5

Validierung von Habitatmodellen

Boris Schröder[1] & Björn Reineking[2]

[1] Universität Potsdam, Institut für Geoökologie, Postfach 601553, 14415 Potsdam, Email: boschroe@rz.uni-potsdam.de
[2] Umweltforschungszentrum Leipzig/Halle - UFZ GmbH, Sektion Ökosystemanalyse, Permoserstr. 15, 04318 Leipzig, bjoern.reineking@ufz.de

Mit der zunehmenden Bedeutung von Habitatmodellen in der ökologischen, naturschutzbiologischen und biogeographischen Forschung und Anwendung gewinnt die Validierung dieser Modelle eine besondere Relevanz. Die Modellvalidierung zielt auf zwei verschiedene Aspekte: zum einen dient sie der objektiven Einschätzung der Prognosegüte, zum anderen der Überprüfung des Gültigkeitsbereichs. In diesem Beitrag werden sowohl verschiedene Methoden der internen Validierung als auch die externe Validierung durch Modellübertragungen vorgestellt und Empfehlungen zu ihrer Durchführung gegeben.

5.1 Einleitung

Habitatmodelle haben sich zu einem wertvollen Werkzeug in der Ökologie, Naturschutzbiologie und Biogeographie entwickelt. Sie erlauben die funktionale Beschreibung der Art-Habitat-Beziehung und ermöglichen die Prognose der potentiellen räumlichen Verteilungsmuster von Organismen in Abhängigkeit von aktuellen und zukünftigen Umwelteigenschaften (Morrison et al. 1998; Scott et al. 2002). Die Modellvalidierung zielt auf zwei verschiedene Aspekte: zum einen dient sie der adäquaten Einschätzung der Prognosegüte, zum anderen der Überprüfung des Gültigkeitsbereichs. Sie ist ein unverzichtbarer Bestandteil des Modellierungsprozesses, zumal wenn es um ihre Modellanwendung in Fragestellungen der Entscheidungsunterstützung in Naturschutz und Habitatmanagement geht – etwa wenn mit Habitatmodellen erstellte Prognosen die Grundlage von Reservatserweiterungen oder Wiedereinbürgerungsprogrammen für gefährdete Arten sind (Fielding & Haworth 1995; Morrison et al. 1998; Noon 1986). So wichtig die Modellvalidierung ist, so selten ist sie in der Literatur zur Habitatmodellierung dokumentiert (Fleishman et al. 2003; James & McCulloch 2002; Schröder & Richter 2000).

5.1.1 Begriffsdefinitionen

Wir folgen der Ansicht von Levins (1966), wonach „Modelle nicht hinsichtlich der Kriterien ‚wahr' oder ‚falsch' bewertet werden können, sondern allein dahingehend, ob sie gute, testbare Hypothesen generieren und im Rahmen ihres Anwendungskontexts zufrieden stellende Prognosegüten erreichen".

Rykiel (1996) kommentiert in einem Übersichtsartikel das begriffliche Durcheinander zum Thema Validierung. Seine Begriffsdefinitionen für verschiedene Aspekte der Validierung geben wir an dieser Stelle gekürzt wieder. Wenngleich sich diese Definitionen bei Rykiel (1996) auf Simulationsmodelle in der Ökologie beziehen, lassen sie sich auch auf den Kontext statistischer Habitatmodelle übertragen (s. Tab. 5.1).

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.

Tabelle 5.1. Begriffsdefinitionen für verschiedene Aspekte der Validierung nach Rykiel (1996) und ihre Übertragung auf den Kontext der Habitatmodellierung. Die grau markierten Aspekte werden in diesem Beitrag behandelt.

Validierungsaspekte	Definition bei Rykiel (1996)	Übertragung auf den Kontext der Habitatmodellierung
■ Verifikation (*verification*)	Demonstration, dass der Modellformalismus korrekt ist und keine logischen Fehler enthält.	Verifikation und Kalibrierung (i.S. der Parameteridentifikation) sind wichtige Bestandteile der statistischen Modellbildung, wobei letztere die Schätzung der Modell-
■ Kalibrierung (*calibration*)	Schätzung und Anpassung der Parameter und Konstanten im Modell mit dem Ziel, die Übereinstimmung von Modellvorhersage und Daten zu erhöhen.	parameter auf der Basis eines Datensatzes, den sog. Trainingsdaten, beschreibt. Im weiteren Text bezieht sich in Anlehnung an Reineking & Schröder (2004a) Kalibrierung als ein Gütekriterium auf das Zueinanderpassen von Vorhersagen und Beobachtungen (s.u.).
■ Validierung (*validation*)	Demonstration, dass ein Modell im Rahmen seines Anwendungskontexts eine zufrieden stellende Prognosegüte aufweist.	Validierung beschreibt die Modellbewertung auf Grundlage von Stichproben, die nicht zur Modellschätzung verwendet wurden (Testdaten); extern oder intern.
■ Glaubwürdigkeit (*credibility*)	ausreichendes Maß an Vertrauen in die Validität des Modells, so dass es in der Forschung oder als Entscheidungsunterstützung angewendet werden kann.	Glaubwürdigkeit zielt v.a. auf die adäquate Einschätzung der Modellgüte ab, da Gütemaße für Modelle zu optimistisch geschätzt werden, wenn sie allein auf den Trainingsdaten berechnet werden (Verbyla & Litvaitis 1989).
■ Qualifikation (*qualification, applicability*)	Festlegung des Gültigkeitsbereichs, in dem das validierte Modell angewendet werden kann, ohne seine Validität zu verlieren.	Unter Qualifikation sind z.B. Übertragbarkeitstests (Bonn & Schröder 2001; Schröder 2000) für Modellübertragungen in Raum und/oder Zeit zur Festlegung des Gültigkeitsbereichs zu verstehen.

5.1.2 Ziele der Validierung von Habitatmodellen

In der Regel ist die Prognosegüte eines Habitatmodells für Flächen besser, die im Trainingsdatensatz berücksichtigt, besser als für solche Flächen, die bei der Schätzung der Regressionskoeffizienten nicht einbezogen worden waren (Verbyla & Litvaitis 1989). Dies gilt vor allem dann, wenn die Modelle auf der Grundlage kleiner Datensätze geschätzt werden (Steyerberg et al. 2001). Die scheinbare Modellgüte (*apparent performance*) ist also zu optimistisch, was an der Überanpassung (*overfitting*, s. Abb. 5.1) des Modells liegen kann (Harrell 2001).

Die Modellvalidierung hingegen zielt auf eine unverzerrte, nicht-optimistische Schätzung der Modellgüte, um festzustellen, ob und mit welcher Güte die Prognosen des Modells auch für neue Fälle verwendet werden können. Dieser Punkt ist auch die Grundlage der von Rykiel (1996) beschriebenen Glaubwürdigkeit (*credibility*). Validierung beschreibt also die Modellbewertung auf Grundlage von Stichproben, die nicht zur Modellschätzung verwendet wurden (Testdaten).

