Vogelwarte 43, 2005: 19 – 38

Einführung in die multivariate Statistik für Feldornithologen: Hauptkomponentenanalyse, Diskriminanzanalyse und Clusteranalyse

Ortwin Elle

Elle O 2005: Introduction to multivariate statistics for field ornithologists: Principal Component Analysis, Discriminant Analysis and Cluster Analysis. Vogelwarte 43: 19 – 38.

This practically oriented introduction deals with potentialities and limitations of the use of multivariate statistics in field ornithology. Principal Component Analysis, Discriminant Analysis and Cluster Analysis belong to the most important multivariate statistical techniques in ecological research. This paper introduces the theoretical basics and also offers guidance on the application of these techniques. Furthermore, indicators of the quality of analysis and possibilities of interpretation are discussed for each technique and demonstrated on the basis of an empirical example.

OE: Universität Trier, Biogeographie, D-54286 Trier; E-Mail: elle@uni-trier.de

1. Einleitung

1.1. Zielsetzung

Die weite Verbreitung leistungsfähiger Statistik-Computerprogramme hat dazu geführt, dass die Anwendung statistischer Verfahren auch in der Feldornithologie zur Selbstverständlichkeit geworden ist. Die einfache Bedienbarkeit dieser Programme ist Segen und Fluch zugleich: Auf der einen Seite stehen damit auch dem Nicht-Fachmann komplexere statistische Verfahren aus dem Bereich der multivariaten Statistik zur Verfügung. Auf der anderen Seite besteht die Gefahr einer unkritischen Anwendung solcher Verfahren, denn die (technische) Durchführung multivariater Analysen am Computer ist selbst ohne größere Kenntnisse der dazugehörigen theoretischen Grundlagen in kürzester Zeit möglich. Häufig werden vorgegebene Standardeinstellungen des Statistikprogramms vom Bearbeiter unreflektiert übernommen, oder die Wirkungsweise alternativer Teilschritte in den meist komplexen Verfahren ist nicht bekannt. Indikatoren für die Qualität einer Analyse können oft nicht richtig interpretiert werden oder werden vollkommen ignoriert. Ein Verfahren bekommt somit eher den Charakter einer „Blackbox", bei der Input und Output zwar bekannt sind, der Weg zum Ziel aber weitgehend im Dunkeln bleibt.

Seit einigen Jahren existieren zahlreiche, auch für den Laien verständliche Einführungen in die Grundlagen der Statistik, die z.T. sogar speziell auf Ornithologen zugeschnitten sind (z. B. Niemeyer 1980; James & McCulloch 1985; Fliege 1986; Lorenz 1992; Bärlocher 1999; Kesel et al. 1999; Ashcroft & Pereira 2003; Fowler & Cohen o. D.). Die multivariate Statistik hat dagegen den Schritt in das Bewusstsein einer „breiteren Öffentlichkeit" noch nicht vollzogen, so dass die Einsatzmöglichkeiten dieser hilfreichen Verfahren nur einer kleinen Minderheit bekannt sein dürften. Dieser Artikel wendet sich in erster Linie an den interessierten Laien mit statistischen Grundkenntnissen. Er soll in die Lage versetzt werden, die Anwendung multivariater statistischer Verfahren in Publikationen kritisch beurteilen zu können. Dieser Artikel kann kein Ersatz für die einschlägigen Lehrbücher zur multivariaten Statistik sein (z. B. Gauch 1982; Deichsel & Trampisch 1985; Digby & Kempton 1987; Bahrenberg et al. 1992; Backhaus et al. 1994; Legendre & Legendre 1998; McGarigal et al. 2000), wird aber mit Sicherheit den Einstieg in diese komplexe Materie erleichtern. Hauptkomponentenanalyse, Diskriminanzanalyse und Clusteranalyse gehören zu den wichtigsten multivariaten Verfahren in der biologischen Forschung. Jedes Verfahren bietet ganz unterschiedliche Ansatzpunkte zur Bearbeitung einer Fragestellung. Häufig gibt es nicht „die" einzig richtige Entscheidung für ein ganz bestimmtes Verfahren, sondern mehrere alternative Ansätze. Manchmal ist auch eine Kombination mehrerer multivariater Verfahren angebracht. Entscheidend dabei ist, dass der Bearbeiter die Arbeitsweise und die „Stellschrauben" der verschiedenen Verfahren kennt und seine Methodenauswahl genau begründen kann. Im folgenden werden – soweit wie möglich von „statistischem Ballast" befreit – Möglichkeiten und Grenzen der multivariaten Statistik für die Feldornithologie vorgestellt.

1.2. Wozu multivariate Statistik?

Vereinfacht gesagt sprechen wir von multivariaten Analyseverfahren, sobald mehr als zwei Variablen gleichzeitig untersucht werden. Angenommen, Sie gehen der „klassischen" Frage nach, wie die Koexistenz von drei nahe verwandten Vogelarten in einem reich strukturierten Lebensraum möglich ist. Beispielsweise untersuchen

Sie das Habitatwahlverhalten von Mönchsgrasmücken, Dorngrasmücken und Gartengrasmücken in einem Lebensraum, der durch ein kleinräumiges Mosaik aus unterschiedlichen Gebüschen, Hecken, Streuobstwiesen und Baumgruppen gekennzeichnet ist.

Habitatwahl ist ein wichtiger Aspekt der ökologischen Nische. Die ökologische Nische beschreibt die funktionelle Rolle einer Art in einer Gemeinschaft (z. B. ihre Stellung in der Nahrungskette) sowie ihr Verhalten zu den Umweltgradienten wie Temperatur, Feuchtigkeit etc. (vgl. z. B. Odum 1983). Hutchinson (1965) betrachtet die ökologische Nische als multidimensionalen Raum oder Hypervolumen, worin die Umwelt einem Individuum oder einer Art erlaubt, auf unbestimmte Zeit zu überleben und macht die ökologische Nische damit mathematischen Analysen zugänglich. Jede Dimension in diesem Hypervolumen entspricht einem relevanten Umweltfaktor. Nach dem Konkurrenz-Ausschluss-Prinzip (Hardin 1960) können zwei Arten nicht exakt die gleiche ökologische Nische besetzen, sofern bestimmte Faktoren limitierend wirken.

Die getrennte Analyse von jeweils nur einem Merkmal der ökologischen Nische wird mit hoher Wahrscheinlichkeit nicht zu einer Artunterscheidung führen, denn häufig ist ein mehr oder weniger breiter Überschneidungsbereich zwischen zwei Vogelarten entlang eines Umweltgradienten zu beobachten. Die Stärke der hier vorgestellten Verfahren liegt in der gleichzeitigen Untersuchung mehrerer Variablen, weil dieses die Multidimensionalität ökologischer Systeme besser widerspiegelt. Außerdem können redundante Informationen (d. h. überflüssige Variablen oder Daten, die keine zusätzliche Information liefern) im Datensatz aufgespürt und eliminiert werden. Dieses ist besonders hilfreich, weil der Bearbeiter in der Planungsphase einer Untersuchung oft noch gar nicht wissen kann, welche Variablen wirklich wichtig für die Beantwortung der Fragestellung sind. Um auf der „sicheren Seite" zu sein, wird deshalb häufig bewusst eine zu große Zahl von Variablen erhoben und deren Zusammenspiel im Nachhinein analysiert.

1.3. Überblick

Vögel nehmen ihre Umwelt nicht zwangsläufig so wahr, wie wir sie mit unserer spezifisch menschlichen Sichtweise für uns messbar machen. Die Vegetationsstruktur eines Lebensraumes, beispielsweise, kann in den seltensten Fällen durch eine einzige Messgröße dargestellt werden (vgl. Bell et al. 1991). Will man mit dieser Habitateigenschaft quantitativ arbeiten, um beispielsweise Unterschiede in der Habitatwahl der drei Grasmückenarten herauszustellen, so kann die Vegetationsstruktur durch einen ganzen Satz von unterschiedlichen Variablen beschrieben werden, die vom Bearbeiter als bedeutsam eingeschätzt werden (z. B. Stammumfang, Schichtung, Deckungsgrade, Art der Belaubung, Verzweigungstypen etc.). Mit Hilfe einer Hauptkomponentenanalyse wird aus solchen unterschiedlich gewichteten „Stellvertreter-Variablen" die komplexe Variable „Vegetationsstruktur" konstruiert und kann danach für weitergehende Analysen verwendet werden. Solche abgeleiteten Variablen werden als Hauptkomponenten bezeichnet. Wenig „greifbare" Größen, die nicht direkt messbar sind, werden auf diese Weise operationalisiert. Aber auch der umgekehrte Fall ist denkbar, wenn die komplexe Größe a priori gar nicht bekannt ist, sondern sich erst im Rahmen der Hauptkomponentenanalyse ergibt. Es gibt häufig unerwartete, wechselseitige Abhängigkeiten (positive oder negative Korrelationen) zwischen verschiedenen Umweltvariablen. In der Hauptkomponentenanalyse geht man davon aus, dass diese interagierenden Variablen nicht einzeln für sich, sondern als latenter, gleichsam hinter den Variablen stehender Umweltfaktor (in Form einer Hauptkomponente) wirken. Beispielsweise könnte sich aus einer Analyse ergeben, dass eine bestimmte Kombination von Variablen zur Kennzeichnung unterschiedlicher Sukzessionsstadien der Vegetation geeignet ist. Die mit dieser Informationsverdichtung stets verbundene Reduzierung der ursprünglichen Variablenzahl auf wenige Hauptkomponenten ist in der Praxis oft der entscheidende Beweggrund für den Bearbeiter, eine Hauptkomponentenanalyse durchzuführen. Handhabbarkeit und Interpretierbarkeit der „Urdaten" können dadurch im allgemeinen wesentlich verbessert werden.

