscheinlichkeit beispielsweise 90% beträgt, dann ist unsere beste Vorhersage, dass der Organismus immer vorkommt, und wir werden in 10% der Fälle eine falsche Vorhersage treffen.

Dormann CF, Blaschke T, Lausch A, Schröder B, Söndgerath D (Hrsg.)(2004) Habitatmodelle – Methodik, Anwendung, Nutzen. Tagungsband zum Workshop vom 8.-10. Oktober 2003 am UFZ Leipzig. UFZ-Berichte 9/2004.

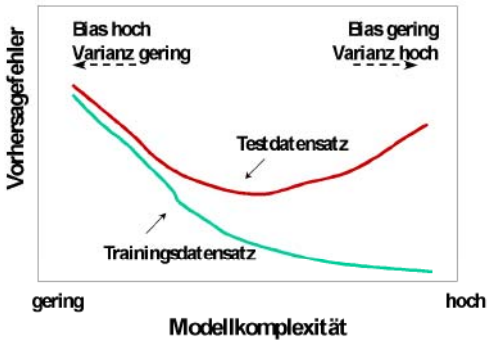


Abb. 4.1. Veranschaulichung des *Bias-Variance-Tradeoff* (verändert nach Hastie et al. 2001). Mit zunehmender Modellkomplexität (z.B. durch zunehmende Anzahl Prädiktorvariablen) sinkt die Fehlerrate auf den Daten, an die das Modell angepasst wurde. Auf Testdaten ist der Fehler zum einen höher als auf den Trainingsdaten, zum anderen nimmt der Optimismus, d.h. die Differenz zwischen den Fehlerraten auf Trainings- und Testdaten, mit steigender Modellkomplexität zu.

Die Verzerrung ist der systematische Fehler des Modells, d.h. der systematische Anteil an der Abweichung der vorhergesagten Vorkommenswahrscheinlichkeiten von den wahren Vorkommenswahrscheinlichkeiten (s. auch den Beitrag zu Gütemaßen von Reineking & Schröder 2004, in diesem Band). Die Ursache ist, dass das Modell die zugrundeliegende Verteilung nicht abbilden kann, weil entweder relevante Umweltvariablen nicht berücksichtigt wurden oder weil der Zusammenhang zwischen den Ausprägungen der Umweltvariablen und der Vorkommenswahrscheinlichkeit nicht richtig widergespiegelt wird.

Die Varianz eines Modells ergibt sich aus der Tatsache, dass die Beobachtungen, die uns zur Verfügung stehen, um die Modellparameter zu schätzen, lediglich eine Stichprobe aus der Grundgesamtheit ist. Wenn wir zu einem anderen Zeitpunkt die Aufnahmen wiederholten, ergäbe sich ein etwas anderes Muster der Vorkommen und Nichtvorkommen. Entsprechend fielen die Parameterschätzwerte und die Vorhersagen etwas anders aus. Je flexibler das Modell ist, desto stärker werden die spezifischen Besonderheiten der verschiedenen Stichproben auf die Parameterschätzungen und Modellvorhersagen durchschlagen.

Je komplexer ein Modell ist, d.h. je mehr Umweltvariablen berücksichtigt werden und je flexibler die Zusammenhänge zwischen Umweltvariablen und Vorkommen beschrieben werden, desto geringer ist die systematische Abweichung des Modells, aber umso größer ist seine Varianz. Das flexiblere Modell wird eine geringere Fehlerrate auf den Daten zeigen, an die es angepasst wurde, aber nicht notwendigerweise auf neuen Daten (s. Abb. 4.1).

Wenn man vor der Wahl zwischen einem komplexeren und einem einfacheren Modell steht, hängt es entscheidend von den zur Verfügung stehenden Daten ab, ob der Modellfehler von der Verzerrung oder von der Varianz dominiert wird (vgl. Abb. 5.2 in Schröder & Reineking 2004a, in diesem Band), und ob entsprechend das komplexere oder das einfachere Modell vorzuziehen ist. Eine adäquate Anzahl von Beobachtungen wird z.B. bei Steyerberg et al. (2001) für einen ausgeglichen Datensatz (Prävalenz, d.h. Anteil der Vorkommen ungefähr 50%) mit ca. zehn Präsenzen pro Variable angegeben.

4.2.1 Grundidee und Vorgehen der Variablenselektion

Die Grundidee der Variablenselektion ist es, *Bias* für Varianz einzutauschen. Wenn die Reduktion in Varianz die Erhöhung der Verzerrung übertrifft, dann sollte dem einfacheren Modell der Vorzug gegeben werden, weil es im Schnitt eine bessere Vorhersageleistung zeigen wird.