Habitatmodelle, die auf der Datengrundlage eines Untersuchungsgebietes und eines Untersuchungsjahres erstellt wurden, erlauben - wenn ihre Übertragbarkeit nicht untersucht und festgestellt wurde - nur Aussagen, die räumlich und zeitlich auf die der Modellierung zugrunde liegenden Daten begrenzt sind (Fielding & Haworth 1995). Um diese lokale Aussagekraft der entwickelten Modelle in Richtung allgemeiner Aussagen zu erweitern, bedarf es ihrer Validierung in räumlicher und zeitlicher Dimension (Fielding & Haworth 1995;

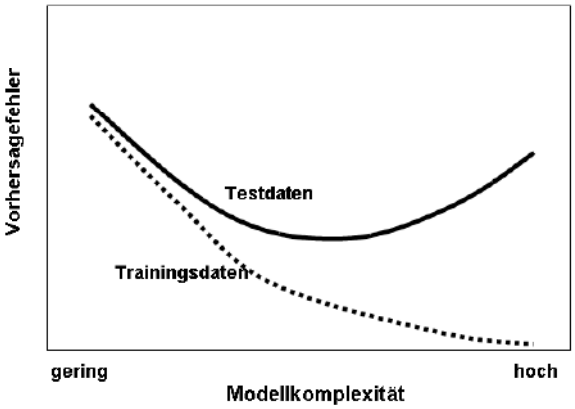


Abb. 5.1. Überanpasung/*Overfitting*: mit zunehmender Modellkomplexität (z.B. durch zunehmende Anzahl Prädiktorvariablen) verringert sich zwar der Vorhersagefehler auf den Trainingsdaten immer weiter; auf den Testdaten hingegen erreicht er ein Minimum und nimmt dann wieder zu (verändert nach Hastie et al. 2001).

Schamberger & O'Neil 1986; Schröder 2000; Schröder & Richter 2000), was dem Aspekt der Qualifikation bei Rykiel (1996) entspricht.

Die Gütekriterien, die im Zuge der Validierung adäquat quantifiziert werden, sind Thema eines weiteren Beitrags in diesem Band (Reineking & Schröder 2004a). Validierung betrachtet sowohl die Kalibrierung eines Modells, d.h. die Übereinstimmung zwischen den beobachteten relativen Häufigkeiten der Vorkommen und den vorhergesagten Vorkommenswahrscheinlich-

keiten, als auch die Diskriminanz, also die Fähigkeit des Modells, zwischen Vorkommen und Nichtvorkommen zu unterscheiden. Als Kriterien eignen sich beispielsweise: R^2_N nach Nagelkerke (1991), *AUC* (Fläche unter der *Receiver-Operating-Characteristic*-(ROC-)Kurve), sowie y-Achsenabschnitt und Steigung der Kalibrierungskurve. Sie sind sowohl in R als auch S-Plus implementiert.

Für die schon seit den 80er Jahren in den USA institutionalisierten *Habitat-suitability-index*-(HSI)Modelle werden mitunter Studien zur Validierung durchgeführt (z.B. Brooks 1997; Clark & Lewis 1983; Prosser & Brooks 1998), für die Roloff & Kernoban (1999) Hinweise zur Durchführung geben.

5.2 Validierungsverfahren

Es gibt grundsätzlich zwei verschiedene Ansätze zur Validierung von Habitatmodellen – externe und interne Validierung. Bei der externen Validierung werden unabhängige Testdaten zur Einschätzung der Modellgüte verwendet, während bei der internen Validierung die Testdaten aus dem Trainingsdatensatz mittels Resamplingverfahren erstellt werden (Chatfield 1995; Harrell 2001; Manly 1997; Verbyla & Litvaitis 1989).

5.2.1 Interne Validierung

Bei der internen Validierung wird nur ein Datensatz zur Kalibrierung des Modells verwendet, dessen Modellgüte dann durch Anwendung eines Resamplingverfahrens abgeschätzt wird.

Data-Splitting

Die einfachste Form der Validierung ist die einmalige Teilung des zur Verfügung stehenden Datensatzes in einen Trainingsdatensatz zur Modellbildung und einen Testdatensatz zur Modellvalidierung (*data-splitting*). Die Teilung erfolgt zufällig, wobei die Verteilung der Response- und Prädiktorvariablen berücksichtigt werden kann oder nicht. Diskriminierung und Kalibrierung werden dann auf der Basis der Testdaten ermittelt. Das Verfahren kann auch als externe Validierung interpretiert werden, bei der der „quasi-"unabhängige Testdatensatz nicht aus unabhängig durchgeführten Untersuchungen, sondern aus derselben Erhebung stammt (Harrell 2001).

Das Verfahren ist nur anwendbar, wenn eine Stichprobe ausreichenden Umfangs zur Verfügung steht. Harrell (2001) empfiehlt für den Testdatensatz eine Größe von mindestens 100 Beobachtungen für die im Datensatz mit geringerer Häufigkeit vertretene Kategorie der Responsevariable (Präsenz bzw. Absenz), um die Validierung extremer Vorhersagen gewährleisten zu können. In der Literatur tauchen verschiedene Vorschläge für das Verhältnis zwischen Trainings- und Testdatensatzgröße auf. Steyerberg et al. (2001) testen beispielsweise 50:50 und 66,67:33,33, Manel et al. (1999) 80:20, was der Faustregel bei Huberty (1994) entspricht. Hastie et al. (2001) schlagen vor, den Datensatz dreizuteilen. Auf dem ersten Teil werden die verschiedenen Modelle angepasst, anhand der Modellleistung auf dem zweiten Teil wird ein Modell ausgewählt, und dessen Leistung auf dem letzten Teil bewertet. Hastie et al. (2001) betonen, dass die Aufteilung von den Daten abhängt, insbesondere vom Signal/Rauschen-Verhältnis; eine typische Aufteilung sei 50:25:25 (vgl. auch Hurvich & Tsai 1990) .

Data-Splitting hat den Nachteil, dass die Datensatzgröße sowohl für die Modellbildung als auch für die Modellevaluation stark verringert wird, was einen unnötigen Verlust von Information bedeutet. Um dieselbe Genauigkeit bei der Schätzung der Prognosegüte zu erhalten, muss bei diesem Verfahren eine viel größere Datenmenge verwendet werden als bei Resamplingverfahren. Letztlich wird durch Aufteilung des Datensatzes nicht das eigentliche Modell validiert, sondern nur ein Modell, das auf einer Teilmenge der Daten geschätzt wurde (Harrell 2001). Osborne & Suarez-Seoane (2002) empfehlen, keine Partitionierung der Daten nach geographischen Gesichtspunkten vorzunehmen, wenn Modelle für größere Skalen geschätzt werden sollen, sondern die Einteilung zufällig vorzunehmen.