Es liegt in der Natur des Menschen, Erscheinungen in der Umwelt zu klassifizieren, indem Objekte mit ähnlichen Eigenschaften zusammengefasst werden. Ist die Gruppierung aus sachlogischen oder theoretischen Überlegungen heraus bekannt (z. B. Artzugehörigkeit, Geschlecht, Altersklasse usw.) und soll mit multivariaten Verfahren bestätigt werden, so spricht man von „Strukturen-prüfenden Verfahren". Dazu gehört die Diskriminanzanalyse. Mit Hilfe dieses Verfahrens kann beispielsweise die Hypothese untersucht werden, dass die drei Grasmückenarten unterschiedliche Habitatnischen einnehmen. Es könnte untersucht werden, ob das Vorkommen der drei Arten jeweils an ein bestimmtes Sukzessionsstadium der Vegetation gebunden ist. Die Territorien der drei Arten sollten sich dann aufgrund geeigneter Vegetationsstrukturvariablen unterscheiden lassen. Dafür wird die Variabilität der Vegetationsstrukturen innerhalb der durch die Artzugehörigkeit vorgegebenen Gruppen mit der Variabilität zwischen den Gruppen in Beziehung gesetzt und sollte intraspezifisch deutlich geringer sein als interspezifisch. Andernfalls müsste die Hypothese verworfen werden.

Ist dagegen die Zusammensetzung der einzelnen Gruppen a priori nicht bekannt und soll erst durch die Analyse herausgefunden werden, so spricht man von „Strukturen-entdeckenden Verfahren". Die Clusteranalyse verfolgt diesen Zweck. Möglicherweise nutzen die drei Grasmückenarten im untersuchten Lebensraum teilweise ein ähnliches Spektrum von Mikrohabita-

ten (z. B. Sukzessionsstadien) und trennen sich bei syntopem Vorkommen über andere Dimensionen der ökologischen Nische. In einem solchen Fall würde eine Clusteranalyse der Territorien auf der Grundlage ihrer Vegetationsstruktur zu interspezifisch gemischten Clustern (Gruppierungen) führen. Die Cluster entsprechen in diesem Fall einem durch ein bestimmtes Strukturprofil gekennzeichneten Mikrohabitattyp. Eine Analyse der einzelnen Cluster könnte Aufschluss darüber bringen, in welchen Mikrohabitaten die einzelnen Grasmückenarten ihren Schwerpunkt haben oder ob sich zwei der drei Arten untereinander in ihrer Habitatwahl ähnlicher sind und deshalb häufiger gemeinsam gruppiert wurden.

Danksagung: Ich danke Herrn Prof. P. Berthold und Herrn Dr. H.-W. Ley von der Vogelwarte Radolfzell für ihr großes Engagement für ein so unpopuläres Thema wie die multivariate Statistik, ohne das dieser Artikel wahrscheinlich nicht publiziert worden wäre. Herrn Prof. F. Bairlein von der Vogelwarte Helgoland danke ich für die kritische Durchsicht des Manuskriptes und wertvolle Anregungen und Verbesserungsvorschläge.

2. Vorbemerkungen zu den verwendeten Daten

2.1. begleitendes Fallbeispiel

Zur Veranschaulichung der drei behandelten mulitvariaten Verfahren wird im folgenden ein Datensatz verwendet, der auf realen Freilanderhebungen beruht. Es geht dabei um das Habitatwahlverhalten von Mönchsgrasmücken *Sylvia atricapilla*, Dorngrasmücken *S. communis* und Gartengrasmücken *S. borin* in einem durch unterschiedliche Gehölzstrukturen gekennzeichneten Lebensraum (Elle 2003). Allerdings wurden aus didaktischen Gründen aus den insgesamt 397 untersuchten Territorien 37 Territorien ausgewählt. Die Ergebnisse werden durch diese (nicht zufällige) Selektion erheblich beeinflusst und entsprechen nicht mehr der Realität, die durch eine größere intraspezifische Variabilität und interspezifische Überlappung bei der Habitatwahl gekennzeichnet ist. Die autökologischen Aussagen, die aus diesem Rumpfdatensatz abgeleitet werden, sind deshalb zu simplifizierend. Sie dienten lediglich der Veranschaulichung der drei vorgestellten multivariaten Verfahren und sollten in dieser Form nicht mit den drei Grasmückenarten assoziiert werden. Zu diesem Zweck sollte ausschließlich die Originalarbeit (Elle 2003) herangezogen werden.

Die Habitatwahl einer Vogelart wird in diesem Untersuchungsansatz durch die Vegetationsstruktur in den Vogelterritorien bestimmt. Die Gehölzstrukturen in den Territorien werden auf der Grundlage von 12 Variablen gekennzeichnet. Acht dieser Strukturvariablen beschäftigen sich mit der Menge der Vegetation in acht gedachten, horizontalen Querschnitten durch den Vegetationsraum in verschiedenen Höhen – hier als Schichtvariablen S1 bis S8 bezeichnet (S1 = 0,3 m, S2 = 0,8 m, S3 = 1,5 m, S4 = 2,5 m, S5 = 3,5 m, S6 = 4,5 m, S7 = 7 m, S8 = 12 m Höhe). Je höher der Wert einer solchen Variable ist, desto mehr belaubte Vegetation ist in der entsprechenden Höhenschicht in einem Territorium vorhanden. Vier weitere Variablen geben Auskunft darüber, aus welchen Untereinheiten sich der belaubte Vegetationsraum in den Territorien zusammensetzt. Unterschieden wird zwischen punkthaften (PUNKT) und linearen Randstrukturen (LINIE), flächig ausgeprägten Vegetationseinheiten (FLÄCHE) und fehlender Vegetation (OHNE). Diese vier Variablen werden als Strukturtypen bezeichnet. Innerhalb eines gedachten, dreidimensionalen Gitters mit einer horizontalen Maschenweite von 20 m, welches das Untersuchungsgebiet lückenlos abdeckt, wird an jedem Kreuzungspunkt in den gleichen 8 Höhen wie für die Schichtvariablen jeweils einer dieser vier Strukturtypen angetroffen. Der komplette Vegetationsraum wird somit nach einem „Bausteinprinzip" abgebildet. Hohe Werte dieser Variablen stehen für häufiges Vorkommen des entsprechenden Strukturtyps in einem Territorium. Die Grenzen der Territorien wurden mit der Revierkartierungsmethode ermittelt (vgl. Bibby et al. 1995) und mit Hilfe eines Geographischen Informationssystems mit den Daten der Vegetationsstruktur verschnitten (methodische Details in Elle 2003). Die Rohdaten für die 37 untersuchten Territorien sind im Anhang zu finden.

2.2. Standardisierung der Daten

Merkmale mit großer Varianz bestimmen die Ergebnisse multivariater Verfahren stärker als Merkmale mit geringerer Varianz (s. unten). Die Varianz eines Merkmals korreliert aber nicht zwangsläufig mit der Wichtigkeit dieses Merkmals für das untersuchte System, sondern kann auch andere Ursachen haben. Das trifft insbesondere zu, wenn z. B. ganz unterschiedliche Eigenschaften der ökologischen Nische gemeinsam untersucht werden, die zudem in verschiedenen Einheiten gemessen werden oder ganz unterschiedliche absolute Wertebereiche abdecken. Beispielsweise hat die Eigenschaft eines Laubblattes (z. B. Länge oder Fläche) als möglicher Bestandteil der Vegetationsstruktur naturgemäß einen engeren biologisch möglichen Wertebereich als eine Baumeigenschaft (z. B. Stammhöhe oder Kronendurchmesser). Die Blatt-Eigenschaften würden von den Baum-Eigenschaften quantitativ „erschlagen" werden. Die Intensität des einfallenden Lichtes in verschiedenen Vegetationsschichten könnte eine weitere ergänzende Variable der Vegetationsstruktur sein (vgl. Fox 1979) und wird in einer vollkommen anderen Einheit gemessen. Ist ein Helligkeits-Unterschied von 1000 Lux an zwei Standorten entscheidender als eine Stammhöhendifferenz von 1 Meter? Sollten Längenangaben überhaupt in Metern oder besser in Millimetern erfolgen?