Ein Verfahren der Variablenselektion besteht üblicherweise aus einem Kriterium der Modellgüte, nach dem die verschiedenen Modellkandidaten bewertet werden, und einem Algorithmus, der aus der Fülle der Modellkandidaten das Modell herauszufinden sucht, das dieses Kriterium optimiert.

Wie oben diskutiert wurde, zeigen komplexere Modelle eine bessere Anpassung an die Trainingsdaten als einfachere Modelle. Daher wird für die Variablenselektion nicht allein auf die Modellanpassung (gemessen z.B. über die *Likelihood*, s. Schröder & Reineking 2004a, in diesem Band) geachtet, sondern die Modellkomplexität wird ebenfalls berücksichtigt. Die Modellkomplexität wird dabei in der Regel entweder über die Anzahl verwendeter Parameter gemessen oder über die Beträge der Modellparameterwerte.

Selektionskriterien

Ein großer Teil der Selektionskriterien lässt sich als ein gewichtetes Mittel aus einem Maß für die Modellanpassung (*model fit*) und einem Maß für die Modellkomplexität auffassen. Als Maß für die Modellanpassung wird die Devianz $D = -2\log(Likelihood) = -2LL$ verwendet. Das relative Gewicht der Modellkomplexität zur Modellanpassung wird durch einen Koeffizienten λ festgelegt. Ein solches Kriterium hat daher folgenden Aufbau:

$$\begin{aligned} \text{Kriterium} &= -2LL + \lambda \cdot \text{Modellkomplexität} \\ &= D + \lambda \cdot \text{Modellkomplexität} \end{aligned} \quad (4.1)$$

Die verschiedenen Methoden unterscheiden sich zum einen in der Wahl von λ und zum anderen in der Wahl des Maßes für die Modellkomplexität.

Die populärsten Selektionskriterien verwenden die Anzahl der Freiheitsgrade des Modells als Maß für die Modellkomplexität. Im Fall der logistischen Regression ist dies identisch mit der Anzahl geschätzter Parameter. Der AIC (*An Information Criterion*, oftmals nach seinem Autor auch als *Akaike Information Criterion* bezeichnet, Akaike 1974) ist gegeben durch $\lambda = 2$, während der BIC (*Bayes Information Criterion* oder *Schwarz Information Criterion*, Schwarz 1978) $\lambda = \ln n$ verwendet, wobei n die Anzahl der Beobachtungen ist. Der BIC bestraft die Modellkomplexität also für $n > 7$ stärker und führt daher zu sparsameren Modellen; oftmals haben die resultierenden Modelle aber eine zu große Verzerrung, d.h. sie enthalten zu wenige Variablen.

Für geringe Stichprobenumfänge wird ein modifizierter AIC_C vorgeschlagen, bei dem der Bestrafungsterm $2p$ mit dem Wert $1 + \frac{p+1}{n-p-1}$ multipliziert wird. Hierbei ist n die Anzahl der Beobachtungen und p die Anzahl der geschätzten Parameter (Harrell 2001; Hurvich & Tsai 1989).

Das wohl am häufigsten verwendete Verfahren zur Variablenselektion ist die schrittweise Selektion mit einem kritischen Signifikanzniveau für die hinzuzufügenden oder zu entfernenden Variablen (Hosmer & Lemeshow 2000). Dieses Verfahren lässt sich so verstehen, dass als Komplexitätsmaß die Anzahl der Freiheitsgrade verwendet wird. Der Wert von λ ist dabei abhängig von dem gewählten kritischen Signifikanzniveau α. Es ist oft darauf hingewiesen worden, dass der AIC einem Signifikanzniveau von 0,157 entspricht, wenn für die betrachtete Variable ein linearer Parameter geschätzt wird (z.B. Harrell 2001). Die Äquivalenz ergibt sich aus dem *Likelihood-Ratio* Test, bei dem die Teststatistik gleich der Differenz zwischen dem Modell ohne bzw. mit der zu testenden Variablen ist (vgl. Schröder & Reineking 2004a; Hosmer & Lemeshow 2000). Unter der Nullhypothese, dass der Parameter 0 ist, d.h. kein Unterschied zwischen den beiden Modellen besteht, ist die Teststatistik χ^2- verteilt mit einem Freiheitsgrad. Ein Wert von 2 für die χ^2-Statistik mit einem Freiheitsgrad entspricht dann gerade einem α von 0,157. Umgekehrt lässt sich über die Quantilsfunktion der χ^2-Verteilung berechnen, welcher Wert für die Statistik einem Quantil von 1-α bei einem Freiheitsgrad entspricht. Dies ist dann gleich dem äquivalenten Wert von λ. So entspricht ein kritisches Signifikanzniveau $\alpha = 0{,}05$ einem $\lambda = 3{,}84$.