Die geschätzte Prognosegüte ist eine Kennzahl für eine Stichprobe. Das bedeutet, dass ihr Wert in Abhängigkeit von der zugrunde liegenden Stichprobe schwankt. Die verschiedenen Validierungsverfahren zielen darauf ab, die tatsächliche Prognosegüte des Modells zu bestimmen. Zur Bewertung der Verfahren ist es nützlich, zwei Eigenschaften der Schätzung zu betrachten – die Verzerrung (*Bias*) und die Varianz (vgl. Reineking & Schröder 2004a, in diesem Band). Die Verzerrung ist der systematische Anteil an der Abweichung der vorhergesagten von den wahren Vorkommenswahrscheinlichkeiten. Die Schätzung der Prognosegüte auf den Originaldaten ist, wie bereits dargelegt, verzerrt – sie fällt im Mittel zu positiv aus (Steyerberg et al. 2001; Verbyla & Litvaitis 1989). Der andere Aspekt der Schätzung ist die Varianz, die beschreibt, wie sehr die Werte der geschätzten Prognosegüte zwischen verschiedenen Stichproben schwanken. Die Varianz sollte möglichst klein sein – wenn das Verfahren im Mittel zwar den richtigen Wert bestimmt, aber häufig vom tatsächlichen Wert abweicht und die Prognosegüte mal zu optimistisch und mal zu pessimistisch einschätzt, dann ist es ebenfalls nur von eingeschränktem Nutzen.

Im Hinblick auf die Verzerrung geht es bei der internen Validierung darum, eine Balance zwischen zwei gegenläufigen Interessen zu finden: Zum einen sollte die

Modellgüte möglichst auf Daten geschätzt werden, die nicht zur Anpassung des Modells verwendet wurden, damit die Schätzung nicht zu optimistisch wird. Auf der anderen Seite sollten möglichst alle Daten für die Modellanpassung verwendet werden, denn die Güte eines Modells nimmt im Fall eines konsistenten Schätzers für die Modellparameter zunächst mit zunehmender Datenmenge zu. Wenn nun ein beträchtlicher Teil der Daten für die Validierung zur Seite gelegt wird und nicht für die Anpassung des Modells zur Verfügung steht, erhält man ein schlechteres Modell, als wenn der komplette Datensatz zur Modellanpassung verwendet würde, und die Abschätzung der Modellgüte ist unter Umständen zu pessimistisch.

Die Varianz in der Schätzung hat zwei Ursachen. Zum einen wird die Güte aufgrund einer endlichen Stichprobe aus den Daten geschätzt – wenn beispielsweise 10% der Daten zur Seite gelegt wurden. Zum anderen sind die Daten selbst lediglich eine Stichprobe, eine neue Erhebung würde zu anderen beobachteten Werten führen.

Die nachfolgend vorgestellten *Resampling*-Verfahren unterscheiden sich in der Art, wie die Daten in Test- und Trainingsdaten aufgeteilt werden, und wie die Modellleistung auf Test- und Trainingsdaten zu einer Schätzung der Modellgüte zusammengeführt wird. In der Folge unterscheiden sie sich im Hinblick auf *Bias* und Varianz der Schätzung der Modellgüte.

Kreuzvalidierung

Kreuzvalidierung bezeichnet eine Verallgemeinerung des *Data-Splittings*, die einige Probleme dieses Verfahrens umgeht. Hierbei werden die Gesamtdaten zufällig in gleich große Teilmengen aufgeteilt. Wenn diese Teilmengen jeweils den Umfang eines Zehntels des Datensatzes aufweisen, spricht man von zehnfacher Kreuzvalidierung. Die ersten neun Zehntel werden dann als Trainingsdatensatz zur Modellbildung verwendet (inkl. Test auf Interaktionen, Variablenselektion etc., s. Schröder & Reineking 2004) und dann die Modellgüte auf dem restlichen Zehntel bestimmt. Nach diesem Durchgang werden Test- und Trainingsdatensätze gewechselt und die Prozedur wiederholt. Nach zehn Wiederholungen, wenn also jede Beobachtung einmal zur Evaluierung im Testdatensatz beigetragen hat, können dann Mittelwerte der Gütekriterien gebildet werden. Um die Stabilität der Ergebnisse zu erhöhen, kann das gesamte Verfahren mit neuen Aufteilungen mehrfach wiederholt werden.

Wenn die Prävalenz deutlich von 50% abweicht, d.h. wenn also beispielsweise wesentlich mehr Nichtvorkommen in den Daten enthalten sind als Vorkommen, dann ist es sinnvoll, geschichtete Stichproben zu ziehen (*stratified sampling*), d.h. die Aufteilung der Daten so zu wählen, dass der Anteil von Vorkommen in jedem der Teile etwa gleich ist.

Harrell (2001) führt als Nachteil dieses Verfahrens an, dass auch hier die Wahl der Datensatzgrößen beliebig ist und häufig mehr als 200 Wiederholungen nötig sind, um genaue Aussagen zur Modellgüte treffen zu können. Zudem repräsentiert die Kreuzvalidierung nicht zwangsläufig die Variabilität der Variablenselektion, etwa wenn die gewählten Trainingsdatensätze ähnlich groß sind wie der Originaldatensatz und dementsprechend nur kleine Testdatensätze verwendet werden. Letztlich wird auch bei der Kreuzvalidierung nicht das eigentliche Modell für alle Beobachtungen validiert. Beispiele für die Durchführung der Kreuzvalidierung in der Habitatmodellierung finden sich u.a. bei Bio et al. (2002), Boyce et al. (2002), Lehmann et al. (2002), Manel et al. (1999) und Schröder (2000).

Leave-One-Out-Kreuzvalidierung

Ein Extremfall der Kreuzvalidierung ist das *Leave-one-out*-Verfahren, bei der jeweils nur eine Beobachtung zum Testen des Modells verwendet wird, das auf der Grundlage aller anderen $n-1$ Beobachtungen geschätzt wird. Dies wird n-mal wiederholt, bis jede Beobachtung einmal zum Testen verwendet wurde. Nach Efron (1983) ist aber das oben erläuterte Verfahren gruppierter Kreuzvalidierung vorzuziehen, da es eine genauere Abschätzung der Modellgüte ermöglicht. Manel et al. (1999) zeigen ein Beispiel für die *Leave-One-Out*-Kreuzvalidierung in der Habitatmodellierung.