Wegen dieser fehlenden Vergleichbarkeit verschiedener Variablen in einem Gesamtsystem, die zu willkürlichen Verzerrungen der Ergebnisse führen würde, sollten

statt der Originalvariablen stets die z-standardisierten Variablen verwendet werden. Dabei wird von jedem Original-Variablenwert der arithmetische Mittelwert der Stichprobe subtrahiert und dieser Wert durch die Standardabweichung dividiert. Diese Transformation der Originaldaten bewirkt, dass jede Variable einen Mittelwert von 0 und eine Standardabweichung von 1 bekommt. Auf diese Weise erhält man gleichwertige Variablen, auf deren Grundlage die multivariaten Verfahren durchgeführt werden können. Derartig transformierte Variablen werden hier durch die Voranstellung eines „z" vor dem Variablennamen gekennzeichnet.

3. Vorstellung der multivariaten Verfahren

3.1. Hauptkomponentenanalyse

Ziel der Analyse

Jedes Untersuchungsobjekt (z. B. das Habitat einer Vogelart), kann durch eine große Anzahl verschiedener Merkmalsvariablen beschrieben werden. Das Ziel einer Hauptkomponentenanalyse (HKA) besteht darin, auf der Grundlage von Korrelationen zwischen den Variablen zusammenhängende Variablengruppen zu identifizieren und deren Informationsgehalt auf eine geringere Anzahl von komplexeren Variablen (sog. Hauptkomponenten, HK) zu übertragen. Die Dimensionalität des untersuchten Systems wird auf diese Weise herabgesetzt, wodurch es leichter handhabbar oder interpretierbar werden kann. Der damit verbundene Informationsverlust sollte möglichst gering gehalten werden. Für jedes Untersuchungsobjekt können die Werte für die neu definierten HK berechnet und somit die HK als „normale" Merkmalsvariablen weiterverwendet werden.

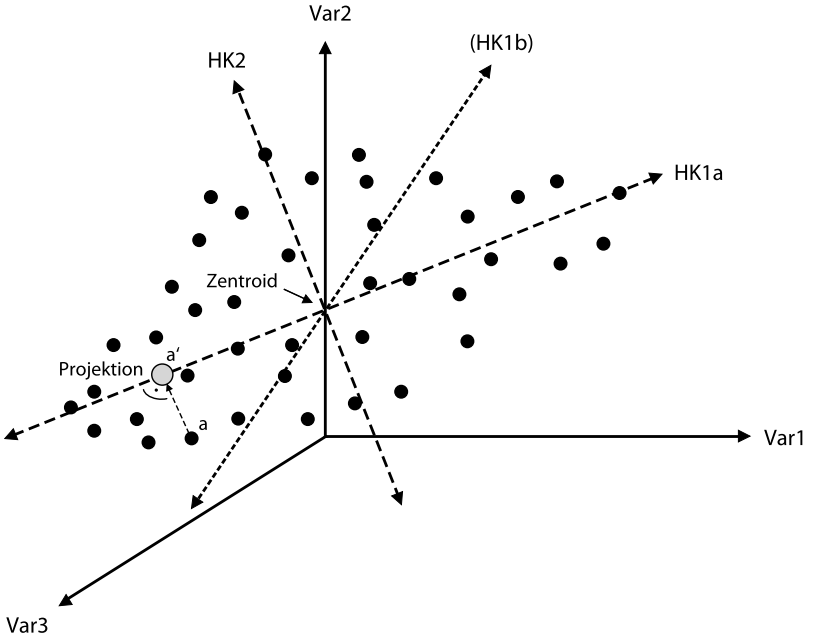


Abb. 1: Graphische Ableitung der ersten beiden Hauptkomponenten (HK) aus drei Originalachsen (Var). HK1a reproduziert das Maximum an Varianz, das durch Projektion der Punktwolke entlang einer Dimension möglich ist. Jede andere Lösung (z. B. HK1b) liefert deshalb ein schlechteres Ergebnis (nach McGarigal et al. 2000). HK2 liegt orthogonal zu HK1a und reproduziert das Maximum der noch nicht erklärten Varianz. – Graphic derivation of the first two Principal Component Axes (HK) from three original axes (Var). HK1a reproduces the maximum amount of variance that is possible by projection through the cloud of sample points in a single dimension. Any other solution (for example HK1b) yields worse results. HK2 is constrained by orthogonality to HK1a and reproduces the maximum of the variance yet unexplained.

Was sind Hauptkomponenten?

Zur graphischen Veranschaulichung der Definition von Hauptkomponenten diene ein hypothetischer Datensatz, in dem jedes Objekt durch drei Variablen gekennzeichnet wird und in Form einer dreidimensionalen Punktwolke dargestellt werden kann (Abb. 1). Jeder Punkt könnte beispielsweise einem Vogelterritorium entsprechen, das durch die Deckungsgrade der Baumschicht (Var. 1), Strauchschicht (Var. 2) und Krautschicht (Var. 3) gekennzeichnet wird. Wenn diese Punktwolke nicht in alle drei Dimensionen des Raumes gleichermaßen ausstrahlt (d. h. nicht sphärisch ist), sondern eine bestimmte Vorzugsrichtung aufweist, kann ein großer Teil des Informationsgehaltes der Punktwolke durch Projektion auf eine neu definierte Achse (d. h. Hauptkomponente) übertragen werden. Aus einem dreidimensionalen System ist somit ein eindimensionales System geworden. Die neue Achse würde dann eine Kombination aus Kraut-, Strauch- und Baumschicht beschreiben. Es ist leicht nachvollziehbar, dass diese drei Vegetationsschichten nicht unabhängig voneinander sind: Eine dichte Baumschicht unterdrückt möglicherweise eine Kraut- oder Strauchschicht, eine dichte Krautschicht kann die Ansiedlung von Gehölzen hemmen oder wird möglicherweise künstlich durch den Menschen gehölzfrei gehalten, usw.. Allerdings werden durch die Projektion auf die neue Achse nicht alle Objekte, die im dreidimensionalen System deutlich getrennt sind auch im eindimensionalen System getrennt dargestellt. Darin besteht der Informationsverlust. Dabei stellt für den gegebenen Datensatz die HK 1a aus Abb. 1 die optimale Lösung dar, denn sie verläuft durch den Mittelpunkt (Zentroid) der Punktwolke entlang ihrer größten Ausdehnung und reproduziert somit einen maximalen Anteil der Varianz des Originaldatensatzes. Dagegen ist die Lage der HK 1b (als eine von unendlich vielen weiteren denkbaren Lösungen) wesentlich ungünstiger, weil sie zu einem größeren Informationsverlust führt.

Diese an einem dreidimensionalen Modell demonstrierten Grundprinzipien zur Dimensionsreduktion lassen sich rechnerisch problemlos auf höherdimensionale Modelle übertragen. Man hat es dann mit einer N-dimensionalen Punktwolke des Datensatzes zu tun (wobei N der Anzahl der Variablen entspricht), die bei N>3 freilich graphisch nicht mehr darstellbar ist und gewisse Ansprüche an unser Abstraktionsvermögen stellt. Auch solche viel-dimensionalen Punktwolken weisen, wie im dreidimensionalen Merkmalsraum, meistens bestimmte Vorzugsrichtungen auf. Die erste Hauptkomponente wird stets so definiert, dass sie das Maximum an Varianz erklärt, das durch Projektion entlang einer Dimension (d. h. HK) möglich ist. Sie läuft damit durch den viel-dimensionalen Raum entlang der Hauptausdehnung der viel-dimensionalen Punktwolke. Die zweite HK wird so definiert, dass sie orthogonal (rechtwinklig) zur ersten HK liegt und gleichzeitig die durch die erste HK nicht erklärte Varianz maximal reproduziert. Sie verläuft somit in Richtung der zweitgrößten Ausdehnung der Punktwolke. Die dritte HK wird unter den gleichen Bedingungen definiert, d. h. Orthogonalität und maximale Erklärung der noch nicht erklärten Varianz. Dieser Vorgang ist so lange wiederholbar, bis genauso viele HK wie Originalvariablen definiert sind. Normalerweise werden jedoch deutlich weniger HK als Originalvariablen verwendet, denn eines der wichtigsten Ziele einer HKA ist ja die Reduzierung der Dimensionalität eines Systems. Die Notwendigkeit der z-Standardisierung wird aus dieser geometrischen Ableitung der HK unmittelbar deutlich: Die Positionierung der HK wird auf diese Weise unabhängiger von Zufällen aufgrund menschgemachter Einheiten und Wertebereiche der Originalvariablen.

Die Tatsache, dass alle HK orthogonal zueinander liegen, hat zur Folge, dass die einzelnen HK vollkommen unabhängig voneinander sind. Es gibt somit keine inhaltliche Redundanz (bzw. Korrelation) zwischen den HK, welche bei den Originalvariablen durchaus vorhanden ist und als Multikollinearität bezeichnet wird. Bei manchen multivariaten Verfahren (z. B. Diskriminanzanalyse) erschwert Multikollinearität die inhaltliche Interpretation der Ergebnisse. Die HKA wird deshalb häufig auch anderen multivariaten Verfahren vorgeschaltet, um Multikollinearität aus der Rohdatenmatrix zu eliminieren.