Anstelle der Anzahl der Modellparameter können auch Parameterwerte zu einem Komplexitätsmaß zusammengefasst werden. Beim sogenannten Lasso (*least absolute shrinkage and selection operator*) Verfahren (Tibshirani 1996) wird mit der Summe der Absolutbeträge der j Modellparameter gearbeitet:

$$\text{Lasso} = -2LL + \lambda \sum |\beta_j|$$

Das Lasso-Kriterium führt dazu, dass u.U. für einige Variablen Parameterwerte gleich Null geschätzt werden können, so dass Variablenselektion und Parameterschätzung im gleichen Schritt stattfinden.

Warum werden einige Parameterwerte gleich Null geschätzt? Bei der Parameterkombination mit dem optimalen, d.h. geringsten Wert für das Lasso-Kriterium, ist die partielle Ableitung des Kriteriums nach den Parameter größer oder gleich 0, d.h. eine Veränderung der Parameterschätzwerte würde den Wert des Kriteriums verschlechtern. Das Lasso-Kriterium besteht – wie AIC und BIC auch – aus dem *Likelihood*-Term und dem Bestrafungsterm. Die partielle Ableitung des *Likelihood*-Terms nach dem Parameter β_j an der Stelle $\beta_j = 0$ wird praktisch immer ungleich Null sein, so dass entweder ein Wert für β_j größer oder kleiner Null zu einem günstigeren Wert für die *Likelihood* führen würde. Die Ableitung des Bestrafungsterms nach dem Parameter β_j ist jedoch gleich $\text{sign}(\beta_j) \cdot \lambda$, für $\beta_j \neq 0$, d.h. wenn der Betrag des Parameters zunimmt, ist die Ableitung positiv. Parameterschätzwerte ungleich Null werden also nur dann geschätzt, wenn die Ableitung des *Likelihood*-Terms nach dem Parameter dem Betrag nach größer ist als λ.

Alternativ zum Absolutbetrag der Modellparameter kann auch das Quadrat der Parameter zu einem Komplexitätsmaß zusammengefasst werden. Dies wird in der Regel *Penalized maximum likelihood*-Verfahren bezeichnet (le Cessie & van Houwelingen 1992):

$$PML = -2LL + \lambda \sum \beta_j^2$$

Das *Penalized maximum likelihood*-Verfahren führt im Gegensatz zum Lasso-Verfahren nicht zu einer Selektion von Variablen, da keine Parameterwerte gleich Null geschätzt werden. Die Ableitung des Bestrafungsterms nach dem Parameter β_j ist gleich $2\lambda\beta_j$, also proportional zum Wert des Parameters. Wenn dieser nahe bei Null ist, ist auch die Ableitung des Bestrafungsterms nahe Null. Eine geringe Verbesserung der log likelihood reicht daher aus, damit der Parameter ungleich Null geschätzt wird. Daher gehört das *Penalized maximum likelihood*-Verfahren nicht zu den Selektionsverfahren, wohl aber in die umfassendere Klasse von Verfahren zur Beschränkung der Modellkomplexität. Es kann mit anderen Verfahren, z.B. der schrittweisen Variablenselektion, kombiniert werden.

Wenn die Schätzwerte der Parameter zu einem Komplexitätsmaß zusammengefasst werden, spielt die Skalierung der Prädiktorvariablen eine entscheidende Rolle. Zwei Standardisierungsverfahren werden üblicherweise verwendet: Entweder werden alle erklärenden Variablen

so skaliert, dass sie einen Mittelwert von 0 und eine Standardabweichung von 1 haben (Harrell 2001; Tibshirani 1996), oder sie werden so skaliert, dass die Werte zwischen 0 und 1 liegen (Ripley 1996).

Sowohl beim Lasso als auch beim *Penalized maximum likelihood*-Verfahren muss ein geeignetes λ gewählt werden. Dazu gibt es jeweils unterschiedliche Ansätze. Zum einen kann dazu Kreuzvalidierung verwendet werden. Alternativ wird ein modifiziertes GCV (*generalized crossvalidation*) Kriterium oder ein modifizierter AIC verwendet (Harrell 2001; Tibshirani 1996). Aufgrund theoretischer Überlegungen im Kontext von neuronalen Netzwerken leitet Ripley (1996) einen Bereich von 0,001 bis 0,1 als Grundlage für die Wahl geeigneter Werte für den Fall der *Penalized maximum likelihood*-Verfahren ab.

