3

Gütemaße für Habitatmodelle

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

3.1 Einleitung

Sobald ein Habitatmodell an die beobachteten Daten angepasst worden ist, stellen sich Fragen nach der Modellgüte: Wie gut ist das Modell? Ist das Modell gut genug? Diese Fragen, insbesondere die zweite, lassen sich nicht allein durch die Berechnung statistischer Gütemaße beantworten, sondern die Beantwortung hängt wesentlich vom Anwendungskontext ab. Statistische Gütemaße helfen jedoch bei der Beurteilung, indem sie unterschiedliche Aspekte des Modells quantifizieren; zu den bekanntesten Gütemaßen zählen die Fehlerrate bei Modellen, die Vorkommenswahrscheinlichkeiten vorhersagen, oder das R^2 bei linearen Regressionsmodellen.

In diesem Beitrag stellen wir verschiedene Gütemaße vor, die sich für die Beurteilung von Habitatmodellen eignen. Wir betrachten die Situation, in der die Beobachtungen *binär* sind, d.h. in eine von zwei Klassen fallen (Vorkommen/Nichtvorkommen bzw. Erfolg/Nichterfolg). Modelle für solche Beobachtungen sagen entweder *Wahrscheinlichkeiten* für das Vorkommen, oder direkt das Vorkommen bzw. Nichtvorkommen voraus. Die meisten statistischen Verfahren in der Habitatmodellierung zählen zur ersten Gruppe, zum Beispiel logistische Regression, *Generalized additive models* (GAMs) oder neuronale Netze. Vertreter der zweiten Gruppe sind lineare Diskriminanzanalyse und Entscheidungsbäume (vgl. Schröder und Reineking 2004a).

Die Auswahl an Gütemaßen ist sehr groß und die korrespondierende Literatur umfangreich. Wahrscheinlichkeitsvorhersagen spielen in der Medizin und der Meteorologie eine große Rolle, daher stammt ein wichtiger Teil der Literatur aus diesen Fachgebieten (Harrell 2001; Harrell et al. 1996; Steyerberg et al. 2000, 2001; Brier 1950; Dawid 1986; Murphy & Winkler 1992). Für den Bereich der Habitatmodellierung sind die Übersichten von Fielding & Bell (1997); Manel et al. (2001) sowie von Pearce & Ferrier (2000) hervorzuheben.

Wir beschränken uns hier auf solche Maße, die entweder besonders häufig verwendet werden oder die wir für besonders vielversprechend halten. Insbesondere diskutieren wir: Fehlerrate, Sensitivität, Spezifizität, *AUC* (Fläche unter der *Receiver Operating Characteristic* (ROC) Kurve), Cohen's Kappa κ, R^2 nach Nagelkerke (1991), das im folgenden mit R^2_N bezeichnet wird, Brier Index, sowie Achsenabschnitt und Steigung der Kalibrierungsgeraden.

Bevor wir mit den Einzeldarstellungen beginnen, stellen wir einen Interpretationsrahmen für die Beurteilung von Wahrscheinlichkeitsvorhersagen vor. Er wurde von Murphy & Winkler (1992) im Rahmen der Vorhersage von Niederschlagsereignissen entwickelt. In der Diskussion gehen wir dann auf die Problematik der Schätzung der Gütemaße ein. Abschließend geben wir einige Empfehlungen für die Auswahl eines geeigneten Satzes von Gütemaßen.

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.

3.2 Die diagnostische Beurteilung von Wahrscheinlichkeitsvorhersagen

3.2.1 Grundidee

Ausgangspunkt für den Ansatz von Murphy & Winkler (1992) ist die gemeinsame Verteilung (*joint distribution*) der Vorhersagen und der beobachteten Vorkommen. Sie beschreibt, wie wahrscheinlich die verschiedenen Kombinationen von vorhergesagter Wahrscheinlichkeit und beobachtetem Vorkommen sind. Diese Verteilung selbst ist unbekannt; wir können lediglich versuchen, sie anhand der beobachteten Vorkommen und der dazugehörigen Vorhersagen zu rekonstruieren bzw. mit einigen Kennzahlen zu charakterisieren. Man kann alle der im folgenden diskutierten Gütemaße so auffassen, dass sie verschiedene Aspekte dieser gemeinsamen Verteilung beschreiben. Sämtliche dieser Gütemaße ließen sich aus dieser gemeinsamen Verteilung berechnen (vgl. Tab. 3.1).

Wir bezeichnen im folgenden die Vorhersagen mit f, die Beobachtungen mit y und die gemeinsame Verteilung mit $p(f, y)$. Die Werte von f liegen alle im Bereich von [0, 1], während y entweder den Wert 0 (Nichtvorkommen) oder 1 (Vorkommen) annimmt.

Murphy und Winkler schlagen nun vor, die gemeinsame Verteilung $p(f, y)$ in das Produkt aus einer bedingten Verteilung und einer Randverteilung zu zerlegen. Dies kann auf zwei Weisen geschehen. Zum einen kann man sich die bedingte Wahrscheinlichkeit der Vorkommen, gegeben die Vorhersagen anschauen, die dann mit der Verteilung der Vorhersagen multipliziert wird: $p(f,y) = p(y|f)p(f)$. Diese Zerlegung nennen Murphy und Winkler Kalibrierungs-Verfeinerungs-Faktorisierung (*calibration-refinement factorization*). Analog dazu ist die *likelihood-base rate* Faktorisierung gegeben durch: $p(f,y) = p(f|y)p(y)$, d.h. es wird die Verteilung der Vorhersagen, gegeben die tatsächlichen Vorkommen betrachtet, multipliziert mit der Verteilung der Vorkommen ($p(y)$ ist bernoulliverteilt).

Die Kalibrierung entspricht der bedingten Verteilung der Beobachtungen, gegeben die Vorhersagen, $p(y \mid f)$. Bei einem gut kalibrierten Modell entspricht der Anteil der beobachteten Vorkommen jeweils der vorhergesagten Wahrscheinlichkeit, d.h. $p(y=1| f) = f$. Einen guten Eindruck von der Kalibrierung eines Modells erhält man, wenn man die empirischen relativen Häufigkeiten der Vorkommen gegen die vorhergesagten Wahrscheinlichkeiten aufträgt, wie in Abb. 3.1. Die erste Diagonale (Identität) entspricht einem perfekt kalibrierten Modell. Wenn ein Modell gut kalibriert ist, dann kann man die vorhergesagten Wahrscheinlichkeiten als zu erwartende relative Häufigkeiten interpretieren. Modelle sind häufig schlecht kalibriert und machen zu extreme Prognosen (*overconfident*). Das bedeutet beispielsweise, dass bei einer vorhergesagten Vorkommenswahrscheinlichkeit von 0,9 lediglich in 80% der Fälle empirisch ein Vorkommen beobachtet wird, oder dass 30% Vorkommen beobachtet werden, obwohl das Modell lediglich eine Vorkommenswahrscheinlichkeit von 0,12 vorhersagt. Wir gehen auf die Kalibrierungskurve noch einmal detaillierter ein, wenn wir die Kalibrierungsmaße diskutieren.