Inhaltliche Interpretation

Die inhaltliche Interpretation der durch die Hauptkomponenten neu definierten Achsen erfolgt über ihre relative Lage zu den Original-Variablenachsen. Je ähnlicher der Verlauf einer HK mit dem Verlauf einer Originalachse ist, desto stärker wird diese HK auch inhaltlich von der entsprechenden Originalvariablen bestimmt (vgl. Abb. 1). Der rechnerische Erklärungsgehalt jeder Originalvariablen für eine Hauptkomponente ist aus der sog. Komponenten-Ladungsmatrix ersichtlich. Die Komponenten-Ladung ergibt sich aus der Korrelation (Wertebereich zwischen -1 und +1) zwischen einer HK und einer Ausgangsvariablen als Maß für die Stärke und Richtung des Zusammenhangs. Konventionsgemäß ordnet man eine Originalvariable ab einem Ladungswert von 0,5 aufwärts (bzw. von –0,5 abwärts bei negativer Korrelation) einer HK zu (Backhaus et al. 1994: 228). Es gibt jedoch auch Abweichungen von diesem Richtwert (z. B. McGarigal et al. 2000).

Niemand würde eine Variable akzeptieren, von deren Relevanz für das untersuchte System er nicht überzeugt ist oder deren inhaltliche Bedeutung nicht bekannt ist. Die gleichen Ansprüche sollten auch an die HK gestellt werden – die ja nichts anderes als komplexere Variablen sind – denn die bloße Tatsache, dass der Computer eine HK berechnet, ist noch keine Garantie für ihre Qualität. Entscheidend ist deshalb die Interpretierbarkeit einer HK, denn es besteht die Gefahr von Scheinkorrelationen (ein „Klassiker“: Rückgang des „Klapperstorchs“ und der Geburtenrate in Industrieländern). Wenn also der Informationsgehalt der HK nicht nachvollziehbar ist, weil sie sich aus Variablengruppen zusammensetzen, deren Zusammenwirken inhaltlich keinen Sinn macht, sollten diese nicht für weitergehende Analysen verwendet werden.

Eine Optimierung bzgl. der Interpretierbarkeit von HK wird durch eine sog. Rotation erreicht. Die verbreitetste Methode in der ökologischen Forschung ist die Varimax-Rotation, bei der die Orthogonalität der Achsen aufrechterhalten wird. Ungünstig für die Interpretation ist, wenn eine Originalvariable mit mittleren Ladungswerten auf mehrere HK ähnlich hoch lädt. Wünschenswert wäre ein klares „ja“ oder „nein“ der Variablen zu den HK in Form von hohen (größer 0,6 bzw. kleiner -0,6) oder niedrigen (zwischen –0,3 und +0,3) Ladungswerten. Treten solche ungünstigen Fälle vermehrt in einer Ladungsmatrix auf, ist eine Rotation

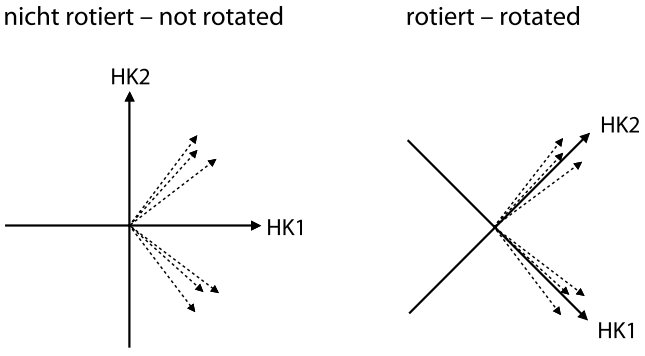


Abb. 2: Rotation der Hauptkomponentenachsen (HK). Die gestrichelten Pfeile repräsentieren die Komponentenladungen der Originalvariablen (nach Backhaus et al. 1994). Je näher eine Originalvariable bei einer HK liegt, desto höher der Ladungswert. – Rotation of Principal Component Axes (HK). Dotted arrows represent component loadings of the original variables. The closer an original variable lies to a HK the higher the loading.

angebracht. Aus Gründen, die für uns Nicht-Mathematiker schwer nachvollziehbar sind, wird die Aussagekraft einer Hauptkomponentenanalyse nicht verändert, wenn das durch die HK definierte Koordinatensystem in seinem Ursprung gedreht wird (Abb. 2). Diese Drehung bewirkt, dass die Originalvariablen sich über ihre Komponentenladungen deutlicher (d. h. mit höheren Ladungs-Werten) für eine bestimmte HK „entscheiden". Auf diese Weise sind die HK inhaltlich schärfer definiert und besser voneinander abgrenzbar.

Für jedes Untersuchungsobjekt (z. B. Vogelterritorium) können die Werte, die es bzgl. der neu definierten Hauptkomponenten annimmt, vom Statistikprogramm berechnet werden (Komponentenwert oder „score"). Häufig werden diese Komponentenwerte dann im durch die HK definierten Merkmalsraum (bei 3 HK) oder in der Merkmalsebene (bei 2 HK) in Form eines sog. Streudiagramms („scatter plot") dargestellt. Manchmal lassen sich aus dieser graphischen Darstellung Beziehungen zwischen einzelnen Untersuchungsobjekten (z. B. Territorien unterschiedlicher Vogelarten) herauslesen, die im ursprünglichen, viel-dimensionalen Datenmaterial der Originalvariablen nicht erkennbar waren. Selbstverständlich lassen sich damit aber immer nur maximal die ersten drei HK einer Analyse darstellen.

Indikatoren für die Qualität einer Hauptkomponentenanalyse

Die geometrische Ableitung der Hauptkomponenten diente der Veranschaulichung eines komplexen mathematischen Vorgangs, der hinter der Hauptkomponentenanalyse tatsächlich steht und an dieser Stelle nicht vertieft werden kann. Grundlage für die HKA und Ausgangspunkt für weitere berechnete Matrices ist im allgemeinen die Korrelationsmatrix (andere Unähnlichkeitsmatrices wie z. B. die Kovarianzmatrix spielen in der ökologischen Forschung eine untergeordnete Rolle). In dieser Korrelationsmatrix wird für jedes Variablenpaar der Produkt-Moment-Korrelationskoeffizient als Maß für die Stärke und Richtung des Zusammenhangs von zwei Variablen berechnet. Es ist nachvollziehbar, dass eine Korrelationsmatrix, in der nur niedrige Korrelationswerte vorhanden sind, nicht für eine HKA geeignet ist. Auf unser geometrisches Beispiel übertragen würde das bedeuten, dass die Punktwolke unkoordiniert in alle Dimensionen ausstrahlt und keine Lösung durch projizierte, neue Achsen besser als die Originallösung wäre. In der Praxis sind jedoch „griffigere" Indikatoren nötig, die Auskunft darüber geben, ob eine HKA überhaupt sinnvoll ist oder ob bestimmte Variablen aus der Analyse ausgeschlossen werden sollten, um die Qualität der HKA zu verbessern.

MSA	Beurteilung
≥ 0,9	erstaunlich
≥ 0,8	verdienstvoll
≥ 0,7	ziemlich gut
≥ 0,6	mittelmäßig
≥ 0,5	kläglich
< 0,5	untragbar

Tab.1: Beurteilung der „Measure of Sampling Adequacy" (nach Backhaus et al. 1994). – Assessment of „Measure of Sampling Adequacy".

Das beste zur Verfügung stehende Maß ist das sog. „Kaiser-Meyer-Olkin-Kriterium" (KMO), auch „Measure of Sampling Adequacy" (MSA) oder „Maß der Stichprobeneignung" genannt. Es deckt einen Wertebereich von 0 bis 1 ab (Tab.1) und ist sowohl für die Gesamtmatrix als auch für jede einzelne Variable verfügbar. Es sollte den Wert 0,5 nicht unterschreiten. Variablen mit geringeren Werten sollten aus der Analyse ausgeschlossen werden, falls ihr Verbleib in der Analyse nicht aus sachlogischen Überlegungen heraus erwünscht wird. Bei einem MSA-Wert unter 0,5 für die Gesamtmatrix sollte die ganze HKA in Frage gestellt werden, sofern der Wert durch den sukzessiven Ausschluss ungeeigneter Variablen aus der HKA nicht erhöht werden kann. Der Ausschluss einer Variablen impliziert nicht, dass diese Variable generell „schlecht" sei. Er besagt lediglich, dass diese Variable zu eigenständig für eine HKA ist, bei der sich ja verschiedene Variablen gegenseitig erklären sollen. Eine ausgeschlossene Variable besitzt somit möglicherweise (!) schon für sich genommen einen hohen Informationsgehalt.