Bootstrap

Einfacher *Bootstrap* (*regular bootstrap*)

Bootstrapping liefert fast unverzerrte Schätzungen der Modellgüte ohne Daten bei der Schätzung der Parameter des endgültigen Modells zurückzuhalten (Efron & Tibshirani 1993). Das Verfahren läuft wie folgt ab: es wird eine *Bootstrap*-Stichprobe durch zufälliges Ziehen von n Beobachtungen aus dem Gesamtdatensatz mit Zurücklegen erzeugt. Dieser *Bootstrap*-Datensatz wird als Trainingsdatensatz verwendet, während die Originaldaten als Testdatensatz fungieren. Die Differenz zwischen der Vorhersagegüte auf Trainings- und Testdatensatz ist dann eine Schätzung für den Optimismus des Modells. Diese Prozedur wird häufig (meist $\geq$ 100-mal) wiederholt und abschließend werden Mittelwerte aller Gütekriterien berechnet. Der durchschnittliche Optimismus wird dann von den naiven Gütekriterien des eigentlichen Modells subtrahiert und man erhält eine Schätzung der Modellgüte, die hinsichtlich der Überanpassung korrigiert ist. Durch *Bootstrapping* wird also der Prozess, der zur Schätzung des Originalmodells geführt hat, validiert (Harrell 2001).

Ein weiterer Vorteil des *Bootstrapping* ist die Möglichkeit, schrittweise Variablenselektion hinsichtlich der Beliebigkeit der ins Modell aufgenommenen Variablen zu überprüfen (Wisnowski et al. 2003). Oftmals weisen einzelne Variablen sehr ähnliche Beiträge zum Modell auf, und es hängt nur von einzelnen Beobachtungen im Trainingsdatensatz ab, welche von ihnen bei schrittweiser Selektion im Modell verbleibt (vgl. Reineking & Schröder 2004b). Diese werden für jede *Bootstrap*-Stichprobe dokumentiert, so dass bei Mustern, die von einem stochastischen Prozess generiert worden sind, anstelle schrittweiser Variablenselektion auf andere Selektions- oder Datenreduktionsverfahren zurückgegriffen werden kann.

Beispiele für die Anwendung des *Bootstrapping* in der Habitatmodellierung finden sich bei Oppel et al. (2004), Peppler-Lisbach & Schröder (2004) sowie Reineking & Schröder (2003). Guisan & Zimmermann (2000) zeigen eine Anwendung für Habitatmodelle mit ordinalskalierter Responsevariablen.

***Bootstrap*-Variante I:.632**

Bei der .632 Variante des *Bootstrap-Resamplings* werden die nicht in den *Bootstrap*-Stichproben (d.h. den jeweiligen Trainingsdatensätzen) enthaltenen Beobachtungen anstelle des Originaldatensatzes zum Testen der Modellgüte herangezogen. Seinen Namen hat das Verfahren daher, dass im Durchschnitt 63,2% der Beobachtungen mindestens einmal in einer *Bootstrap*-Stichprobe enthalten sind. Die geschätzte Modellgüte entspricht dann einer gewichteten Kombination: $0,368 \cdot smg + 0,632 \cdot mg$, wobei *smg* die scheinbare Modellgüte und *mg* die Modellgüte auf den Testdaten bezeichnet. Weitere Gewichte können die Häufigkeit der Aufnahme in die Trainingsdaten berücksichtigen.

***Bootstrap*-Variante II: .632+**

Eine Erweiterung dieser Variante des Bootstrappings ist die .632+ Variante, bei der die Gewichte vom Maß der Überanpassung des Modells abhängen (Efron & Tibshirani 1997). Die geschätzte Modellgüte errechnet sich dann als $(1-w) \cdot smg + w \cdot mg$, mit $w = 0,632/(1 - 0,368 \cdot R)$, wobei R die relative Überanpassung bezeichnet (vgl. Steyerberg et al. 2001).

Großer Optimismus, also eine hohe Differenz im Zähler, ergibt hohe Werte von R, was bedeutet, dass dann w nahe bei 1 liegt und die Modellgüte fast ausschließlich durch die Güte auf den Testdaten bestimmt wird. Bei geringer Überanpassung liegt R nahe bei 0, w nahe bei 0,632 und die geschätzte Modellgüte entspricht der mittels der .632-Variante errechneten.

Die Motivation für die Entwicklung des .632-*Bootstraps* und anschließend des .632+*Bootstraps* war jeweils, dass die vorher bestehenden Verfahren bei Klassifikationsmethoden, die zu Überanpassung neigen, eine optimistische Verzerrung zeigten.

5.2.2 Externe Validierung

Bei der externen Validierung wird ein unabhängiger Datensatz zur Überprüfung der Modellgüte verwendet. Dieser kann im einfachsten Fall aus dem Originaldatensatz zurückgehalten werden (*hold-out* Verfahren), was dem *Data-Splitting* entspricht (s.o.). Günstigenfalls liegen „wirklich" unabhängige Daten vor, die nach demselben Verfahren erhoben wurden und für die viele der in Fielding & Bell (1997) oder Reineking & Schröder (2004b, in diesem Band) aufgeführten Kriterien berechnet werden können.

Wenn die unabhängigen Daten aus anderen Regionen oder Untersuchungsjahren stammen, kann durch diese Form der Validierung auch der Gültigkeitsbereich des Modells evaluiert werden. Ebenso lässt sich dadurch das Auftreten lokaler Adaptationen (Kuussaari et al. 2000) oder untypischer Untersuchungsjahre (Kindvall 1995) analysieren. Brzeziecki et al. (1993), Capen et al. (1986), Guisan et al. (1998, 1999), Fielding & Haworth (1995), Schröder (1997) sowie Zimmermann & Kienast (1999) zeigen Beispiele der externen Validierung von Habitatmodellen.

Randomisierungsverfahren zur Abschätzung der Überanpassung

Harrell (2001) schlägt ein Randomisierungsverfahren vor, das zur heuristischen Abschätzung der Überanpassung (*overfitting*) geeignet ist. Hierbei wird die Responsevariable zufällig permutiert – d.h. Präsenzen und Absenzen zufällig auf die Datensätze verteilt – und dann ein Modell mit unveränderten Prädiktorvariablen und permutierter Responsevariablen geschätzt. Die Modellgüte wird sowohl für diesen Trainingsdatensatz als auch für den Originaldatensatz mit unveränderter Responsevariable bestimmt. Wenn Überanpassung kein Problem darstellt, sollten beide Gütekriterien ungefähr dieselben Werte aufweisen und eine schlechte Modellqualität anzeigen. Je besser das Modell an die zufällig permutierte Responsevariable angepasst werden kann, desto ausgeprägtere Überanpassung ist zu erwarten.

Räumliche und zeitliche Übertragbarkeit

Stehen Datensätze aus Erhebungen in verschiedenen Untersuchungsgebieten und -jahren zur Verfügung, so kann durch externe Validierung die Übertragbarkeit von Habitatmodellen untersucht und ihr Gültigkeitsbereich festgelegt werden.