Die Verfeinerung (*refinement*) ist die Randverteilung $p(f)$ der vorhergesagten Wahrscheinlichkeiten. Sie ist ebenfalls in Abb. 3.1 dargestellt und sagt etwas darüber aus, wie nahe die vorhergesagten Wahrscheinlichkeiten bei Null oder Eins liegen.

Die Zerlegung in die bedingte Verteilung der Vorhersagen, gegeben die empirischen Vorkommen bzw. Nichtvorkommen $p(f|y)$ und die Prävalenz $p(y)$(*a priori* Wahrscheinlichkeit eines Vorkommens, *base rate*) ist die Grundlage für die Abschätzung der Diskriminanz des Modells, also der Fähigkeit, zwischen Vorkommen und Nichtvorkommen zu unterscheiden. Das *Likelihood*-Diagramm (Abb. 3.2) stellt zwei Häufigkeitsverteilungen vorhergesagter Wahrscheinlichkeiten dar. Zum einen für die Fälle, in denen tatsächlich ein Vorkommen beobachtet wird, und zum anderen für die Fälle, in denen ein Nichtvorkommen beobachtet wurde. Je stärker die beiden Verteilungen überlappen, desto schlechter ist das Modell in der Lage, Vorkommen von Nichtvorkommen zu unterscheiden. Das *Likelihood*-Diagramm diskutieren wir noch einmal eingehender im Kontext der Maße zur Diskriminanz.

Wenn ein Modell perfekt kalibriert und vollständig verfeinert (*refined*) ist, dann wird für alle tatsächlichen Vorkommen eine Vorkommenswahrscheinlichkeit von 1 vorhergesagt, und für alle tatsächlichen Nichtvorkommen eine Vorkommenswahrscheinlichkeit von null. Ein solches Modell zeigt entsprechend perfekte Diskriminanz, d.h. die Fehlerrate ist 0 und der *AUC* ist 1 (s.u.). Der Umkehrschluss gilt jedoch nicht: Ein Modell mit einer Fehlerrate von 0 muss weder besonders gut kalibriert sein, noch müssen die vorhergesagten Wahrscheinlichkeiten alle 0 oder 1 betragen, oder auch nur nahe 0 oder 1 liegen.

3.2.2 Grenzen des Ansatzes

Sämtliche der im folgenden diskutierten Gütemaße lassen sich in diesen Interpretationsrahmen einordnen. Er hat jedoch Grenzen, von denen wir zwei aufzeigen möchten.

Zum einen tauchen in der Betrachtung keine Bezüge zu den Habitatvariablen mehr auf. Daher lassen sich Modelldefizite nicht hinsichtlich bestimmter Habitateigenschaften verorten.

Zum anderen gehen Nachbarschaftsbeziehungen zwischen den einzelnen Vorhersagen verloren. Diese Nachbarschaftsbeziehungen können beispielsweise räumlicher oder zeitlicher Art sein. Muster räumlicher

Autokorrelation in den Residuen der Modellvorhersagen werden somit nicht betrachtet, ähnliches gilt für zeitliche Autokorrelation, die beispielsweise dann auftreten kann, wenn die Vorhersagen für mehrere Zeitschritte an den gleichen Orten gemacht werden.

3.3 Kleines Bestiarium verschiedener Gütemaße

Im folgenden stellen wir einige der wichtigsten Gütemaße vor. Wir beginnen mit Maßen, die mit den vorhergesagten Wahrscheinlichkeiten arbeiten. Danach kommen Maße, die mit Klassifikationen arbeiten, in der also die Wahrscheinlichkeiten auf Vorkommensprognosen abgebildet wurden. Die Maße werden jeweils in Bezug zu dem oben eingeführten Ansatz von Murphy & Winkler (1992) gesetzt.

3.3.1 Maße, die mit vorhergesagten Wahrscheinlichkeiten arbeiten

Kalibrierung: Steigung und Achsenabschnitt der Kalibrierungsgerade.

Die Kalibrierung misst das Ausmaß, in dem die vorhergesagten Wahrscheinlichkeiten mit den empirischen relativen Häufigkeiten der Vorkommen übereinstimmen. Ein typisches Phänomen von Modellen, die zu sehr an die Daten angepasst wurden (Problem des *overfitting*, vgl. Schröder & Reineking 2004a,b, in diesem Band), ist eine zu hohe anscheinende Prognosesicherheit; die vorhergesagten Wahrscheinlichkeiten sind zu extrem. Kleine Vorkommenswahrscheinlichkeiten sind kleiner als die beobachteten relativen Häufigkeiten, und große Vorkommenswahrscheinlichkeiten sind höher als die beobachteten relativen Häufigkeiten.

Man kann die Fehlkalibrierung quantifizieren, indem man eine Kalibrierungskurve – im logit-Maßstab als Kalibierungsgerade – berechnet. Dabei handelt es sich um eine logistische Regression, die als einzige erklärende Variable den logit der vorhergesagten Wahrscheinlichkeiten verwendet und als zu erklärende Variable die beobachteten Vorkommen und Nichtvorkommen (Steyerberg et al. 2001). Im Fall eines perfekt kalibrierten Modells hat die Kalibrierungskurve die Form einer Geraden (s. Abb. 3.1).

Der Achsenabschnitt dieser Kalibrierungsgeraden gibt an, ob sich die vorhergesagten Wahrscheinlichkeiten im Mittel von der beobachteten Vorkommenshäufigkeit unterscheiden, d.h. ob das Modell das Vorkommen systematisch unter- oder überschätzt. Die Abweichung der Steigung der Kalibrierungsgeraden von 1 (45°-Gerade) misst hingegen, ob die Vorhersagen zu extrem sind, das Modell also beispielsweise eine zu hohe Vorhersagesicherheit vorspiegelt. Dies ist bei Steigungen < 1 der Fall. Mit diesen beiden Größen können natürlich nicht alle Aspekte der Kalibrierung abgedeckt werden; es handelt sich lediglich um eine lineare Näherung an die Kalibrierungseigenschaften des Modells. So kann die Abweichung vorwiegend bei hohen oder bei geringen vorhergesagten Wahrscheinlichkeiten auftreten. Je nachdem, wie wichtig Fehler bei der Vorhersage von Vorkommen oder Nichtvorkommen sind, haben diese verschiedenen Formen fehlender Kalibrierung unterschiedliches Gewicht aus Sicht der Modellanwender.