Der Anteil der durch eine Hauptkomponente erklärten Varianz wird als Eigenwert der HK bezeichnet. Dieser errechnet sich aus der Summe der quadrierten Komponentenladungen einer Komponente über alle Variablen. Sofern als Grundlage der HKA die Korrelationsmatrix verwendet wurde (dieses ist i. A. zu empfehlen) entspricht der Gesamt-Eigenwert des Systems der Anzahl der Originalvariablen. Die Eigenwerte der HK sind deshalb für sich genommen wenig aussagekräftig und nur vor dem Hintergrund dieses Gesamtwertes interpretierbar. Eine weit verbreitete Regel besagt, dass nur solche HK akzeptiert werden, die einen Eigenwert >1 haben, weil sie dann einen größeren Erklärungsgehalt als eine Originalvariable haben. Man bezeichnet diese Regel als „Kaiser-Kriterium". Es gibt noch andere Kriterien, nach denen die Anzahl der zu akzeptierenden HK festgelegt wird, die jedoch in der Praxis weniger häufig eingesetzt werden (vgl. McGarigal et al. 2000). Man kann davon ausgehen, dass in einem viel-dimensionalen Datensatz mit deutlich korrelierenden Originalvariablen bereits die ersten zwei bis vier HK den größten Teil (häufig >90 %) der Varianz des Originaldatensatzes reproduzieren.

Die sog. Kommunalitäten sind eine weitere Entscheidungshilfe für eine mögliche Elimination einzelner Variablen aus der Analyse. Sie sind für jede Ausgangsvariable verfügbar und geben Auskunft über den Umfang der Varianzerklärung, den die Hauptkomponenten gemeinsam für eine Ausgangsvariable liefern. Der Kommunalitäten-Wert einer Variablen ergibt sich aus der Summe der quadrierten Kompo-

nentenladungen dieser Variablen über die akzeptierten Hauptkomponenten (vgl. Kaiser-Kriterium). Ein Wert von 0,42 für eine Variable besagt, dass lediglich 42 % der ursprünglich in dieser Variable steckenden Varianz durch die akzeptierten HK reproduziert werden. Ein niedriger MSA-Wert gemeinsam mit einem niedrigen Kommunalitäten-Wert sollte zum Ausschluss der entsprechenden Variable aus der HKA führen. Eine HKA bringt deshalb u.U. erst nach mehreren Durchläufen unter leicht veränderten Rahmenbedingungen zufriedenstellende Ergebnisse.

Hauptkomponentenanalyse oder Faktorenanalyse?
Diese beiden Verfahren werden begrifflich nicht immer sauber getrennt. In manchen Statistik-Programmen (z. B. SPSS) wird die HKA als Option innerhalb der Faktorenanalyse (FA) geführt und dann der sog. Hauptachsenanalyse gegenübergestellt. In den meisten Fällen wird jedoch die HKA als eigenständiges Verfahren geführt. Das „Gegenstück" ist dann die Faktorenanalyse. Rechnerisch unterscheiden sich beide Verfahren nur in einem wesentlichen Punkt, nämlich in der Berechnung der Kommunalitäten. Ansonsten sind sie rechnerisch als identisch zu bezeichnen (Backhaus et al. 1994: 221).

Die FA (bzw. Hauptachsenanalyse) geht davon aus, dass in jeder Ausgangsvariablen auch eine sog. Einzelrestvarianz (spezifische Varianz + Messfehlervarianz einer Variable) vorhanden ist, die von den anderen Variablen nicht reproduziert werden kann. Die Startwerte der Kommunalitäten – d. h. die Kommunalitäten, die sich ergeben, wenn genauso viele Faktoren (so werden die Gegenstücke der Hauptkomponenten genannt) wie Ausgangsvariablen definiert werden – werden dann nach bestimmten Kriterien geschätzt und liegen immer unter dem Wert 1. Nur dieser Teil der Varianz kann maximal durch die Faktoren erklärt werden. Der Endwert der Kommunalitäten nimmt noch einmal ab, weil i.d.R. nicht alle rechnerisch möglichen Faktoren akzeptiert werden (vgl. Kaiser-Kriterium).

Die HKA geht dagegen davon aus, dass die Varianz der Ausgangsvariablen vollständig durch die Extraktion der Hauptkomponenten erklärt werden kann. Wenn also genauso viele HK wie Ausgangsvariablen definiert werden, liegen die Startwerte der Kommunalitäten für jede Variable bei dem Wert 1. Da aber i.d.R. weniger HK akzeptiert werden (s.o.), liegen auch bei diesem Verfahren die Kommunalitäten normalerweise unter dem Wert 1. Der nicht reproduzierte Varianzanteil wird als bewusst in Kauf genommener Informationsverlust deklariert. Das Ziel der HKA ist die möglichst umfassende Reproduktion der Varianz der Ausgangsvariablen durch die Hauptkomponenten. Die HKA erlaubt deshalb streng genommen im Gegensatz zur FA auch keine kausale Interpretation der Hauptkomponenten (Backhaus et al. 1994: 221-222). Für weitere Informationen verweise ich auf die in der Einleitung zitierten Lehrbücher zur multivariaten Statistik.

Fallbeispiel Teil 1, Hauptkomponentenanalyse
Ausgangssituation: Jedes Grasmücken-Territorium ist u.a. durch die Vegetationsmenge in 8 verschiedenen horizontalen Höhenschichten zwischen 0,3 und 12 m gekennzeichnet (vgl. Kap. 2.1). Es ist anzunehmen, dass in diesen Schichtvariablen ein großes Maß an Redundanz (d. h. überflüssige Information) steckt, denn die Unterscheidung von 8 Höhenschichten ist eigentlich sehr willkürlich. Es hätten auch 20 Schichten oder auch nur 3 Schichten definiert werden können. Die Wahrscheinlichkeit, dass in unserem konkreten Beispiel benachbarte Schichten ähnliche Informationen liefern, wird umso größer, je näher die untersuchten Schichten vertikal beieinander liegen.

Ziel: Es soll untersucht werden, ob der Informationsgehalt dieser 8 Schichtvariablen bei einem akzeptablen Informationsverlust auf eine geringere Anzahl von (komplexeren) Hauptkomponenten übertragen werden kann. Ohne eine solche Elimination redundanter Informationen besteht die Gefahr, dass die schichtbezogene Vegetationsmenge, die in diesem Ansatz ja nur einen Aspekt der Vegetationsstruktur eines Vogelterritoriums darstellt, durch immerhin 8 Variablen unverhältnismäßig stark gewichtet wird gegenüber den anderen hier berücksichtigten Aspekten der Vegetationsstruktur (Aufbau des Vegetationsraumes aus unterschiedlichen Untereinheiten), die durch lediglich 4 Variablen repräsentiert werden.

Vorgehensweise (vgl. Output 1): Eine an dieser Stelle nicht weiter ausgeführte Voranalyse hat ergeben, dass die Schichtvariablen S7 (7 m Höhe) und S8 (12 m Höhe) für sich genommen sehr eigenständig sind und deshalb nicht in eine HKA eingehen sollten. Beide Variablen zeigten niedrige Werte bei MSA und Kommunalitäten. Der Grund dafür ist, dass die Höhenschichten S7 und S8 vertikal wesentlich isolierter sind (2,5 bzw. 5 m vertikaler Abstand zur nächsten untersuchten Schicht) als die Schichten S1 bis S6, deren vertikaler Abstand untereinander zwischen 50 cm und 1 m liegt. Es gehen also nur die Variablen S1 bis S6 in die HKA ein. In diesem Stadium der Untersuchung wird die Artzugehörigkeit der Territorien noch nicht berücksichtigt.

Die Variablenwerte wurden zuvor z-standardisiert (z-S1 bis z-S6). Grundlage der HKA ist die Korrelationsmatrix. Es treten teilweise hohe Korrelationen zwischen den Variablenpaaren auf (v.a. zwischen vertikal benachbarten Vegetationsschichten). Die MSA für die Gesamtmatrix ist mit 0,66 annehmbar („mittelmäßig"). Die MSA-Werte für die Einzelvariablen sind für die Variablen S3 bis S6 „mittelmäßig" bis „verdienstvoll". Bei den Variablen S1 und S2 liegen sie zwar noch über dem kritischen Wert von 0,5, gelten aber als „kläglich" (vgl. Tab.1). Da die Werte der Kommunalitäten bei beiden Variablen mit über 97 % sehr hoch sind und aus inhaltlichen Überlegungen der Verbleib der untersten beiden Schichtvariablen in der

Analyse erwünscht wird, werden sie trotz der geringen MSA-Werte beibehalten. Zwei Hauptkomponenten mit einem Eigenwert >1 („Kaiser-Kriterium“) wurden extrahiert. Sie reproduzieren gemeinsam 93,9 % der Gesamtvarianz des untersuchten Systems. Die erste Hauptkomponente erklärt stets den entlang einer Dimension maximal möglichen Anteil der Varianz (vgl. Abb. 1) und hat deshalb den größten Eigenwert (3,79 bzw. 63,2 % der Gesamtvarianz). Nach Durchführung einer Varimax-Rotation zum Zwecke einer verbesserten inhaltlichen Interpretierbarkeit der HK ist der Eigenwert der ersten HK kleiner geworden. Dieses Phänomen ist ein Automatismus nach jeder Rotation und lässt sich sehr einfach erklären: Da die (unrotierte) Originallösung immer die Optimallösung für die Definition der ersten HK ist, muss jede Abweichung davon zwangsläufig zu einer Verkleinerung des Eigenwertes führen. Man erkennt dieses daran, dass ein Teil des Eigenwertes der ersten HK nach der Rotation auf die zweite HK übertragen wurde (Eigenwert der ersten HK nach Rotation: 3,33 bzw. 55,5 %). Der Wert von 93,9 % erklärter Gesamtvarianz bleibt dagegen unverändert.