In unserer Darstellung scheinen die Wahl des Komplexitätsmaßes und des Wertes von λ beliebig, was allerdings den zugrunde liegenden theoretischen Überlegungen nicht gerecht wird. Uns geht es aber an dieser Stelle in erster Linie darum, die strukturellen Ähnlichkeiten der verschiedenen Selektionskriterien aufzuzeigen.

Selektionsalgorithmus

Bei p Variablen beträgt die Anzahl verschiedener Variablenkombinationen $2^p - 1$. Das sind beispielsweise bei 21 Variablen mehr als zwei Millionen verschiedene Kombinationen. Es würde zu lange dauern, einzelne Modelle für all diese Kombinationen anzupassen. Vollständige Enumeration ist demnach für größere Werte von p auszuschließen. Daher werden andere Verfahren verwendet, um aus der Menge aller möglichen Modelle ein adäquates auszuwählen. Am häufigsten werden schrittweise Verfahren angewendet. Dabei werden jeweils die Modelle betrachtet, die sich vom Referenzmodell um eine Variable unterscheiden, und es wird ein neues Modell gewählt, wenn sich eines findet, das einen besseren Wert für das Selektionskriterium aufweist. Wenn mehrere Varianten zu einer Verbesserung des Selektionskriteriums führen, wird diejenige gewählt, die die größte Verbesserung bringt. Spezielle Spielarten dieses Ansatzes sind die schrittweise Rückwärts- und die schrittweise Vorwärtsauswahl. Bei der Rückwärtsauswahl bildet das (volle) Modell mit allen Variablen den Ausgangspunkt, d.h. das erste Referenzmodell, und es wird solange eine Variable entfernt, bis keine Variable mehr entfernt werden kann, ohne den Wert des Selektionskriteriums wieder signifikant zu verschlechtern. Die Vorwärtsauswahl verfährt spiegelbildlich; ausgehend von einem (Null-)Modell, indem keine Umweltvariablen enthalten sind, wird jeweils solange eine Variable hinzugefügt, bis keine signifikante Verbesserung des Kriteriums mehr erreicht wird.

Die schrittweisen Selektionsverfahren kann man als *greedy algorithms* verstehen, in ein lokales Optimum zu laufen, indem sie in jedem Schritt diejenige Wahl treffen, die sie dem Ziel am nächsten bringt. Solche Verfahren laufen Gefahr, leicht in lokalen Optima gefangen zu werden. Schrittweise Verfahren sind früh dafür kritisiert worden, nicht das Modell zu finden, das den besten Wert für das Selektionskriterium hat (Harrell 2001). Als Alternativen sind erweiterte schrittweise Verfahren vorgeschlagen worden, in denen nicht nur jeweils eine Variable hinzugefügt oder eliminiert wird, sondern Gruppen aus zwei oder drei Variablen (Lucic & Trinajstic 1999). Eine weitere Alternative ist die Verwendung von evolutionären Algorithmen (Kubinyi 1994, 1996). Schließlich ist es möglich, die für lineare Regression entwickelten *all-subset*-Algorithmen für die logistische Regression zu verwenden (Hosmer et al. 1989). Dabei wird jedoch lediglich ein approximativ lineares Problem gelöst. Keine der Alternativen zu den schrittweisen Verfahren hat sich etabliert, wobei ein Grund sein könnte, dass sie nicht im wünschenswerten Umfang in den gängigen Softwaresystemen implementiert sind.

Einen Sonderfall stellt die Variablenselektion mit dem Lasso-Verfahren dar, bei dem im Zuge der Parameterschätzung einige Parameterwerte als 0 geschätzt werden, so dass Parameterschätzung und Variablenselektion zusammenfallen.

4.2.2 Probleme der Variablenselektion

Die Probleme der Variablenselektion betreffen sowohl den Anwendungsbereich Erklärung als auch den Anwendungsbereich Prognose. Im Anwendungsbereich Erklärung treten insbesondere die folgenden Probleme auf (Harrell 2001):