Wenn man feststellt, dass das Modell schlecht kalibriert ist, gibt es mehrere Möglichkeiten fortzufahren: Bei einer geringen Abweichung kann die Kalibrierungskurve verwendet werden, um die Modellvorhersagen zu korrigieren. Anstelle der vorhergesagten Wahrscheinlichkeiten $\hat{P}$ verwendet man die korrigierten Wahrscheinlichkeiten P_c, die man durch Verwendung des Achsenabschnitts γ_0 und der Steigung γ_1 der logistischen Kalibrierungskurve erhält:

$$P_c = \left[1 + \exp\left(-\gamma_0 - \gamma_1 \ln \frac{\hat{P}}{1-\hat{P}}\right)\right]^{-1} \qquad (3.1)$$

Falls die Abweichungen erheblich sind, sollte ein anderes Modell gewählt werden. Bei Überanpassung ist es sinnvoll, die Modellkomplexität zu beschränken. Dafür eignet sich bei logistischen Regressionsmodellen und bei neuronalen Netzwerken ein Verfahren, das als *shrinkage* oder *regularization* bezeichnet wird. Siehe hierzu auch den Beitrag von Reineking & Schröder (2004) zur Variablenselektion in diesem Band.

Kalibrierung und Refinement: R^2 nach Nagelkerke

Ein Modell sollte gut kalibriert und verfeinert sein. Allein betrachtet, sind die Eigenschaften zwar wünschenswert, aber nicht hinreichend. So ist ein Modell, das immer die gleiche Wahrscheinlichkeit vorhersagt, die der tatsächlichen Prävalenz entspricht, zwar perfekt kalibriert, aber es ist nutzlos für das Trennen von Vorkommen und Nichtvorkommen. Andererseits ist ein Modell, dessen Vorhersagen immer 0 oder 1 sind, zwar perfekt verfeinert und sagt klare Klassenzugehörigkeiten voraus, wenn aber die tatsächlichen Beobachtungen davon abweichen, d.h. das Modell schlecht kalibriert ist, ist es ebenfalls nur eingeschränkt nutzbar.

Der R^2–Wert nach Nagelkerke (1991) (R^2_N) ist ein Maß, das die Kalibrierung und das *Refinement* $p(f)$ eines Modells gemeinsam bewertet. R^2_N ist eine Verallgemeinerung des Bestimmtheitsmaßes R^2, das standardmäßig für lineare Regressionsmodelle verwendet wird und den Anteil der vom Modell repräsentierten Varianz quantifiziert (z.B. Sokal & Rohlf 1995). Ein R^2_N von eins wird erreicht, wenn das Modell nur

Tabelle 3.1. Definition von Aspekten der Vorhersagequalität sowie Angabe relevanter Graphen und Gütemaße. Verändert nach Murphy & Winkler (1992).

Name	Definition	Graphen und Maße
Kalibrierung (*calibration*)	Ausmaß, in dem die vorhergesagten Wahrscheinlichkeiten mit den empirischen Häufigkeiten der Vorkommen übereinstimmen, $p(y\|f)$. Ein spezieller Aspekt ist der *unconditional bias*, der die Abweichung der mittleren vorhergesagten Vorkommenswahrscheinlichkeit von der empirischen Häufigkeit bezeichnet.	Kalibrierungsdiagramm, Achsenabschnitt und Steigung der Kalibrierungsgeraden. R^2 nach Nagelkerke (berücksichtigt auch *refinement*)
Verfeinerung, Läuterung (*refinement*)	Ausmaß, in dem die vorhergesagten Wahrscheinlichkeiten bei null oder eins liegen, $p(f)$	R^2 nach Nagelkerke (berücksichtigt auch Kalibrierung)
Diskriminanz (*discrimination*)	Ausmaß, in dem die Vorhersagen zwischen Vorkommen (y=1) und Nichtvorkommen (y=0) unterscheiden, $p(f\|y)$.	*AUC*, D_{xy}, Fehlerrate, Sensitivität, Spezifizität, *likelihood ratios*
accuracy	Durchschnittliche Übereinstimmung zwischen Vorhersagen und Beobachtungen	Brier Index (quadratische Gewichtung der Fehler), $D = (\text{Modell}\chi^2 - 1)/N$ (logarithmische Gewichtung der Fehler)

Wahrscheinlichkeiten von 0 und 1 vorhersagt (also vollständig *refined* ist) und alle Beobachtungen korrekt vorhergesagt werden (das Modell also perfekt kalibriert ist). In diesem Fall repräsentiert das Modell sämtliche Varianz in den Daten – das Modell macht deterministische und korrekte Vorhersagen. Ein R^2_N von 0 wird erreicht, wenn das Modell konstant die beobachtete Prävalenz vorhersagt. Die Skalierung zwischen diesen beiden Referenzpunkten geschieht über eine logarithmische Gewichtungsregel (*scoring rule*). Ein $R^2_N < 0$ ist möglich.

Wenn LL die Log Likelihood des Modells ist, LL_0 die Log Likelihood des Nullmodells, d.h. des Modells, das für alle Beobachtungen den Wert der Prävalenz voraussagt, und N die Anzahl Beobachtungen, dann ist $R^2{}_N$ gegeben durch:

$$R^2_N = \frac{1 - e^{-\frac{2}{N}(LL - LL_0)}}{1 - e^{\frac{2}{N}LL_0}} \tag{3.2}$$

R^2_N hat bislang nicht die gleiche Verbreitung gefunden wie das Bestimmtheitsmaß der linearen Regression. Eine mögliche Ursache dafür ist, dass es keine vergleichbar eingängige graphische Darstellung der Übereinstimmung von Vorhersagen und Beobachtungen gibt. Ein weiteres, eher psychologisches Problem besteht darin, dass die Werte notorisch klein sind, wesentlich kleiner als die aus der gewöhnlichen linearen Regression bekannten Werte (Hosmer & Lemeshow 2000).

Ein hohes $R^2{}_N$ bedeutet, dass das Modell gut kalibriert ist und dass es gut zwischen Vorkommen und Nichtvorkommen unterscheiden kann.

Diskriminanz: AUC

Der *AUC*-Wert (*Area Under the ROC-Curve*) gibt den Flächeninhalt unter der ROC-Kurve (*Receiver-Operating-Characteristics*, s. Hanley und McNeil 1982, Abb. 3.4) an. Er kann als die Wahrscheinlichkeit interpretiert werden, dass bei einem zufällig gezogenen Paar von Beobachtungen, von denen eine ein Vorkommen und die andere ein Nichtvorkommen ist, das Modell eine höhere Vorkommenswahrscheinlichkeit für das tatsächliche Vorkommen vorhersagt (Fielding & Bell 1997).

ROC-Kurven werden auf die folgende Weise konstruiert: Jede vom Modell vorhergesagte Vorkommenswahrscheinlichkeit wird als Schwellenwert verwendet, so dass alle Beobachtungen mit einer kleineren vorhergesagten Vorkommenswahrscheinlichkeit als dieser Schwellenwert als Nichtvorkommen vorhergesagt werden und alle anderen als Vorkommen. Ausgehend von dieser Klassifikation (s.u.) werden die Sensitivität und die Spezifizität zu diesem Schwellenwert berechnet. Die Sensitivität ist der Anteil der tatsächlichen Vorkommen, der korrekt klassifiziert wurde, während die Spezifizität den Anteil der korrekt vorhergesagten Nichtvorkommen beschreibt (vgl. Abb. 3.5). Die Sensitivitätswerte werden dann gegen die entsprechenden Werte von (1−Spezifizität) aufgetragen (Hanley & McNeil 1982).