In einigen Arbeiten fokussieren die Autoren eher auf die Robustheit der geschätzten Modelle (Freeman et al. 1997, vgl. z.B.), vergleichbar mit dem Einsatz des

Bootstrapping bei der Variablenselektion. Hierbei werden die Datensätze eines Jahres oder Untersuchungsgebiets jeweils als Trainingsdatensätze verwendet. Untersucht wird dann, ob in diese Originalmodelle dieselben Prädiktorvariablen mit vergleichbaren Abhängigkeiten eingehen, d.h. dieselben Vorzeichen für die Regressionskoeffizienten geschätzt werden bzw. diese dieselben Werte aufweisen (vgl. Strobach 2002).

Weitere Analysen zur Übertragbarkeit fokussieren auf den Aspekt der Diskriminanz. Dabei geht es um die Frage, ob und mit welcher Qualität das Modell in der Lage ist, die Inzidenz für Flächen anderer Untersuchungsgebiete oder -jahre vorherzusagen.

Fielding & Bell (1997) oder (Reineking & Schröder 2004a, in diesem Band) führen einige Gütekriterien an, mit denen die Klassifikationsgüte eines Modells in Abhängigkeit vom Klassifikationsschwellenwert bestimmt werden kann. Mit diesen wird dann die Güte der Modellübertragung bestimmt. Derzeit wird häufig Cohen's Kappa (Cohen 1960; Monserud & Leemans 1992) verwendet, weil κ nicht so stark wie andere schwellenwertabhängige Gütemaße – z.B. der Anteil korrekter Klassifikationen – von der Prävalenz abhängt (Manel et al. 2001) und die Klassifikationsgüte eines Zufallsmodells berücksichtigt. Uns ist aber kein Beispiel aus der Literatur bekannt, in dem für eines dieser Gütekriterien ein Signifikanztest zur Übertragbarkeit durchgeführt wird.

Thomas & Bovee (1993); Freeman et al. (1997) sowie Schröder & Richter (2000) führen Signifikanztests für Klassifikationsmatrizen durch, in denen beobachtete Vorkommen und Nichtvorkommen der vorhergesagten Inzidenz, über die Anwendung eines Klassifikationsschwellenwertes (*cutoff-value*) abgeleitet aus den geschätzten Vorkommenswahrscheinlichkeiten, gegenüber gestellt werden (vgl. Reineking & Schröder 2004a).

Mittels eines *G*-Tests – Thomas & Bovee (1993) sowie Freeman et al. (1997) verwenden den traditionellen χ^2-Test (Unterschiede und Vorschläge zur adäquaten Verwendung s. Sokal & Rohlf 1995, S. 729) – wird der Zusammenhang zwischen Daten und Prognosen in den Klassifikationsmatrizen überprüft. Getestet wird die Nullhypothese, dass geeignete (Vorhersage: Vorkommen) und ungeeignete Flächen (Vorhersage: Nichtvorkommen) in demselben Verhältnis besetzt sind. Überschreitet die aus der Belegung der Klassifikationsmatrix berechnete Teststatistik einen kritischen Wert, so ist die Nullhypothese zu verwerfern, d.h., dass auf geeigneten Flächen proportional häufiger Vorkommen beobachtet werden. Kann die Nullhypothese für eine Modellübertragung nicht verworfen werden, so ist das übertragene Modell nicht besser als ein Zufallsmodell und gilt damit als nicht übertragbar. Dieses Verfahren hat den Nachteil, dass sein Ergebnis vom gewählten Klassifikationsschwellenwert abhängt (vgl. Reineking & Schröder 2004a, in diesem Band).

Einen vergleichbaren Signifikanztest der Modellübertragung, der unabhängig von der getroffenen Wahl des Klassifikationsschwellenwertes ist, wird bei Schröder (2000) aus ROC-Kurven abgeleitet. ROC-Kurven erhält man durch Auftragung der Sensitivität – Anteil der korrekten Vorkommensprognosen – über dem Term (1−Spezifizität) – Anteil der falschen Nichtvorkommensprognosen – für alle denkbaren Klassifikationsschwellenwerte (s. Hanley & McNeil 1982; Reineking & Schröder 2004a; Zweig & Campbell 1993). Nach Beck & Shultz (1986) testet die in Gleichung 5.1 gezeigte Statistik die Nullhypothese, dass sich die Fläche unter der ROC-Kurve (*Area Under Curve*, *AUC*) nicht signifikant von einem vorher festzulegenden kritischen *AUC*-Wert unterscheidet.

$$Z = (AUC - AUC_{krit})/SE_{AUC} \qquad (5.1)$$

mit *AUC*: Fläche unter der ROC-Kurve; AUC_{krit}: kritischer *AUC*-Wert, z.B. $AUC_{krit} = 0{,}7$; SE_{AUC}: Standardfehler der Fläche unter der ROC-Kurve.

Da $Z \sim N(0,1)$ gilt (Beck & Shultz 1986), z also standardnormalverteilt ist, kann man durch den Vergleich mit den Quantilen der Standardnormalverteilung Aussagen darüber treffen, ob die Fläche unter der ROC-Kurve signifikant vom kritischen *AUC*-Wert – etwa eines Zufallsmodells ($AUC_{krit} = 0{,}5$) oder eines akzeptablen Modells ($AUC_{krit} = 0{,}7$ vgl. Hosmer & Lemeshow 2000) – abweicht. Der Vorteil dieses Signifikanztests ist, dass der gesamte Bereich möglicher Klassifizierungen berücksichtigt wird und sich dadurch unterschiedliche Prävalenzen in den Trainings- und Testdatensätzen weniger stark auf die Bewertung der Übertragbarkeit auswirken. Alternativ kann auch nach Berechnung des Konfidenzintervalls für *AUC* angegeben werden, ob es AUC_{krit} bei vorgegebener Konfidenzzahl überdeckt oder nicht.

Allerdings ist anzumerken, dass das in vielen Statistikprogrammen zur Berechnung des Standardfehlers des *AUC* eingesetzte Verfahren nach DeLong et al. (1988) nicht in jedem Fall geeignet ist, vor allem bei kleinen Datensätzen und hohen *AUC*-Werten. Obuchowski & Lieber (1998) geben in diesem Zusammenhang einen Mindeststichprobenumfang von 200 Fällen an.

Schwerer wiegt, dass das mit diesem Verfahren berechnete (asymptotische, auf Approximation einer Normalverteilung beruhende) Konfidenzintervall symmetrisch um den *AUC* liegt, während es eigentlich asymmetrisch ist, vor allem dann, wenn *AUC* sehr hohe Werte nahe 1 annimmt. So kann das Konfidenzintervall Werte größer 1 umfassen (vgl. Obuchowski & Lieber 2002; Vida 2001).