- *Mögliche Auswahl von Variablen, die nur zufällig einen Zusammenhang mit der Zielvariablen aufweisen, und Verlust der Interpretierbarkeit von Signifikanzniveaus einzelner Variablen.*
  Aufgrund zufälliger Streuung werden einige Variablen einen signifikanten Zusammenhang mit der Zielvariablen zeigen, obwohl dieser in der Realität (in der Grundgesamtheit) vorhanden ist (Fehler 1. Art). Die Wahrscheinlichkeit eines solchen Falls ist gleich dem Signifikanzniveau α: wählt man beispielsweise $\alpha = 0{,}05$, so werden im Mittel 5% aller Variablen, die in Wirklichkeit *nicht* mit der Zielvariablen korrelieren, eine Korrelation mit der Zielvariablen zeigen. Das Ziel der Variablenselektion ist es nun, denjenigen Satz von Variablen zu identifizieren, der die beobachtete Varianz in den Daten zu einem möglichst hohen Anteil repräsentiert. Dies führt jedoch dazu, dass Variablen, die nur scheinbar mit der Zielvariablen korrelieren, mit höherer Wahrscheinlichkeit ausgewählt werden,

als die nominellen Signifikanzniveaus angeben. Das bedeutet, dass von den Variablen, die in den ausgewählten Modellen vorkommen und nur zufällig mit der Zielvariable korrelieren, mehr als 5% auf dem 5%-Niveau signifikant sind. Wie hoch der Anteil ist, ist jedoch nicht bekannt und variiert von Fall zu Fall. Verbunden mit dem Problem der verfälschten Signifikanzniveaus ist das Problem der verfälschten Schätzung der Konfidenzintervalle; diese werden zu schmal geschätzt.

- *Verfälschung der Parameterschätzwerte für die tatsächlich wichtigen Variablen.*
  Die in einem Modell vorhandenen Variablen beeinflussen die Parameterschätzwerte derjenigen anderen Variablen in dem Modell, mit denen sie korreliert sind. Korreliert eine fälschlicherweise berücksichtigte Variable ebenfalls mit einer tatsächlich relevanten Variable, so wird sich der Parameterschätzwert der relevanten Variable verändern. Dadurch kann die Größe des Effektes der relevanten Variablen auf die Zielvariable nicht mehr zuverlässig bestimmt werden. Da Variablenselektion dazu führt, dass die Variablen in dem ausgewählten Modell möglichst „signifikant" sind, werden die Schätzwerte der Parameter tendenziell stärker von 0 verschieden sein als deren tatsächliche Werte.
- *Übersehen von tatsächlich relevanten Variablen.*
  Man kann nicht generell davon ausgehen, dass Variablen, die nicht ausgewählt wurden, auch tatsächlich nicht relevant sind (Fehler 2. Art).
- *Instabilität der Auswahl.*
  Wenn die Variablenauswahl auf einem nur leicht veränderten Datensatz ausgeführt wird, kann eine andere ausgewählte Kombination von Variablen resultieren.

Im Bereich der Prognose liegt die Gefahr in der

- *Überanpassung*/Overfitting.
  Variablenselektion führt tendenziell dazu, dass zu einem wesentlichen Teil nicht mehr nur die in der Grundgesamtheit tatsächlich vorliegende Struktur, sondern die zufällige Streuung in der betrachteten Stichprobe beschrieben wird. Diese Überanpassung führt dazu, dass die Übertragbarkeit der Modelle stark eingeschränkt wird (vgl. Abb. 4.1).
- *Vorspiegelung hoher Modellgüte (Optimismus).*
  Auch in Fällen, in denen keine Beziehung zwischen den unabhängigen Variablen und der Zielvariablen besteht, führt Variablenselektion zu solchen Variablenkombinationen, die einen hohen Anteil der in den Daten vorhandenen Varianz erklären und dadurch die Vermutung nahe legen, dass in der Grundgesamtheit die durch das statistische Modell beschriebene Struktur tatsächlich vorhanden sei.
- *Verfehlen des besten Modells.*
  Die Methoden zur Bestrafung der Modellkomplexität über die Anzahl der Modellparameter (z.B. AIC) gehen davon aus, dass die Anzahl der Freiheitsgrade eines Modells mit der Anzahl Modellparameter übereinstimmt. Das ist dann gegeben, wenn jeweils nur ein Modell pro Komplexitätsstufe (d.h. Anzahl Modellparameter) getestet wird. Bei der Variablenselektion wird jedoch in der Regel nicht genau ein Modell mit k Modellparametern betrachtet, sondern es wird das beste Modell mit k Modellparametern aus einer größeren Anzahl von Modellen mit ebenfalls k Parametern ausgewählt und dieses dann mit Modellen anderer Komplexität verglichen. Durch den Auswahlprozess ist die effektive Anzahl von Freiheitsgraden dann höher als die nominelle Anzahl. Wird dies nicht berücksichtigt, werden diejenigen Modellkomplexitätsstufen begünstigt, bei denen eine große Anzahl verschiedener Modelle getestet wurde.