AUC kann auf verschiedene Weisen berechnet werden. Zum einen liegt es nahe, eine ROC-Kurve zu konstruieren und die Fläche unter dieser Kurve zu berechnen. Die von uns favorisierte Variante nutzt aus, dass der *AUC* der skalierten Mann-Whitney-U-Statistic bzw. Wilcoxon-Statistik entspricht (Harrell 2001). Anstelle der vorhergesagten Wahrscheinlichkeiten werden deren Ränge verwendet. Dann wird der mittlere Rang der Vorhersagen für die Beobachtungen mit tatsächlichem Vorkommen, $\bar{R}_1$ berechnet. *AUC* lässt sich dann berechnen als:

$$AUC = \frac{1}{N - N_1}\left(\bar{R}_1 - \frac{N_1 + 1}{2}\right) \tag{3.3}$$

wobei N die Gesamtzahl der Beobachtungen und N_1 die Zahl der Vorkommen bezeichnet.

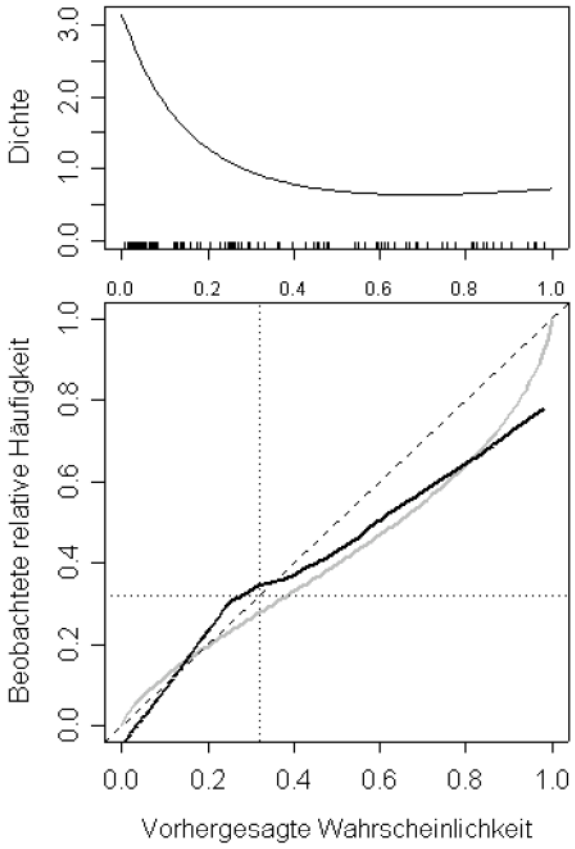


Abb. 3.1. Attributdiagramm zur Visualisierung der Kalibrierungs-*Refinement*-Zerlegung. Im oberen Teil der Graphik wird das *Refinement*, d.h. die Dichte der vorhergesagten Wahrscheinlichkeiten dargestellt. Der untere Teil der Graphik ist eine Kalibrierungskurve, in der die bedingte Verteilung der tatsächlichen Vorkommen für die verschiedenen Vorhersagen $p(y|f)$ aufgetragen wird. Die graue Kurve beschreibt die logistische Kalibrierungskurve. Die durchgezogene schwarze Kurve bezieht sich auf eine loess-Glättungskurve, d.h. eine robuste, lokal gewichtete Regression zwischen den beobachteten Vorkommen und vorhergesagten Wahrscheinlichkeiten, deren Form stärker von den Daten beeinflusst wird als die modellgeleitete logistische Kalibrierungskurve (Cleveland 1981). Die gestrichelte Winkelhalbierende markiert die Kalibrierungsgerade eines perfekt kalibrierten Modells; die strichpunktierten Linien bezeichnen die beobachtete Prävalenz.

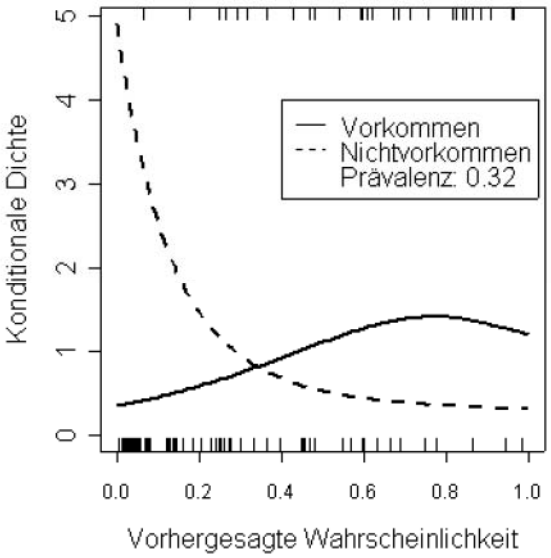


Abb. 3.2. Likelihooddiagramm zur Visualisierung der *Likelihood-Base rate*-Zerlegung. Es zeigt die Dichten der Wahrscheinlichkeitsvorhersagen für die Vorkommen und für die Nichtvorkommen, d.h. die bedingte Verteilung der Vorhersagen gegeben die Vorkommen $p(f|y)$.

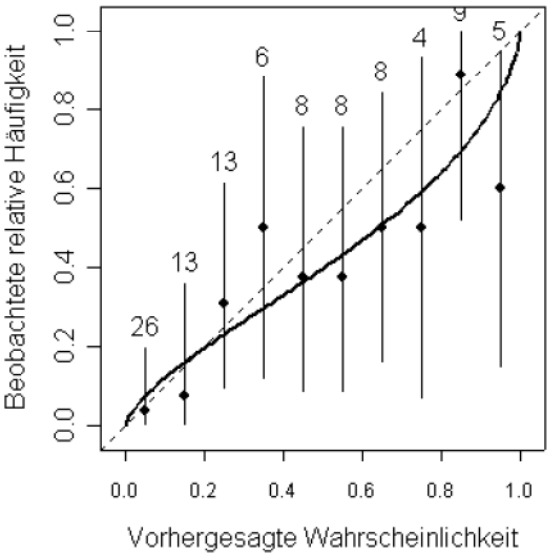


Abb. 3.3. Kalibrierungskurve. Die Idealline ist durch die gestrichelte Winkelhalbierende gegeben. Die durchgezogene Linie gibt die logistische Kalibrierungskurve an. Zusätzlich sind die Beobachtungen eingezeichnet. Die Beobachtungen wurden entsprechend der vorhergesagten Wahrscheinlichkeiten gruppiert, die in zehn Intervalle mit einer Breite von jeweils 0,1 eingeteilt wurden. Die Zahlen bei den Punkten geben jeweils die Anzahl der Beobachtungen an, die in das Intervall fielen, die Striche markieren das 95% Konfidenzintervall, das unter Annahme einer Binomialverteilung berechnet wurde. Die Steigung der Kalibrierungskurve ist kleiner eins, d.h. das Modell macht zu extreme Vorhersagen, es ist *overconfident*.