Diese Probleme treten nicht auf, wenn man zur Berechnung des Konfidenzintervalls *Resampling-*

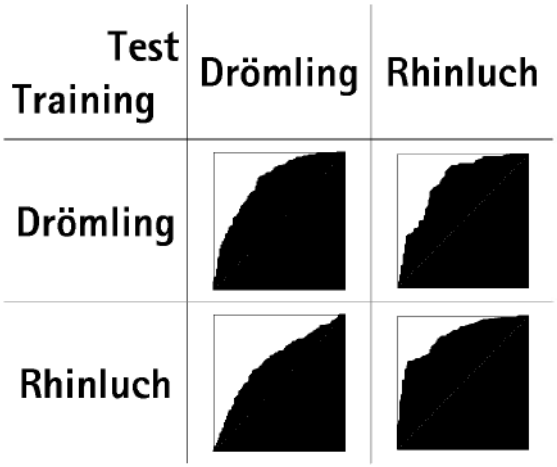


Abb. 5.2. ROC-Kurven für Original- und räumliche übertragene Habitatmodelle für die Kurzflügelige Schwertschrecke *Conocephalus dorsalis* in zwei Niedermooren, Drömling und Rhinluch (aus Schröder 2000). Alle vier AUC-Werte sind signifikant von AUC_{krit}= 0,5 verschieden.

Verfahren verwendet. Während Schröder (2003) unter http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html ein Programm zur Verfügung stellt, welches u.a. Konfidenzintervalle für *AUC* mittels der *Bootstrap*-Perzentilmethode berechnet, die u.U. eine Verzerrung aufweisen können, ermöglicht die aus dem medizinischen Bereich stammende Software AccuROC – zu beziehen unter http://www.accumetric.com/accurocw.htm– verschiedene Verfahren zur Berechnung des Intervalls (Vida 2001). Ein weiteres, auf der Mann-Whitney-Statistik beruhendes Verfahren, stellt Mee (1990) vor.

Beispiele für Untersuchungen hinsichtlich der zeitlichen Übertragbarkeit von Habitatmodellen finden sich u.a. bei Dennis & Eales (1999); Rice et al. (1986); Schröder & Richter (2000) sowie unter Verwendung der beiden oben beschriebenen Übertragbarkeitstests bei Schröder (2000). Die räumliche Übertragbarkeit ihrer Modelle untersuchen z.B. Fleishman et al. (2003); Freeman et al. (1997); Glozier et al. (1997); Lavers & Haines-Young (1997); Leftwich et al. (1997); Özesmi & Mitsch (1997); Schröder & Richter (2000); Schröder (2000, s. Abb. 5.2) sowie Thomas & Bovee (1993).

Bozek & Rahel (1992); Fielding & Haworth (1995); Fleishman et al. (2003) sowie Schröder & Richter (2000) und Schröder (2000) leiten aus ihren Untersuchungen auch Aussagen zum gültigen Anwendungsbereich ihrer Habitatmodelle ab. Bonn & Schröder (2001) wenden dieses Verfahren an, um die Übertragbarkeit von Habitatmodellen einzelner Arten auf andere Arten und Artengruppen zu testen und damit den Schirmeffekt einzelner Zielarten nachzuweisen.

5.2.3 Vergleich der Validierungsverfahren und Empfehlung

Data-Splitting ist sicherlich das einfachste Validierungsverfahren, das aber nur dann sinnvoll ist, wenn sehr große Datensätze – z.B. auf der Basis von Fernerkundungsdaten – vorhanden sind Guisan & Zimmermann (2000). Erst mit der Verfügbarkeit ausgereifter Software finden die rechenintensiven Resamplingverfahren der internen Validierung häufigere Anwendung (Reineking & Schröder 2003). Guisan & Zimmermann (2000) schlagen vor, die interne Validierung als Ergänzung zur externen Validierung zu verwenden, da erstere stärker auf den Aspekt der unverzerrten Schätzung der Modellgüte zielt. Dagegen stellen Lehmann et al. (2002) zur Diskussion, ob externe Validierung der internen in allen Fällen vorzuziehen ist, da durch die Verwendung unabhängiger Daten die Gefahr steigt, verschiedene *Sampling*-Strategien anstelle von verschiedenen Modellen zu vergleichen. Auch steigt durch die Berücksichtigung unabhängiger Daten das Risiko, durch Kartierungsfehler, unadäquate Kartenauflösungen usf. neue Unsicherheiten in die Analyse zu tragen (Zimmermann & Kienast 1999). Was interne Validierungsverfahren allerdings nicht liefern können, ist die Beurteilung des Gültigkeits- oder Anwendungsbereiches eines Modells, die nur auf Grundlage anderer Erhebungen aus anderen Untersuchungsgebieten und/oder -jahren erfolgen kann. Ob dabei die zeitlichen Abhängigkeiten eine Untersuchung der zeitlichen Übertragbarkeit verfälschen, muss untersucht werden.

In ihrer vergleichenden Studie zeigen Steyerberg et al. (2001), dass in ihrem Fall *Bootstrapping* den anderen Verfahren – *Data-Splitting* und Kreuzvalidierung mit verschiedenen Aufteilungen – bei der Bewertung von logistischen Regressionsmodellen überlegen ist und zu weniger verzerrten und weniger variablen Einschätzungen der Modellgüte führt. Auch die *Bootstrap*-Varianten (.632 und.632+) schnitten schlechter ab als das reguläre *Bootstrapping*. In einer Studie, in der anhand verschiedener Klassifikationsprobleme der .632*Bootstrap*, *Leave-one-Out* Kreuzvalidierung (mit m = 1, 2, oder 5) sowie 2-, 5-, 10 und 20-fache Kreuzvalidierung für zwei Klassifikationsverfahren verglichen wurden, kommt Kohavi (1995) zum Schluss, dass stratifizierte zehnfache Kreuzvalidierung die beste Methode für die Modellwahl ist.

Es hängt also von den Eigenschaften des statistischen Verfahrens und von den Daten ab, wie gut die verschiedenen Methoden abschneiden. Keines der vorgestellten Validierungsverfahren ist unter allen Umständen das Beste.

Optimal wäre es, wenn man Konfidenzintervalle für alle validierten Kenngrößen hätte. Uns sind allerdings keine Anwendungsbeispiele bekannt, in denen das bei interner Validierung durchgeführt worden wäre. Unse-

re Empfehlung ist es, auf jeden Fall eine interne Validierung der Habitatmodelle durchzuführen, um realistische Schätzwerte für die Modellgüte hinsichtlich Kalibrierung und Diskriminanz zu erhalten. Für Habitatmodelle, die mittels logistischer Regression geschätzt wurden, sind der gewöhnliche *Bootstrap* oder die zehnfache Kreuzvalidierung geeignete Verfahren. Wenn Habitatmodelle in der Naturschutzpraxis eingesetzt werden sollen, sollte vorher ihr Gültigkeitsbereich untersucht worden sein. Dafür eignet sich in unseren Augen am besten, analoge Datensätze in anderen Untersuchungsgebieten zu erheben. Die Güte des Modells auf diesen Daten kann dann anhand verschiedener Gütekriterien überprüft werden, für die jeweils Konfidenzintervalle angegeben werden sollten. Bei größeren Datensätzen kann die räumliche oder raumzeitliche Übertragbarkeit der Modelle durch den ROC-Signifikanztest überprüft werden.