4.2.3 Alternativen und Ergänzungen

Dem Problem der instabilen Auswahl von Variablen kann zum Teil begegnet werden, indem das Selektionsverfahren auf *Bootstrap*-Stichproben, die vom ursprünglichen Datensatz durch Ziehen mit Zurücklegen erzeugt werden, wiederholt wird (vgl. Beitrag zur Validierung von Schröder & Reineking 2004b, in diesem Band). Die Häufigkeit, mit der verschiedene Variablen oder Variablenkombinationen selektiert werden (*resampling stability*), kann dann verwendet werden, um eine geeignete Variablenkombination auszuwählen. Ein Beispiel hierfür findet sich bei Wisnowski et al. (2003). Ein Problem des Ansatzes ist, dass er sehr rechenintensiv ist. Da für eine interne Validierung (Schröder & Reineking 2004b) der gesamte Modellbildungsprozess vielfach wiederholt werden muss, stößt dieses Verfahren unter Umständen an die Grenzen des vertretbaren Aufwands. Um mit dem Problem umzugehen, dass bei Variablenselektion die effektive Anzahl geschätzter Parameter eines Modells größer ist als die nominelle Anzahl (d.h. die Anzahl von Parametern des Modells), schlagen Tibshirani & Knight (1999) ein weiteres Kriterium vor, das *Covariance inflation criterion (CIC)*. Dabei wird das Selektionsverfahren auf Versionen des Trainingsdatensatzes angewendet, bei denen die werte der Zielvariablen zufällig permutiert wurde. Die Kovarianz zwischen den Vorhersagen und den beobachteten Werten wird dann als ein Maß für die Überanpassung des Modells an die Daten verwendet, und die Modellanpassung wird damit bestraft.

Eine radikale Antwort auf die Probleme, die bei einer datengeleiteten Variablenselektion auftreten, ist, die Variablenauswahl vollständig aufgrund von

Überlegungen zu treffen, die man aufgrund bereits bestehenden Wissens über die Habitatansprüche der betrachteten Art oder vergleichbarer Arten anstellt. Es ist ohne Frage sinnvoll, die betrachteten Variablen entsprechend dem vorhandenen Vorwissen auf diejenigen zu beschränken, für die ein Zusammenhang mit dem Vorkommen plausibel ist. Das Problem der starken *a priori* Auswahl liegt darin, dass die Zahl "vernünftiger Variablen" oftmals immer noch sehr groß ist, und die weitere Reduzierung der Variablen schwierig zu begründen ist. Es ist eine offene Frage, wie gut das Expertenwissen sein muss, damit eine starke *a priori* Auswahl zu besseren Ergebnissen führt als eine datengeleitete Auswahl, und ob das vorhandene Expertenwissen üblicherweise von der notwendigen Güte ist (Reineking & Schröder 2003).

Eine weitere Alternative ist, die Anzahl Prädiktorvariablen zu reduzieren, ohne dabei die Daten zu Vorkommen und Nichtvorkommen zu berücksichtigen. Die Grundidee ist, die in den vielen Umweltvariablen enthaltene Information so gut wie möglich durch wenige ausgewählte oder neu berechnete Variablen zu erfassen. Eine Variante dieses Vorgehens ist die Durchführung einer Hauptkomponentenanalyse und die Auswahl der ersten k Hauptkomponenten als unabhängige Variablen (*principal component regression*, Quinn & Keough 2002). Alternativ können Clusterverfahren eingesetzt werden, um die Variablen in Gruppen einzuteilen. Innerhalb der Gruppen kann entweder eine repräsentative Variable ausgewählt oder eine neue Variable als gewichtetes Mittel der in der Gruppe enthaltenen Variablen berechnet werden. Harrell (2001) empfiehlt die zweite Variante, da sie stabiler sei. Insbesondere die Verwendung hierarchischer Clusterverfahren ist sehr attraktiv, weil sie dazu anregen, die Beziehungen der unabhängigen Variablen untereinander zu betrachten. Das Potential dieser Aggregationsmethoden für eine verbesserte Modellbildung ist bislang schwierig zu beurteilen, da wenige Vergleichsstudien vorliegen. In den meisten Fällen führt die Aggregation zu Informationsverlusten und Schwierigkeiten bei der Interpretation, wenn neue Variablen aus mehreren Ausgangsvariablen gebildet werden. Da die Vorkommen und Nichtvorkommen nicht bei der Auswahl berücksichtigt werden, sondern lediglich die Beziehung der erklärenden Variablen untereinander, ist nicht gewährleistet, dass der Verlust von Informationen, die für die Vorhersage der Vorkommen relevant sind, minimiert wird.