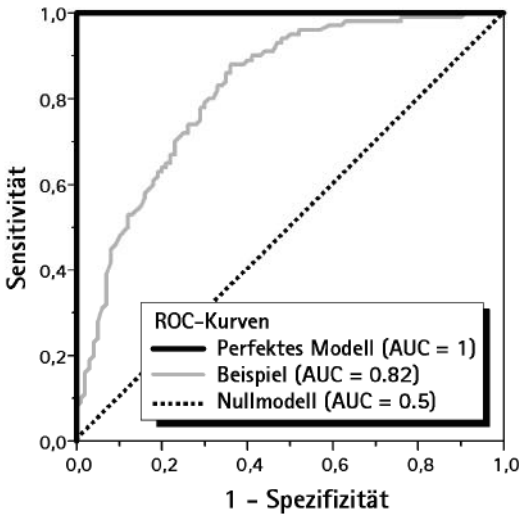


Abb. 3.4. ROC-Kurve.

Anstelle der Bezeichnung *AUC* findet sich in der Literatur auch *concordance index c*. Eng verwandt mit *AUC* ist Somer's D_{xy} (Harrell 2001); es gilt $D_{xy} = 2(AUC - 0{,}5)$. Da *AUC* eine Rangstatistik ist, verändert sich der Wert nicht, wenn die zugrundeliegenden Wahrscheinlichkeitsvorhersagen monoton transformiert werden. Er ist dadurch weitgehend unempfindlich gegenüber schlechter Kalibrierung.

Der bestmögliche Wert von *AUC* ist 1, er entspricht einer perfekten Trennung von Vorkommen und Nichtvorkommen. Ein Wert von 0,5 entspricht einem Modell mit konstant vorhergesagter Wahrscheinlichkeit. Hosmer & Lemeshow (2000) geben folgende Werte zur Einordnung der *AUC*-Werte an:

$0{,}7 \leq AUC < 0{,}8$ = *acceptable*
$0{,}8 \leq AUC < 0{,}9$ = *excellent*
$0{,}9 \leq AUC$ = *outstanding*

Ein *AUC* von 1 wird erreicht, wenn es wenigstens einen Schwellenwert gibt, bei dem der Klassifikationsfehler auf null sinkt. Das Phänomen wird im Englischen als *complete separation* bezeichnet. In diesem Fall lassen sich die Modellparameter bei einer logistischen Regression nicht mehr mit dem Verfahren der *maximum likelihood* bestimmen (s. Schröder & Reineking 2004a, in diesem Band); denn die *Likelihood* wird umso größer, je steiler die Kurve der vorhergesagten Vorkommenswahrscheinlichkeiten von dem Bereich der Vorkommen zu dem Bereich der Nichtvorkommen abfällt, so dass die Absolutbeträge der Parameterschätzungen gegen unendlich streben. Eine Möglichkeit, mit dieser Situation umzugehen, sind die oben bereits erwähnten *shrinkage*-Verfahren, die auch unter vollständiger Trennung von Vorkommen und Nichtvorkommen noch endliche Parameterwerte schätzen und darüber hinaus die Kalibrierung der Modelle häufig verbessern (siehe auch den Beitrag von Reineking & Schröder 2004 über Variablenselektion in diesem Band).

Accuracy: Brier Index und mittlere Devianz

Eine weitere Gruppe von Maßen bestimmt die durchschnittliche Abweichung zwischen Vorhersage und Beobachtung. Diese werden von Murphy & Winkler (1992) dem Bereich *accuracy* zugeordnet. Üblicherweise wird entweder ein quadratisches oder ein logarithmisches Abstandsmaß verwendet.

Der Brier Index wird im Bereich der Meteorologie häufig verwendet. Er entspricht dem mittleren quadratischen Fehler zwischen vorhergesagter Wahrscheinlichkeit und tatsächlichem Vorkommen bzw. Nichtvorkommen:

$$Brier = \frac{1}{N}\sum(y_i - p_i)^2 \qquad (3.4)$$

Der Wert des Brier Index liegt in der Regel zwischen 0 (perfektes Modell) und 0,25 (wertloses Modell). Der höchstmögliche Wert ist 1, aber Werte von mehr als 0,25 werden nur erreicht, wenn das Modell eine schlechtere Anpassung aufweist als ein Modell, das konstant eine Vorkommenswahrscheinlichkeit von 0,5 voraussagt.

Es gibt verschiedene Maße, die auf der logarithmischen Skala die mittlere Abweichung bestimmen. Eine Methode ist, die durchschnittliche log(*Likelihood*) pro Beobachtung zu berechnen (Hilden et al. 1978), d.h.

$$D = \frac{1}{N}LL = \frac{1}{N}\sum y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

Alternativ kann die mittlere Devianz berechnet werden, die gleich $-2D$ ist. Steyerberg et al. (2001) bestimmen D als $-2/n$ log(Likelihood). Die logarithmische Gewichtung entspricht derjenigen, die für die Anpassung der Modelle verwendet wird. Bei der logarithmischen Gewichtung der Fehler kann eine einzelne falsche Vorhersage dazu führen, dass der Index einen sehr ungünstigen Wert annimmt. Das ist dann der Fall, wenn dem eingetretenen Ereignis eine sehr kleine Wahrscheinlichkeit zugeschrieben wurde, also für eine Fläche mit Vorkommen eine sehr geringe Vorkommenswahrscheinlichkeit geschätzt wurde. Dieser Index ist daher wesentlich empfindlicher gegenüber falsch kalibrierten, überangepassten Modellen als der Brier Index.

3.3.2 Maße, die eine Diskretisierung voraussetzen

Eine wichtige Gruppe von Gütemaßen arbeitet nicht mit den vorhergesagten Wahrscheinlichkeiten, sondern mit Klassifizierungen, d.h., dass für jede Beobachtung vorhergesagt wird, ob der Organismus vorkommt oder nicht. Einige Verfahren wie beispielsweise regelbasierte

Entscheidungsbäume liefern ausschließlich diese Information, so dass die im oberen Abschnitt besprochenen Gütemaße nicht angewendet werden können. Die vorhergesagten Vorkommen und Nichtvorkommen werden mit den Beobachtungen verglichen und daraus wird eine Klassifikationsmatrix (engl. *classification matrix* oder *confusion matrix*) erstellt, die einer Vierfeldertafel entspricht (Abb. 3.5): die Anzahl richtig vorhergesagter Vorkommen (a), die Anzahl richtig vorhergesagter Nichtvorkommen (d), die fälschlich als Vorkommen prognostizierten Nichtvorkommen (b) sowie die fälschlich als Nichtvorkommen prognostizierten Vorkommen (c).

Klassifikationsmatrix

		Daten 1	Daten 0	
Prognose	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
		a+c	*b+d*	*n*

% korrekt: $(a+d)/n$
Sensitivität: $a/(a+c)$
Spezifizität: $d/(b+d)$

Abb. 3.5. Klassifkationsmatrix und einige abgeleitete Gütemaße. „% korrekt" bedeutet den Anteil korrekter Prognosen.