5.3 Danksagung

Die Autoren danken Hans-Peter Bäumer, Universität Oldenburg, für die hilfreichen Kommentare und Anmerkungen zum Manuskript.

Literaturverzeichnis

Beck, J. R. & Shultz, E. K. 1986. The use of ROC curves in test performance evaluation. *Archives of pathology and laboratory medicine*, 110:13–20.

Bio, A. M. F., De Becker, P., De Bie, E., Huybrechts, W. & Wassen, M. 2002. Prediction of plant species distribution in lowland river valleys in Belgium: modelling species response to site conditions. *Biodiversity and Conservation*, 11(12):2189–2216.

Bonn, A. & Schröder, B. 2001. Habitat models and their transfer for single and multi species groups: a case study of carabids in an alluvial forest. *Ecography*, 24:483–496.

Boyce, M. S., Vernier, P. R., Nielsen, S. E. & Schmiegelow, F. K. A. 2002. Evaluating resource selection functions. *Ecological Modelling*, 157:281–300.

Bozek, M. A. & Rahel, F. J. 1992. Generality of microhabitat suitability models for young Colorado river cutthroat trout (*Oncorhynchus clarki pleuriticus*) across sites and among years in Wyoming streams. *Canadian journal of fisheries and aquatic sciences*, 49:552–564.

Brooks, R. P. 1997. Improving habitat suitability index models. *Wildlife Society bulletin*, 25(1):163–167.

Brzeziecki, B., Kienast, F. & Wildi, O. 1993. A simulated map of the potential natural forest vegetation of Switzerland. *Journal of Vegetation Science*, 4(4):499508.

Capen, D. E., Fenwick, J. W., Inkley, D. B. & Boynton, A. C. 1986. Multivariate models of songbird habitat in New England forests. In Verner, J., Morrison, M. L. & Ralph, C. J., editors, *Wildlife 2000: modeling habitat relationships of terrestrial vertebrates.*, pages 171–177. University of Wisconsin Press, Madison.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society A*, 158(3):419–466.

Clark, J. & Lewis, J. 1983. A validity test of a habitat suitability index model for clapper rail. *Proceedings of the Annual Conference of the Southeastern Association of Fish and Wildlife Agencies*, 37:95–102.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametetric approach. *Biometrics*, 44(3):837–845.

Dennis, R. L. & Eales, H. T. 1999. Probability of site occupancy in the large heath butterfly *Coenonympha tullia* determined from geographical and ecological data. *Biological Conservation*, 87:295–302.

Efron, B. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331.

Efron, B. & Tibshirani, R. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, London.

Efron, B. & Tibshirani, R. 1997. Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association*, 92:548–560.

Fielding, A. H. & Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24:38–49.

Fielding, A. H. & Haworth, P. F. 1995. Testing the generality of bird-habitat models. *Conservation biology*, 9(6):1466–1481.

Fleishman, E., MacNally, R. & Fay, J. P. 2003. Validation tests of predictive models of butterfly occurrence based on environmental variables. *Conservation biology*, 17(3):806–817.

Freeman, M. C., Bowen, Z. H. & Crance, J. H. 1997. Transferability of habitat suitability criteria for fishes in warmwater streams. *North American Journal of Fisheries Management*, 17(1):20–31.

Glozier, N. E., Culp, J. M. & Scrimgeour, G. J. 1997. Transferability of habitat suitability curves for a benthic minnow, rhinichthys cataractae. *Journal of freshwater ecology*, 12(3):379–394.

Guisan, A., Theurillat, J.-P. & Kienast, F. 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*, 9(1):65–74.

Guisan, A., Weiss, S. B. & Weiss, A. D. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology*, 143(1):107–122.

Guisan, A. & Zimmermann, N. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135:147–186.

Hanley, J. A. & McNeil, B. J. 1982. The meaning and use of the area under a ROC curve. *Radiology*, 143:29–36.

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.

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

Huberty, C. 1994. *Applied Discriminant Analysis*. Wiley, New York.

Hurvich, C. M. & Tsai, C. L. 1990. The impact of model selection on inference in linear regression. *American Statistician*, 44:214–217.

James, F. C. & McCulloch, C. E. 2002. Predicting species presence and abundance. In Scott, J. M., Heglund, P. J., Morrison, M., Haufler, J. B. & Wall, W. A., editors, *Predicting species occurrences: issues of accuracy and scale*, pages 461–465. Island Press, Washington.

Kindvall, O. 1995. The impact of extreme weather on habitat preference and survival in a metapopulation of the bush cricket *Metrioptera bicolor* in Sweden. *Biological Conservation*, 73(1):51–58.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, Montreal.

Kuussaari, M., Hanski, I. & Singer, M. 2000. Local specialization and landscape-level influence on host use in an herbivorous insect. *Ecology*, 81(8):21772187.

Lavers, C. P. & Haines-Young, R. H. 1997. Displacement of dunlin *Calidris alpina schinzii* by forestry in the Flow Country and an estimate of the value of moorland adjacent to plantations. *Biological Conservation*, 79(1):87–90.

Leftwich, K. N., Angermeier, P. L. & Dolloff, C. A. 1997. Factors influencing behavior and transferability of habitat models for a benthic stream fish. *Transactions of the American Fisheries Society*, 126(5):725–734.

Lehmann, A., Overton, J. M. & Austin, M. P. 2002. Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity and Conservation*, 11(12):2085–2092.

Levins, R. O. 1966. The strategy of model building in population biology. *American Scientist*, 54:421–431.

Manel, S., Dias, J. M. & Ormerod, S. J. 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a himalayan river bird. *Ecological Modelling*, 120:337–348.

Manel, S., Williams, H. C. & Ormerod, S. J. 2001. Evaluating presenceabsence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931.

Manly, B. F. J. 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Texts in Statistical Science. Chapman & Hall, London, 2nd edition.

Mee, R. W. 1990. Confidence intervals for probabilities and confidence regions based on a generalization of the Mann-Whitney statistic. *Journal of the American Statistical Association*, 85(411):793–800.

Monserud, R. A. & Leemans, R. 1992. Comparing global vegetation maps with Kappa statistic. *Ecological Modelling*, 62:275–293.

Morrison, M. L., Marcot, B. G. & Mannan, R. W. 1998. *Wildlife-Habitat Relationships - Concepts and Applications*. University of Wisconsin Press, Madison, 2nd edition.

Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Noon, B. R. 1986. Biometric approaches to modeling - the researcher's viewpoint. In Verner, J., Morrison, M. L. & Ralph, C. J., editors, *Wildlife 2000: modeling habitat relationships of terrestrial vertebrates.*, pages 197–201. University of Wisconsin Press, Madison.