Ein Vergleich der nicht-rotierten Komponentenladungsmatrix mit der rotierten Ladungsmatrix zeigt deutlich, dass der Vorgang der Rotation tatsächlich zu einer Verbesserung der Interpretierbarkeit der Hauptkomponenten geführt hat. Dieses ist nicht zwangsläufig in jeder Analyse der Fall. Auf die erste HK laden die oberen Schichtvariablen S3-S6 sehr hoch. Die zweite HK wird v.a. durch die untersten Schichtvariablen S1-S2 (bzw. S3) erklärt. Die erste HK repräsentiert somit die Vegetationsmenge einer oberen Strauchschicht zwischen 1,5 und 4,5 m Höhe, die zweite HK dagegen die Vegetationsmenge einer unteren Strauchschicht bis 1,5 m Höhe.

Die Werte beider HK wurden für jedes Territorium berechnet. Das Streudiagramm der Komponentenwerte zeigt, dass die Dorngrasmücke tendenziell niedrigere Werte in Bezug auf HK1 (obere Strauchschicht) und gleichzeitig tendenziell höhere Werte in Bezug auf HK2 (untere Strauchschicht) hat als die Mönchsgrasmücke. Die durch die Hauptkomponentenanalyse vorgenommene Informationsverdichtung der 6 Schichtvariablen auf zwei Hauptkomponenten deutet also bereits auf eine unterschiedlich realisierte Habitatnische von Dorn- und Mönchsgrasmücke hin. Dennoch gibt es einen großen Überschneidungsbereich zwischen beiden Arten. Die Gartengrasmücke zeigt dagegen aufgrund sehr starker Überschneidung mit den Territorien der Dorngrasmücke und der Mönchsgrasmücke wenig Eigenständigkeit. Sie deckt in Bezug auf beide HK eine große Wertespanne ab. Dabei ist allerdings zu bedenken, dass die beiden HK nur einen kleinen Teil der zur Verfügung stehenden Information zur Vegetationsstruktur in den Vogelterritorien repräsentieren (Vegetationsmenge bis 4,5 m Höhe). Es ist zu erwarten, dass die Verwendung der restlichen 6 Strukturvariablen (Vegetationsmenge in 7 m und 12 m Höhe

Korrelationsmatrix – correlation matrix

	z-S1	z-S2	z-S3	z-S4	z-S5	z-S6
z-S1	1,00	0,98	0,54	0,34	0,18	-0,01
z-S2	0,98	1,00	0,55	0,37	0,23	0,01
z-S3	0,54	0,55	1,00	0,86	0,75	0,60
z-S4	0,34	0,37	0,86	1,00	0,95	0,80
z-S5	0,18	0,23	0,75	0,95	1,00	0,91
z-S6	-0,01	0,01	0,60	0,80	0,91	1,00

MSA

MSA = 0,659

Variable	MSA
z-S1	0,546
z-S2	0,551
z-S3	0,841
z-S4	0,711
z-S5	0,626
z-S6	0,669

Kommunalitäten – communality

Variable	Komm.
z-S1	0,979
z-S2	0,975
z-S3	0,852
z-S4	0,953
z-S5	0,967
z-S6	0,908

Eigenwerte – eigenvalues

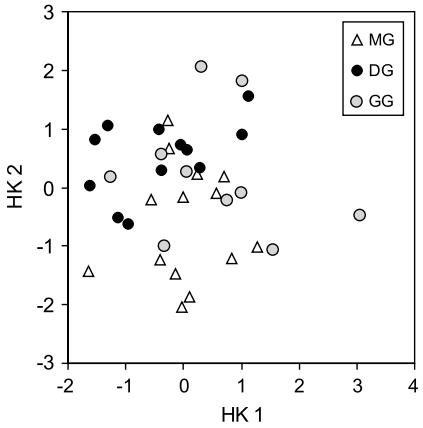
nicht rotiert – not rotated

	Eigenw.	% der Var.	kumul.%
HK 1	3,79	63,2	63,2
HK 2	1,84	30,7	93,9

rotiert – rotated

	Eigenw.	% der Var.	kumul.%
HK 1	3,33	55,5	55,5
HK 2	2,31	38,4	93,9

Streudiagramm der Komponentenwerte – scatter plot of component scores



Komponentenladungsmatrix – component loadings

	nicht rot.		rotiert	
Variable	HK1	HK2	HK1	HK2
z-S1	0,55	0,82	0,08	0,99
z-S2	0,58	0,80	0,11	0,98
z-S3	0,92	0,08	0,76	0,52
z-S4	0,95	-0,22	0,94	0,28
z-S5	0,90	-0,39	0,98	0,10
z-S6	0,77	-0,57	0,95	-0,12

Output 1: Ergebnisse der Hauptkomponentenanalyse, Fallbeispiel Teil 1. HK = Hauptkomponente, MG = Mönchsgrasmücke, DG = Dorngrasmücke, GG = Gartengrasmücke. – Results of Principal Component Analysis, Empirical Example Part 1.

und Anteile der vier unterschiedlichen Strukturtypen) zu einer besseren Abgrenzung der Habitatnische der drei Grasmückenarten führen wird.
Fazit: Die HKA wurde hier also ausschließlich zu einer Reduktion der Variablenanzahl als vorbereitende Maßnahme für weitere multivariate Verfahren (hier: Diskriminanzanalyse und Clusteranalyse) eingesetzt. Statt der ursprünglich 6 Schichtvariablen zur Beschreibung der Vegetationsmenge bis 4,5 m Höhe werden für weitere Analysen nur noch zwei Variablen (HK) verwendet. Der damit verbundene Nachteil in Form eines Informationsverlustes von etwa 6 % ist gegenüber dem Vorteil der Variablenreduktion als sehr gering einzuschätzen. Durch die Unabhängigkeit beider HK („Orthogonalität") wurde außerdem Redundanz vollständig aus dem Original-System der 6 Schichtvariablen entfernt. Die graphische Darstellung der Komponentenwerte machte deutlich, dass die beiden HK durchaus einen Beitrag zur Gruppenunterscheidung leisten.

Grenzen der Hauptkomponentenanalyse
Bei einer HKA werden nur intrinsische (d. h. dem System innewohnende) Beziehungen zwischen den vom Bearbeiter ausgewählten Variablen aufgedeckt und daraufhin die HK definiert. Es kann aber auch noch wichtige Einflussgrößen außerhalb des untersuchten Modells geben, die vom Bearbeiter einfach nicht als solche erkannt wurden und deshalb unberücksichtigt bleiben. Selbstverständlich ist das streng genommen nicht dem Verfahren der HKA selbst vorzuwerfen, aber der Bearbeiter muss sich dieser Tatsache bewusst sein. Die sorgfältige Auswahl der Variablen ist deshalb von größter Wichtigkeit. Außerdem geht die HKA generell von linearen Abhängigkeiten zwischen den Variablen aus (Grundlage ist i.A. die Korrelationsmatrix). Es gibt jedoch auch Beziehungen zwischen Variablen, die nicht-linearer Art sind. Linearität kann noch am ehesten für enge Umweltgradienten angenommen werden, so dass insbesondere bei Variablen mit einer sehr großen Spannweite die Gefahr einer Verzerrung bei der Definition der Hauptkomponenten besteht (vgl. McGarigal et al. 2000: 61).