Ausgehend von der Klassifikationsmatrix können eine Reihe von Gütemaßen berechnet werden. Hierzu gehören z.B. die weit verbreiteten Maße Gesamtvorhersagefehler, Sensitivität und Spezifizität (s.o.) sowie Cohen's Kappa (s.u.). Eine erweiterte Übersicht stellen Fielding & Bell (1997) zur Verfügung.

Da in der Klassifikationsmatrix keine Information mehr über die vorhergesagten Wahrscheinlichkeiten enthalten ist, quantifizieren diese Maße in erster Linie verschiedene Aspekte der Diskriminanz und eignen sich kaum zur Quantifizierung der Kalibrierungseigenschaften des Modells.

Diskriminanz: Fehlerrate, Sensitivität, Spezifizität

Die Fehlerrate bzw. alternativ der Anteil korrekter Klassifikation (*matching index*) ist wohl das am häufigsten verwendete Gütemaß. Ein Problem der Fehlerrate ist, dass sie sehr stark von der Prävalenz abhängig ist (s. Abb. 3.7). So überrascht eine geringe Fehlerrate nicht, wenn eine der beiden Klassen sehr häufig ist und auch häufig vorhergesagt wird. Beträgt beispielsweise der Anteil der Vorkommen 10% und in 15% der Fälle wird ein Vorkommen vorhergesagt, dann ist der erwartete

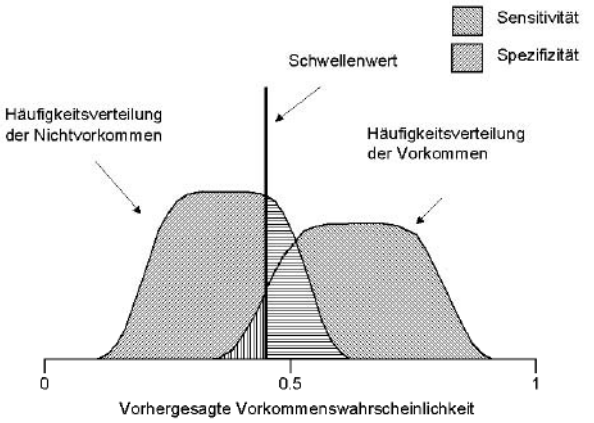


Abb. 3.6. Effekt der Wahl des Klassifikationsschwellenwertes auf Statistiken der Klassifikationsmatrix (verändert nach Pearce & Ferrier 2000). Je niedriger der Schwellenwert ist, desto größer ist der Anteil tatsächlicher Vorkommen, der richtig klassifiziert wird (Sensitivität). Gleichzeitig nimmt aber die Spezifizität ab, d.h. dass der Anteil falsch klassifizierter Nichtvorkommen steigt (siehe auch Abb. 3.7).

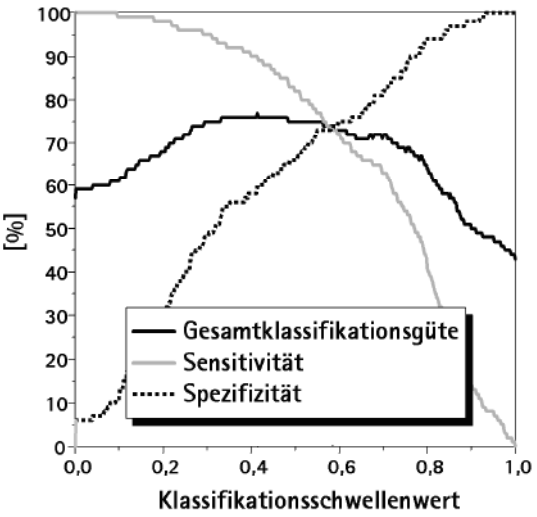


Abb. 3.7. Abhängigkeit des Anteils korrekter Klassifikationen, der Sensitivität und Spezifizität vom Klassifikationsschwellenwert (verändert nach Schröder & Richter 2000).

Anteil richtiger Vorhersagen bereits 78% (0,1 · 0,15 + 0,9 · 0,85), wenn die Vorhersagen rein zufällig und die Ereignisse unabhängig sind. Der Anteil korrekter Vorhersagen muss daher immer unter Berücksichtigung der Prävalenz beurteilt werden.

Ein weiteres Problem ist, dass Fehlerrate und Anteil korrekter Klassifikationen nichts darüber aussagen, welche Fehler gemacht wurden – ob vor allem tatsächliche Vorkommen nicht erkannt wurden oder Vorkommen prognostiziert wurden, wo keine beobachtet wurden.

Sensitivität und Spezifizität quantifizieren die Vorhersagerichtigkeit bezüglich der Vorkommen oder der

Nichtvorkommen (vgl. Abb. 3.5). Diese beiden Maße ergänzen somit die Information der Fehlerrate. Wenn die drei Maße angegeben sind und die Gesamtzahl der Beobachtungen, dann lässt sich die Klassifikationsmatrix rekonstruieren, und andere Gütemaße, die auf der Klassifikationsmatrix beruhen, berechnen.

Diskriminanz: Cohen's Kappa

Wie oben gezeigt, ist der Anteil korrekter Vorhersagen, die bereits zufällig zu erwarten sind, von der Prävalenz und der Häufigkeit der vorhergesagten Vorkommen („Vorhersageprävalenz") abhängig. Die Idee, auf der Cohen's Kappa beruht, ist nun, diesen Anteil erwarteter korrekter Vorhersagen von dem beobachteten Anteil korrekter Vorhersagen abzuziehen, und so zu normieren, dass das Maximum 1 ist. Aufgrund dieses Ansatzes wird Cohen's Kappa oftmals als zufallskorrigiertes Maß für den Anteil korrekter Vorhersagen bezeichnet (Cohen 1960; Pontius 2000).

Kappa wird folgendermaßen berechnet:

$$\kappa = \frac{p_{observed} - p_{expected}}{1 - p_{expected}} = \frac{\%korrekt - [p_D p_V + (1-p_D)(1-p_V)]}{1 - [p_D p_V + (1-p_D)(1-p_V)]} \quad (3.5)$$

wobei %korrekt den Anteil korrekter Vorhersagen, $(a+d)/n$ in Abb. 3.5, bezeichnet; p_D ist der Anteil der Vorkommen an allen Beobachtungen, $(a+c)/n$ in Abb. 3.5 und damit eine Schätzfunktion für die Prävalenz; p_V die „Vorhersageprävalenz", $(a+b)/n$ in Abb. 3.5, bezeichnet den als Vorkommen klassifizierten Anteil der Vorhersagen.