Wenn das Ziel der Modellierung die Vorhersage ist und die verwendeten Variablen auch für die vorherzusagenden Gebiete erhoben worden sind, steht das sogenannte *model averaging* als Alternative zur Verfügung (Hoeting et al. 1999). Dabei wird ein gewichtetes Mittel aus den betrachteten Modellen gebildet. Die Gewichte hängen sowohl von der Devianz der Modelle als auch der Anzahl enthaltener Parameter ab. Modelle, die die Daten zu schlecht beschreiben, können ausgeschlossen werden, d.h. sie erhalten ein Gewicht von Null. Ein Beispiel aus dem Bereich der Habitatmodellierung geben Wintle et al. (2003).

Abschließend sei auf das Verfahren der hierarchischen Partitionierung (*hierarchical partitioning*, Chevan & Sutherland 1991; MacNally 2002) hingewiesen, das die Variablenselektion unterstützen kann. Dieses Verfahren zielt nicht auf das Finden eines einzelnen „besten" Modells, sondern vielmehr darauf, diejenigen Variablen zu finden, die den stärksten Einfluss auf die Zielvariable haben. Auf der Grundlage von Gütemaßen, die für alle $2^p - 1$ möglichen Variablenkombinationen berechnet werden, liefert die hierarchische Partitionierung für alle Prädiktorvariablen einen Überblick über den alleine diesen Variablen (unabhängigen) sowie den diesen Variablen zusammen mit anderen zuzuschreibenden (gemeinsamen) Effekt auf die Responsevariable. Diese Information kann dann die Auswahl der letztlich im Modell verwendeten Prädiktoren unterstützen.

4.3 Zusammenfassung

Der *Bias-Variance-Trade-off* bestimmt die Komplexität der Analyse, die für einen Datensatz gegebener Größe angemessen ist. In kleinen Datensätzen zahlt es sich aus, mit einfachen Methoden zu arbeiten.

Wenn man sich überhaupt für ein automatisches Verfahren der Modellauswahl entscheidet, ist das rückwärts schrittweise Verfahren zu empfehlen (Harrell 2001; Steyerberg et al. 1999). Einen „Strafterm" integrierende *Penalization*-Verfahren sind ein vielversprechender Ansatz, um besser kalibrierte Modelle zu erhalten, und können in Kombination mit schrittweisen Variablenselektionsverfahren verwendet werden. Das Lasso-Verfahren ist als Kombination von *penalization* und Selektion attraktiv.

Wenn in erster Linie die Vorhersage von Interesse ist, stellt *model averaging* eine vielversprechende Alternative zur Variablenselektion dar.

Als grundlegendes Leitmotiv für die Variablenselektion kann der William von Occam zugeschriebene Ausspruch dienen: „*One should not increase, beyond what is necessary, the number of entities required to explain anything.*" Oder einfach: „*Keep it simple!*"

4.4 Danksagung

Die Autoren bedanken sich bei Hans-Peter Bäumer, Universität Oldenburg, sowie Lorenz Fahse und Tamara Münkemüller, UFZ Leipzig-Halle, für die wertvollen, hilfreichen Kommentare zum Manuskript.

Literaturverzeichnis

Akaike, H. 1974. A new look at statistical-model identification. *IEEE Transactions on Automatic Control*, AC19(6):716–723.

Chevan, A. & Sutherland, M. 1991. Hierarchical partitioning. *American Statistician*, 45(2):90–96.

Harrell, Frank E., J. 2001. *Regression Modeling Strategies - with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer, New York.

Hastie, T., Tibshirani, R. & Friedman, J. H. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, Berlin.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. 1999. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401.

Hosmer, D. W., Jovanovic, B. & Lemeshow, S. 1989. Best subsets logistic-regression. *Biometrics*, 45(4):1265–1270.

Hosmer, D. W. & Lemeshow, S. 2000. *Applied Logistic Regression*. John Wiley & Sons, New York, 2nd edition.

Hurvich, C. M. & Tsai, C. L. 1989. Regression and time-series model selection in small samples. *Biometrika*, 76(2):297–307.

Kubinyi, H. 1994. Variable selection in QSAR studies 1. an evolutionary algorithm. *Quantitative Structure-Activity Relationships*, 13(3):285–294.

Kubinyi, H. 1996. Evolutionary variable selection in regression and PLS analyses. *Journal of Chemometrics*, 10(2):119–133.

le Cessie, S. & van Houwelingen, J. C. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41:191–201.

Lucic, B. & Trinajstic, N. 1999. A new efficient approach for variable selection based on multiregression: Prediction of gas chromatographic retention times and response factors. *Journal of Chemical Information and Computer Sciences*, 39(3):610–621.