3.2. Diskriminanzanalyse

Ziel der Analyse
Die Diskriminanzanalyse (DA) ist ein multivariates Verfahren zur Analyse von Gruppenunterschieden auf der Grundlage vorgegebener Merkmalsvariablen. Die Gruppenzugehörigkeit der zu untersuchenden Objekte wird vor der Analyse festgelegt und mit Hilfe der DA quantitativ überprüft. Sie ergibt sich beispielsweise aus der Artzugehörigkeit, dem Geschlecht oder der geographischen Verteilung von Vogelindividuen. Mit Hilfe optimierter, sog. kanonischer Diskriminanzfunktionen (DF) in Form gewichteter, linearer Kombinationen der Merkmalsvariablen werden Diskriminanzwerte für jedes Objekt berechnet. Diese Funktionen sind insofern optimiert, als die resultierenden Diskriminanzwerte innerhalb einer Gruppe möglichst wenig und zwischen den Gruppen möglichst stark streuen. Die Streuung zwischen den Gruppen wird auch als „erklärte Streuung" bezeichnet, weil sie ursächlich auf die unterschiedliche Gruppenzugehörigkeit zurückgeführt wird. Die Streuung innerhalb der Gruppen wird dagegen als „nicht erklärte Streuung" bezeichnet. Sie lässt sich selbstverständlich auch erklären (u.a. mit der natürlichen Varianz, die z. B. in Populationen oder anderen biologischen Einheiten steckt), aber eben nicht mit der Gruppenzugehörigkeit, und nur diese Eigenschaft zählt in der DA. Zur Trennung der Gruppen werden jeweils N minus 1 DF definiert, wobei N der Anzahl der Gruppen entspricht. Das Diskriminanzmodell verteilt auf der Grundlage der berechneten Diskriminanzwerte die untersuchten Objekte erneut auf die festgelegten Gruppen. Die Übereinstimmung zwischen der berechneten („vorhergesagten") und der tatsächlichen Gruppenzugehörigkeit (d. h. die Trefferquote des Modells) ist ein außerordentlich „greifbares" Maß für die Qualität des Modells.

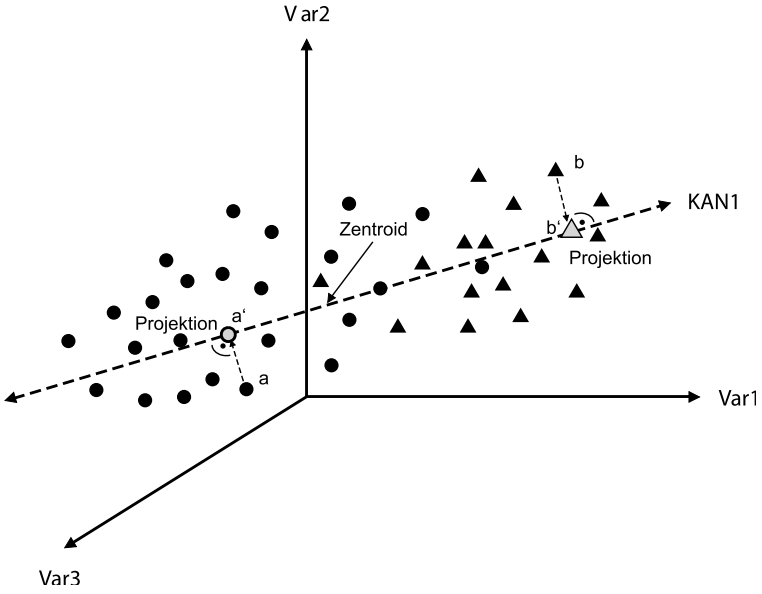


Abb. 3: Graphische Ableitung der kanonischen Achse (KAN1) aus drei Originalachsen (Var) für ein Zwei-Gruppen-Modell. Die kanonische Achse maximiert den Quotienten der Varianz zwischen den Gruppen zur Varianz in den Gruppen (nach McGarigal et al. 2000). Sie repräsentiert die Achse mit dem besten Trennvermögen für beide Gruppen, das entlang einer Dimension möglich ist. – Graphic derivation of the Canonical Axis (KAN1) from three original axes (Var) in a two-group model. The canonical axis maximizes the ratio of among-group to within-group variation, representing the axis of best segregation of both groups that is possible in a single dimension.

Was sind kanonische Diskriminanzfunktionen?
Genauso wie die Hauptkomponenten lassen sich auch die Diskriminanzfunktionen graphisch als neu definierte kanonische Achsen im N-dimensionalen Merkmalsraum darstellen. Angenommen, jeder Punkt in Abb. 3 steht für ein Territorium der Dorngrasmücke und jedes Dreieck für ein Territorium der Mönchsgrasmücke. Jedes Territorium wird z. B. durch die Deckungsgrade der Baumschicht (Var.1), Strauchschicht (Var.2) und Krautschicht (Var.3) gekennzeichnet. Die kanonische Achse wird so gelegt, dass durch Projektion der Territorien auf diese Achse möglichst viele Territorien der Dorngrasmücke unterhalb und möglichst viele Territorien der Mönchsgrasmücke oberhalb eines bestimmten Schwellenwertes auf der Achse liegen. Auf diese Weise kann aufgrund der Position der Territorien auf dieser neuen Achse deren Artzugehörigkeit unterschieden werden. Territorien, die durch Projektion auf die „falsche" Seite gelangen, werden durch das Modell falsch klassifiziert. Wie bei der Hauptkomponentenanalyse ist durch die Definition einer neuen Achse aus einem dreidimensionalen System ein eindimensionales System geworden.

Tatsächlich wird die Gruppenzugehörigkeit selbstverständlich nicht auf graphischem Wege, sondern rechnerisch definiert. Die graphisch abgeleitete, kanonische Achse entspricht rechnerisch einer Diskriminanzfunktion. Die durch Projektion hervorgerufenen Positionen der Territorien auf der neuen Achse ergeben sich aus der Berechnung von Diskriminanzwerten auf der Grundlage der Diskriminanzfunktion, die dahingehend optimiert ist, dass sich die berechneten Diskriminanzwerte der Territorien der beiden Grasmückenarten möglichst stark unterscheiden (hier: Dorngrasmücke: niedrige Werte, Mönchsgrasmücke: hohe Werte). Der Unterschied zwischen einer kanonischen Achse und einer Hauptkomponente besteht also darin, dass die kanonische Achse nicht so gelegt wird, dass sie ein Maximum an Varianz reproduziert, sondern dass der Quotient der Varianz zwischen den Gruppen („erklärte Varianz") zur Varianz innerhalb der Gruppen („nicht erklärte Varianz") maximiert wird.

Wenn bei drei vorgegebenen Gruppen (z. B. durch die zusätzliche Analyse von Gartengrasmücken-Territorien) zwei DF definiert werden (s. oben), wird analog zur Hauptkomponenten-Extraktion die zweite kanonische Achse so gelegt, dass sie orthogonal auf der ersten kanonischen Achse liegt und die durch die erste Achse nicht erklärten Gruppenunterschiede maximal reproduziert. Bei höherer Gruppenanzahl werden weitere kanonische Achsen unter den gleichen Bedingungen definiert. Orthogonalität bedeutet auch hier Unabhängigkeit des Informationsgehaltes der kanonischen Achsen. Die diskriminatorische Arbeit wird somit überschneidungsfrei auf die DF verteilt.

Es gibt zwei Möglichkeiten, die kanonischen DF zu definieren. Normalerweise werden alle Variablen gleichzeitig aufgenommen. Der Bearbeiter muss in diesem Fall besonders sorgfältig bei der Auswahl der Merkmalsvariablen sein und nur solche Variablen auswählen, die mutmaßlich zwischen den Gruppen variieren. Eine alternative Methode ist die schrittweise Aufnahme von Variablen in das Diskriminanzmodell. Hierbei sucht sich das System nach unterschiedlichen, vom Bearbeiter festzulegenden Gütekriterien in mehreren Durchläufen aus den vorgeschlagenen Variablen nur diejenigen mit der besten diskriminatorischen Wirkung aus.

Inhaltliche Interpretation
Ähnlich wie bei den Hauptkomponenten ist auch bei den kanonischen Achsen die relative Lage zu den Originalachsen des Systems entscheidend für die inhaltliche Interpretation. Quantitativer Ausdruck des Erklärungsgehaltes der Ausgangsvariablen für das Diskriminanzmodell sind die Diskriminanzkoeffizienten der DF. Eine DF hat allgemein folgende Form:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_j X_j$$

wobei
Y = Diskriminanzvariable
X_j = Merkmalsvariable j (j = 1,2,…,J)
b_j = Diskriminanzkoeffizient für Merkmalsvariable j
b_0 = konstantes Glied

Hohe Koeffizienten deuten auf eine hohe Trennkraft der entsprechenden Variablen hin. Hierbei ist allerdings Vorsicht geboten. Bei Korrelationen zwischen den Ausgangsvariablen („Multikollinearität") kann nicht ohne weiteres von den einzelnen Koeffizienten auf deren Erklärungsgehalt geschlossen werden, weil möglicherweise der Einfluss einer Variablen teilweise durch den Koeffizienten einer anderen Variablen berücksichtigt wird (Brosius & Brosius 1995: 781). Wenn also nicht sichergestellt ist, dass die Ausgangsvariablen weitgehend unabhängig voneinander sind (z. B. durch eine vorgeschaltete Hauptkomponentenanalyse), sollte auf eine derartige Interpretation verzichtet werden. In einem solchen Fall ist die sog. Strukturmatrix aussagekräftiger, in der die Korrelationen jeder einzelnen Variablen mit den Diskriminanzfunktionen dargestellt werden. Sie ist in ihrer Bedeutung mit der Ladungsmatrix einer Hauptkomponentenanalyse vergleichbar.