Obuchowski, N. A. & Lieber, M. L. 1998. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiology*, 5(8):561–571.

Obuchowski, N. A. & Lieber, M. L. 2002. Confidence bounds when the estimated ROC area is 1.0. *Academic Radiology*, 9(5):526–530.

Oppel, S., Schaefer, H. M., Schmidt, V. & Schröder, B. 2004. Habitat selection by the Pale-headed brush-finch, *Atlapetes pallidiceps*, in southern Ecuador: implications for conservation. *Biological Conservation*, in press.

Osborne, P. E. & Suarez-Seoane, S. 2002. Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling*, 157(2-3):249–259.

Özesmi, U. & Mitsch, W. J. 1997. A spatial habitat model for the marsh-breeding red-winged blackbird (*Agelaius phoeniceus* l.) in coastal Lake Erie wetlands. *Ecological Modelling*, 101(2,3):139–152.

Peppler-Lisbach, C. & Schröder, B. 2004. Predicting the species composition of mat-grass communities (Nardetalia) by logistic regression modelling. *Journal of Vegetation Science*, submitted.

Prosser, D. J. & Brooks, R. P. 1998. A verified habitat suitability index for the Louisiana Waterthrush. *Journal of field ornithology*, 69(2):288–298.

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. 2004a. Gütemaße für Habitatmodelle. *UFZ-Bericht*, 9/2004:27–38.

Reineking, B. & Schröder, B. 2004b. Variablenselektion. *UFZ-Bericht*, 9/2004:39–46.

Rice, J. C., Ohmart, R. D. & Anderson, B. W. 1986. Limits in a data-rich model: modeling experience with habitat management on the Colorado River. In Verner, J., Morrison, M. L. & Ralph, C. J., editors, *Wildlife 2000: Modeling Habitat Relationships of Terrestrial Vertebrates.*, pages 79–86. University of Wisconsin Press, Madison.

Roloff, G. J. & Kernoban, B. J. 1999. Habitat suitability and evaluation - evaluating reliability of habitat suitability index models. *Wildlife Society bulletin*, 27(4):973–985.

Rykiel, E. J., J. 1996. Testing ecological models: the meaning of validation. *Ecological Modelling*, 90(3):229–244.

Schamberger, M. L. & O'Neil, L. J. 1986. Concepts and constraints of habitat-model testing. In Verner, J., Morrison, M. L. & Ralph, C. J., editors, *Wildlife 2000: Modeling Habitat Relationships of Terrestrial Vertebrates*, pages 5–10. University of Wisconsin Press, Madison.

Schröder, B. 1997. Fuzzy Logik und klassische Statistik - ein kombiniertes Habitateignungsmodell für *Conocephalus dorsalis* (Latreille 1804) (Orthoptera: Tettigoniidae). *Verhandlungen der Gesellschaft für Ökologie*, 27:219–226.

Schröder, B. 2000. *Zwischen Naturschutz und Theoretischer Ökologie: Modelle zur Habitateignung und räumlichen Populationsdynamik für Heuschrecken im Niedermoor*. Doktorarbeit, TU Braunschweig.

Schröder, B. 2003. *ROC & AUC-Calculation - evaluating the predictive performance of habitat models*. http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html, Potsdam.

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

Schröder, B. & Richter, O. 1999/2000. Are habitat models transferable in space and time? *Zeitschrift für Ökologie und Naturschutz*, 8:195–205.

Scott, J. M., Heglund, P. J., Morrison, M., Haufler, J. B. & Wall, W. A., editors 2002. *Predicting Species Occurrences: Issues of Accuracy and Scale*. Island Press.

Sokal, R. R. & Rohlf, F. J. 1995. *Biometry*. Freeman, New York, 3rd edition.

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.

Strobach, T. 2002. *Räumlich explizite Quantifizierung von Wasser- und Stoffhaushaltsparametern zur Vegetationsprognose durch digitale Terrainanalyse: eine ArcView Extension*. Diplomarbeit, University of Oldenburg.

Thomas, J. A. & Bovee, K. D. 1993. Application and testing of a procedure to evaluate transferability of habitat suitability criteria. *Regulated Rivers: Research and Management*, 8(3):285–294.

Verbyla, D. L. & Litvaitis, J. A. 1989. Resampling methods for evaluation of classification accuracy of wildlife habitat models. *Environmental Management*, 13(6):783–787.

Vida, S. 2001. AccuROC for Windows 95/98/NT Version 2.5. Technical report, McGill University Health Center, Department of Psychiatry.

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.

Zimmermann, N. & Kienast, F. 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *Journal of Vegetation Science*, 10:469–482.

Zweig, M. H. & Campbell, G. 1993. Receiver-Operating Characteristic (ROC) Plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577.

5.4 Datenblatt

5.4.1 Software

Die Autoren empfehlen die Verwendung von R (*free software*) oder S-Plus (Insightful); bestimmte Verfahren lassen sich auch mit SPSS und vielen anderen Statistikprogrammen umsetzen.

5.4.2 Webresources

- R unter www.r-project.org
- Ein Delphi-Programm zur Berechnung von ROC-Kurven, AUC-Werten und optimalen Klassifikationsschwellenwerten sowie zur Durchführung des im Text beschriebenen Übertragbarkeitstests wird von B. Schröder unter http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html zur Verfügung gestellt.

5.4.3 Kommentierte Literatur

S. umfangreiche Angaben zu beispielhaften Veröffentlichungen in den einzelnen Kapiteln.

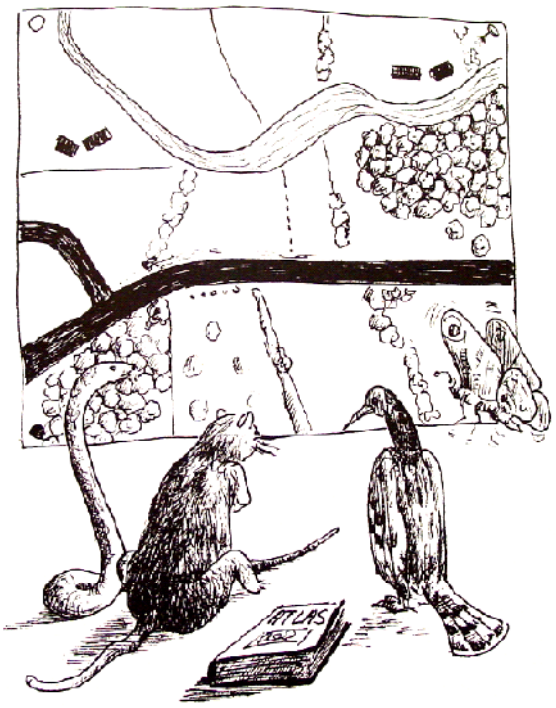
HABITATMODELLE

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