Die Nützlichkeit von Cohen's Kappa wird sehr unterschiedlich bewertet. Beispielsweise wird der Wert von Kappa von der Prävalenz beeinflusst, so dass Kappa-Werte selten über verschiedene Studien hinweg vergleichbar sind (Thompson & Walter 1988; Feinstein & Cicchetti 1990). Zudem herrscht Uneinigkeit darüber, wie verschiedene Werte von Kappa zu interpretieren sind, und es gibt verschiedene Vorschläge für Bewertungen von Kappa:

Nach Landis & Koch (1977):

Kappa index	*Degree of agreement*
0,0 - 0,4	*slight to fair model performance*
0,4 - 0,6	*moderate*
0,6 - 0,8	*substantial and*
0,8 - 1,0	*almost perfect*

Nach Monserud & Leemans (1992):

Kappa index	*Degree of agreement*
< 0,05	*No*
0,05 - 0,20	*Very poor*
0,20 - 0,40	*Poor*
0,40 - 0,55	*Fair*
0,55 - 0,70	*Good*
0,70 - 0,85	*Very good*
0,85 - 0,99	*Excellent*
0,99 - 1,00	*Perfect*

Zur Problematik des Schwellenwertes

Gütemaße, die auf der Klassifikationsmatrix beruhen, werden häufig aus zwei Gründen kritisiert. Zum einen ginge durch die Klassifikation Information verloren, weil nur noch betrachtet wird, ob eine Vorhersage größer oder kleiner als ein bestimmter Schwellenwert ist. Zum anderen hängen die Einträge in der Klassifikationsmatrix, und damit die Werte der abgeleiteten Gütemaße, sehr stark von dem gewählten Schwellenwert ab (s. Abb. 3.7).

Zum ersten Argument lässt sich entgegnen, dass alle Kennzahlen Information „vernichten", um einen bestimmten Aspekt der Modellvorhersagen deutlicher werden zu lassen. Die ursprünglichen Vorhersagen lassen sich aus keinem Gütemaß rekonstruieren; es ist daher nicht wesentlich, dass ein Gütemaß Information vernachlässigt, sondern es kommt darauf an, ob es relevante und interpretierbare Aspekte des Modellverhaltens quantifiziert.

Dem zweiten Argument lässt sich sagen, dass 0,5 ein ausgezeichneter Schwellenwert ist, der beispielsweise bei einem korrekt kalibrierten Modell den Gesamtvorhersagefehler minimiert. Um Vergleichbarkeit zu gewährleisten, sollte dieser Schwellenwert gewählt werden. Wenn Gründe vorliegen, warum die beiden Fehlertypen in der Vorhersage unterschiedlich gravierend sein sollten, dann kann es sinnvoll sein, die Gütemaße zusätzlich für abweichende Schwellenwerte zu berechnen, die die Bedeutung der Fehler berücksichtigen. Oftmals wird jedoch durch die datengeleitete Wahl eines von 0,5 abweichenden Schwellenwertes, etwa durch Optimierung hinsichtlich eines Gütemaßes (vgl. Franklin 1998; Guisan et al. 1998, 1999; Schröder & Richter 2000), in erster Linie eine Rekalibrierung des Modells vorgenommen. Es scheint uns aber sinnvoller zu sein, in einem ersten Schritt das Modell so gut wie möglich zu kalibrieren (s.o.) und den Schwellenwert dann ausschließlich in Abhängigkeit von der Bewertung der verschiedenen Fehler zu wählen. Dieser kann z.B. mit Hilfe des Programms ROC_AUC (Schröder 2003) berechnet werden (`http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html`), das auch eine Berechnung des *AUC*-Wertes samt Konfidenzintervall ermöglicht (s.u.).

Mit dem Nachteil, keine Information über die Wahrscheinlichkeiten zu verwenden, ist andererseits auch der Vorteil verbunden, dass diese Maße auch für Klassifikationsverfahren anwendbar sind, die keine Wahrscheinlichkeiten ausgeben, sondern direkt das Klassifikationsergebnis, wie z.B. Klassifikationsbäume. Sie haben damit ein breiteres Anwendungsfeld.

3.4 Konfidenzintervalle für Gütemaße

Eine sinnvolle Modellbewertung umfasst nicht allein die Berechnung geeigneter Gütemaße, sondern auch eine Abschätzung ihrer Unsicherheit. Eine sehr hoher Wert der Modellgüte sagt nicht besonders viel aus, wenn er sehr unsicher ist. Eine Abschätzung dieser Unsicherheit erfolgt über die Berechnung von Vertrauens- bzw. Konfidenzintervallen, die mit einer vorgebbaren Wahrscheinlichkeit den wahren Wert des Gütemaßes überdecken. Je weiter ein Vertrauenswall geschätzt wird, desto unsicherer ist die Schätzung des Gütemaßes. Beispielsweise ermöglicht das Programm ROC_AUC (Schröder 2003) die Berechnung von Konfidenzintervallen für *AUC* – ermittelt mit dem *Bootstrap*-Perzentil-Verfahren – als auch für Sensitivität und Spezifizität bei ausgewählten Klassifikationsschwellenwerten (s. auch Beitrag von Schröder & Reineking 2004b, in diesem Band).

3.5 Unverzerrte Schätzung der Werte und von Vertrauensintervallen

Bislang haben wir die verschiedenen Gütemaße vorgestellt, aber potentielle Probleme bei ihrer Berechnung nicht diskutiert.

Es ist nicht trivial, eine unverzerrte Schätzung der Gütemaße zu erhalten. Wenn man dieselben Daten zur Berechnung des Gütemaße verwendet, mit denen das statistische Modell angepasst wurde, resultieren zu optimistische Werte und Konfidenzintervalle (Schröder & Reineking 2004b; Verbyla & Litvaitis 1989). Das ist nicht besonders erstaunlich, wurden doch die Modellparameter gerade so gewählt, dass sie die beobachteten Daten besonders gut beschreiben. Beispielsweise ist die Kalibrierung einer logistischen Regression bezüglich der Daten, auf denen die Parameter geschätzt wurden, immer perfekt. Für einige Verfahren und Gütemaße, z.B. die Fehlerrate bei logistischer Regression, lässt sich der Optimismus abschätzen (Efron 1986). Im allgemeinen ist für eine realistische Schätzung der Gütemaße eine interne oder externe Validierung notwendig. Einige wichtige Verfahren der internen Validierung werden im Beitrag von Schröder & Reineking (2004b) in diesem Band vorgestellt.

Ein wichtiger Aspekt bei der Berechnung von Gütemaßen ist die Abschätzung der Unsicherheit, die mit der Berechnung des Gütemaßes verbunden ist. Angenommen, ein Modell hat eine geschätzte Fehlerrate von 10%, und ein zweites Modell hat eine Fehlerrate von 15%. Es bedeutet einen Unterschied, ob die Unsicherheit bei diesen Fehlerraten 1% beträgt oder 10%. Eine Abschätzung der Unsicherheit bei Gütemaßen, die mittels interner Validierung berechnet wurden, haben wir noch in keiner Anwendung gesehen.