MacNally, R. 2002. Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables. *Biodiversity and Conservation*, 11(8):1397–1401.

Quinn, G. P. & Keough, M. J. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Reineking, B. & Schröder, B. 2003. Computer-intensive methods in the analysis of species-habitat relationships. In Breckling, B., Reuter, H. & Mitwollen, A., editors, *Gene, Bits und Ökosysteme - Implikationen neuer Technologien für die ökologische Theorie*, pages 165–182. Peter Lang.

Reineking, B. & Schröder, B. 2004. Gütemaße für Habitatmodelle. *UFZ-Bericht*, 9/2004:27–38.

Ripley, B. D. 1996. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge.

Schröder, B. & Reineking, B. 2004a. Modellierung der Art-Habitat-Beziehung - ein Überblick über die Verfahren der Habitatmodellierung. *UFZ-Bericht*, 9/2004:5–26.

Schröder, B. & Reineking, B. 2004b. Validierung von Habitatmodellen. *UFZ-Bericht*, 9/2004:47–55.

Schwarz, G. 1978. Estimating dimension of a model. *Annals of Statistics*, 6(2):461–464.

Steyerberg, E. W., Eijkemans, M. & Habbema, J. 2001. Application of shrinkage techniques in logistic regression analysis: a case study. *Statistica Neerlandica*, 55(1):76–88.

Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. 1999. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, 52(10):935–942.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):267–288.

Tibshirani, R. & Knight, K. 1999. The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 61:529–546.

Wintle, B. A., McCarthy, M. A., Volinsky, C. T. & Kavanagh, R. P. 2003. The use of bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, 17(6):1579–1590.

Wisnowski, J. W., Simpson, J. R., Montgomery, D. C. & Runger, G. C. 2003. Resampling methods for variable selection in robust regression. *Computational Statistics & Data Analysis*, 43(3):341–355.

4.5 Datenblatt

4.5.1 Software

Die Autoren empfehlen die Verwendung von R (*free software*) oder S-Plus (Insightful).

4.5.2 Webresources

R unter `www.r-project.org`, wichtige Bibliotheken für R und S-Plus sind:
Harrell, F.: `Design` und `Hmisc`: `http://hesweb1.med.virginia.edu/biostat/`
Tibshirani, R.: `Lasso`: `http://www-stat.stanford.edu/~tibs/lasso.html`, `http://lib.stat.cmu.edu/S/lasso`
Walsh, C. and MacNally, R.: `hier.part`: Hierarchical Partitioning.

4.5.3 Kommentierte Literatur

Empfehlenswerte Bücher zum Thema sind: Harrell (2001) und Hosmer & Lemeshow (2000).

Beispielhafte Anwendungen mit Methodenvergleichen finden sich bei: Steyerberg et al. (2001); Reineking & Schröder (2003).

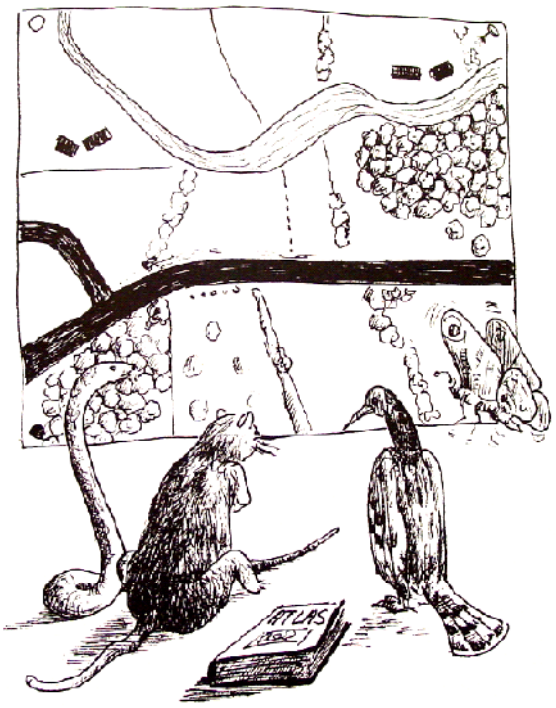
HABITATMODELLE

Methodik, Anwendung, Nutzen

Herausgeber

Carsten F. Dormann
Thomas Blaschke
Angela Lausch
Boris Schröder
Dagmar Söndgerath

Tagungsband
zum Workshop
8.-10. Oktober 2003,
UFZ Leipzig



UFZ - UMWELTFORSCHUNGSZENTRUM
LEIPZIG-HALLE GMBH IN DER HELMHOLTZ-GEMEINSCHAFT