3.6 Diskussion und Zusammenfassung

Die Beurteilung von Modellen, die Wahrscheinlichkeitsvorhersagen machen, beruht auf der gemeinsamen Verteilung $p(f,y)$ von Vorhersagen und beobachteten Vorkommen und Nichtvorkommen. Es gibt mehrere Aspekte dieser gemeinsamen Verteilung, die quantifiziert werden können. In jedem Gütemaß geht Information verloren; kein einzelnes Gütemaß kann alle Aspekte der Modellgüte erfassen. Die graphische Darstellung in Form von Attributdiagrammen und *Likelihood*-Diagrammen ist sinnvoll, weil komplexe Muster gut optisch erfasst werden können.

Die umfangreiche Diskussion über Gütemaße für Modelle, die Wahrscheinlichkeitsvorhersagen machen, konnten wir hier nur in Ansätzen umreißen. Zudem ist die Diskussion keineswegs abgeschlossen, und es ist zu erwarten, dass weitere Gütemaße entwickelt werden, die spezifische Vorzüge aufweisen. Der Ansatz von Murphy & Winkler (1992) scheint uns jedoch als konzeptioneller Orientierungspunkt in dieser Diskussion sehr nützlich zu sein. Insbesondere ist die Verwendung von Attributdiagrammen sehr vielversprechend, da sie eine schnelle visuelle Orientierung erlauben, und die wesentliche Information enthalten, die auf unterschiedliche Weise in den Gütemaßen codiert ist.

Es ist wünschenswert, dass sich ein Satz von Gütemaßen etabliert, der von vielen Autoren angewendet wird. Harrell schlägt vor, *AUC* (d.h. den Konkordanzindex c oder äquivalent das Zusammenhangsmaß Somer's D_{xy}) und Kalibrierungsmaße zu verwenden: „*A good set of indexes to estimate for summarizing a model validation is the c or D_{xy}, indexes and measures of calibration.*“ (Harrell 2001, S. 249).

Als Kalibrierungsmaße schlagen wir den Achsenabschnitt und die Steigung der Kalibrierungsgeraden vor. Zudem halten wir den R^2_N nach Nagelkerke für einen geeigneten Index, da er zusätzlich zur Kalibrierung noch das *Refinement* quantifiziert. Darüber hinaus sollten auch Maße verwendet werden, die auf Grundlage der Klassifikationsmatrix berechnet werden, um eine Vergleichbarkeit mit Verfahren zu gewährleisten, die lediglich eine Klassifikation liefern. Hierfür bietet sich trotz einiger Nachteile das Trio aus Sensitivität,

Spezifizität und Gesamtfehlerrate bzw. Anteil korrekter Klassifikationen an. Als Schwellenwert sollte dabei 0,5 verwendet werden, gegebenenfalls zusätzliche Schwellenwerte, falls die beiden Fehlertypen unterschiedliches Gewicht haben.

Abschließend möchten wir jedoch noch einmal betonen, dass die verschiedenen Gütemaße lediglich Hilfestellungen liefern hinsichtlich der Beurteilung der entscheidenden Frage, ob ein Modell für den beabsichtigten Verwendungszweck geeignet ist. Die Beurteilung der Güte eines Modells für einen gegebenen praktischen Zweck wird von den statistischen Gütemaßen nicht geleistet. Sowohl die relative Bedeutung verschiedener Gütemaße als auch die Relevanz der konkreten Werte hängen vom Kontext der Anwendung ab.

3.7 Danksagung

Vielen Dank an Hans-Peter Bäumer, Universität Oldenburg, und Tamara Münkemüller, UFZ Leipzig-Halle, für die hilfreichen Kommentare zum Manuskript.

Literaturverzeichnis

Brier, G. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Cleveland, W. S. 1981. Lowess: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35(1):54.

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

Dawid, A. P. 1986. Probability forecasting. In Kotz, S., Johnson, N. & Read, C., editors, *Encyclopedia of Statistical Sciences*, pages 210–218. Wiley, New York.

Efron, B. 1986. How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association*, 81(394):461–470.

Feinstein, A. R. & Cicchetti, D. V. 1990. High agreement but low kappa (1) - the problems of 2 paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

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.

Franklin, J. 1998. Predicting the distribution of shrub species in southern california from climate and terrain-derived variables. *Journal of Vegetation Science*, 9(5):733–748.

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.

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

Harrell, F. E., Lee, K. L. & Mark, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

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.

Hilden, J., Habbema, J. D. F. & Bjerregaard, B. 1978. Measurement of performance in probabilistic diagnosis: 3. methods based on continuous-functions of diagnistic probabilities. *Methods Inf. Med.*, 17:238–246.

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

Landis, J. R. & Koch, G. G. 1977. Measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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.

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

Murphy, A. H. & Winkler, R. L. 1992. Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7(4):435–455.

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

Pearce, J. & Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133:225–245.

Pontius, R. G. 2000. Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing*, 66(8):1011–1016.

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

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. 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.

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

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.

Steyerberg, E. W., Eijkemans, M. J. C., Harrell, F. E. & Habbema, J. D. F. 2000. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, 19(8):1059–1079.

Thompson, W. D. & Walter, S. D. 1988. A reappraisal of the kappa-coefficient. *Journal of Clinical Epidemiology*, 41(10):949–958.

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

3.8 Datenblatt

3.8.1 Software

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

3.8.2 Webresources

- R unter `www.r-project.org`
- Ein Delphi-Programm zur Berechnung von ROC-Kurven, *AUC*-Werten samt Konfidenzintervallen und optimalen Klassifikationsschwellenwerten unter Berücksichtigung unterschiedlicher Fehlergewichte wird von B. Schröder unter `http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html` zur Verfügung gestellt.

3.8.3 Kommentierte Literatur

Sehr gute Überblicke über die Gütemaße liefern die folgenden Arbeiten aus dem Bereich der Habitatmodellierung:

Fielding & Bell (1997) sowie Pearce & Ferrier (2000) und das allgemeine Buch zu Modellierungsstrategien von Harrell (2001).

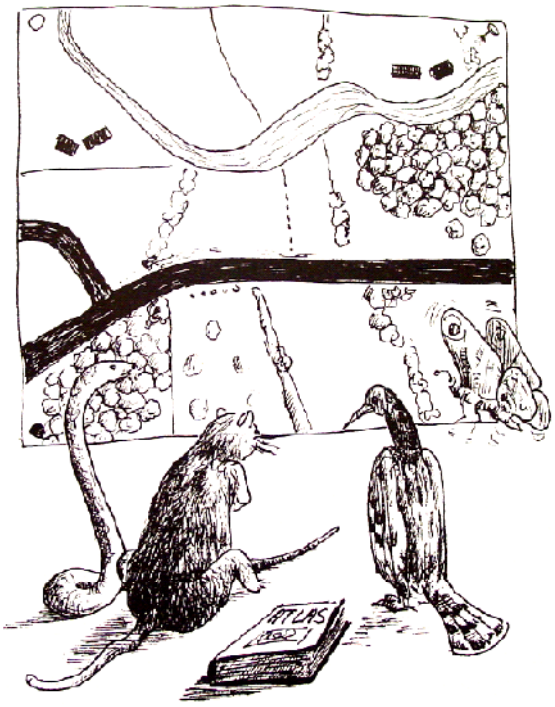
HABITATMODELLE

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