Indikatoren für die Qualität einer Diskriminanzanalyse
Jeder DF wird aufgrund ihrer diskriminatorischen Kraft ein Eigenwert zugeordnet. Der Eigenwert einer DF errechnet sich aus dem Verhältnis der Streuung zwischen den Gruppen zur Streuung in den Gruppen. Er misst somit das Ausmaß der Gruppenunterscheidung entlang der Dimension, die durch die entsprechende DF vorgegeben wird (vgl. Abb. 3). Entsprechend werden für ein gutes Diskriminanzmodell möglichst hohe

Eigenwerte gewünscht. Bei mehreren DF ist es hilfreich, den relativen Anteil jeder DF an der Gesamt-Trennkraft des Diskriminanzmodells (d. h. der Summe aller Eigenwerte) zu berechnen. Definitionsgemäß besitzt die erste DF immer den höchsten Eigenwert und trägt damit am stärksten zur Trennung der Gruppen bei. Dennoch sind diese Werte häufig sehr irreführend, denn auch eine insgesamt sehr schwache diskriminatorische Leistung einer DA bekommt den Gesamtwert 100 %.

Aussagekräftiger ist in diesem Zusammenhang der kanonische Korrelationskoeffizient. Er misst die Strenge des Zusammenhangs zwischen den Funktionswerten der DF und den verschiedenen Gruppen. Er bewegt sich zwischen den Werten 0 und 1, wobei hohe Werte eine große Streuung zwischen den Gruppen (die in einem guten Diskriminanzmodell erwünscht sind) und gleichzeitig geringe Streuung in den Gruppen anzeigen. Der quadrierte kanonische Korrelationskoeffizient repräsentiert den Anteil der erklärten Streuung an der Gesamtstreuung und gibt direkt Auskunft darüber, wie viel Prozent der gesamten kanonischen Varianz des untersuchten Systems durch Gruppenunterschiede erklärt werden kann.

Wilks' Lambda ist ein weiteres Gütemaß des Diskriminanzmodells und errechnet sich aus dem Verhältnis der nicht erklärten Streuung zur Gesamtstreuung, so dass möglichst geringe Werte gefordert werden. Wilks' Lamba ist inhaltlich sehr eng mit dem kanonischen Korrelationskoeffizienten verknüpft. Hohe Werte der einen Größe führen automatisch zu niedrigen Werten der anderen Größe. Die Summe aus dem Wert für Wilks' Lambda und dem quadrierten kanonischen Korrelationskoeffizienten ergibt immer den Wert 1. Somit sind beide Größen gewissermaßen redundant und können je nach Geschmack alternativ zur Interpretation herangezogen werden (Brosius & Brosius 1995: 779).

Am offensichtlichsten zeigt sich die Qualität eines Diskriminanzmodells in seinen Klassifizierungsergebnissen. Dabei wird jedes Objekt (z. B. Grasmücken-Territorium), dessen Gruppenzugehörigkeit (hier: Artzugehörigkeit) ja bekannt ist, auf der Grundlage der in den DF berechneten Diskriminanzwerte erneut auf die festgelegten Gruppen verteilt (z. B. hohe Werte: Mönchsgrasmücke; niedrige Werte: Dorngrasmücke). Die Grenzwerte werden vom System nach bestimmten Optimierungsregeln selbst festgelegt. Je besser das Modell ist, desto höher wird der Anteil der korrekt klassifizierten Territorien sein, bei denen die berechnete mit der tatsächlichen Artzugehörigkeit übereinstimmt.

Dabei ist zu bedenken, dass die Trefferquote eines Diskriminanzmodells unrealistisch hoch ist, wenn die Daten zur Definition der DF mit den Daten zum Testen des Modells identisch sind. Das Modell wird dann quasi an sich selbst getestet. Auf diese Weise könnte z. B. ein Territorium der Dorngrasmücke, das eine für diese Vogelart sehr ungewöhnliche Vegetationsstruktur aufweist, trotzdem vom Modell korrekt als Dorngrasmücken-Territorium klassifiziert werden, weil das Modell durch dessen bekannte Gruppenzugehörigkeit sozusagen „vorgewarnt" wurde, dass solche von der Norm abweichenden Ereignisse möglich sind. Optimalerweise sollten zum Testen des Modells neue Daten verwendet werden. Da die Datenerhebung aber normalerweise aufwendig oder teuer ist, wäre ein solches Vorgehen ein verschwenderischer Luxus, den man sich als Bearbeiter nicht leisten kann. Einen Kompromiss stellt in diesem Zusammenhang die sogenannte Kreuzvalidierung dar. Bei diesem Verfahren wird jedes Objekt durch die DF klassifiziert, die jeweils von den Variablenwerten aller anderen Objekte außer dem gerade zu klassifizierenden Objekt abgeleitet werden. Es gibt also sehr viele Diskriminanzmodelle, die untereinander sehr ähnlich sind, weil jeweils nur das gerade zu klassifizierende Objekt bei der Definition des Modells nicht berücksichtigt wurde. Durch diesen „Trick" ist eine Funktionstrennung der Objekte (Definition vs. Testen des Diskriminanzmodells) ohne den Zwang zur Erhöhung der Stichprobengröße möglich. Als Konsequenz ist die Trefferquote bei der kreuzvalidierten Klassifizierung generell niedriger als bei der direkten Klassifizierung. Der Grund dafür ist, dass bei einer Kreuzvalidierung Werte, die untypisch für eine bestimmte Gruppe sind (Extremwerte), die Lage der kanonischen Achsen nicht zu ihren Gunsten verzerren. Bei der abschließenden Klassifizierung werden solche Werte vom Modell deshalb nicht als gruppenzugehörig erkannt.

Wann liegt eine gute Klassifizierung vor? Es gibt ganz unterschiedliche Gesichtspunkte, unter denen ein Klassifizierungsergebnis eingeordnet werden kann. 1) Der Bearbeiter könnte eine feste Mindest-Trefferquote für ein erfolgreiches Diskriminanzmodell fordern (z. B. 90 %) unterhalb derer ein Modell grundsätzlich verworfen wird. 2) Man könnte bei ungleichen Gruppengrößen berechnen, welche Trefferquote eine Klassifizierung ergeben würde, bei der alle Objekte der größten Gruppe zugeordnet würden. Wenn beispielsweise 70 % aller Objekte einer Gruppe angehören, sollte die Trefferquote eines guten Modells diesen Wert deutlich übertreffen. 3) Ebenso könnte bei ungleichen Gruppengrößen berechnet werden, wie hoch die Trefferquote wäre, wenn alle Objekte zufällig proportional zu den Gruppengrößen auf die Gruppen verteilt würden. Wenn eine Gruppe 80 % und eine zweite Gruppe 20 % aller Objekte enthält, würde eine zufällige, proportionale Verteilung der Objekte zu einer Trefferquote von 68 % führen ($0{,}8^2 + 0{,}2^2 = 0{,}68$) und sollte von einem guten Modell deutlich übertroffen werden. 4) Eine weitere Möglichkeit wäre, nur die Anzahl der Gruppen zu berücksichtigen. Bei drei Gruppen und gleicher a priori-Wahrscheinlichkeit der Gruppenzugehörigkeit hätte eine zufällige Verteilung der Objekte auf die Gruppen eine Trefferquote von 33,3 %. Man würde sich dann nur mit einer Trefferquote zufrieden geben, die deutlich darüber liegt. 5) Manch-

mal steht aber nicht die Gesamt-Trefferquote eines Diskriminanzmodells im Vordergrund, sondern es kommt darauf an, gerade in den kleinen Gruppen oder ganz bestimmten Gruppen gute Klassifizierungsergebnisse zu erzielen. Unter diesem Gesichtspunkt könnte auch eine vergleichsweise niedrige Trefferquote den Bearbeiter zufrieden stellen.

Fallbeispiel Teil 2, Diskriminanzanalyse
Ausgangssituation: Die Vegetationsstruktur jedes Vogelterritoriums wird durch folgende acht (z-standardisierte) Variablen gekennzeichnet (vgl. Kap. 2.1): z-HK1 (Hauptkomponente zur Vegetationsmenge in der oberen Strauchschicht von 1,5 bis 4,5 m Höhe), z-HK2 (dito für die untere Strauchschicht bis 1,5 m Höhe), z-S7 (Vegetationsmenge in 7 m Höhe), z-S8 (dito in 12 m Höhe), z-OHNE, z-PUNKT, z-LINIE, z-FLÄCHE (jeweilige Anzahl der 4 verschiednen Strukturtypen im dreidimensionalen Erfassungsgitter). Die beiden Hauptkomponenten sind im Rahmen der Hauptkomponentenanalyse entstanden und repräsentieren die Originalvariablen z-S1 bis z-S6 (vgl. Fallbeispiel, Teil 1).

Ziel: Es soll mit Hilfe einer Diskriminanzanalyse geklärt werden, ob Mönchsgrasmücke, Dorngrasmücke und Gartengrasmücke im untersuchten Lebensraum eine unterschiedliche Habitatnische einnehmen. Im Falle einer Habitatsonderung dieser drei Vogelarten sollten sich ihre Territorien aufgrund der Vegetationsstruktur interspezifisch deutlicher unterscheiden als intraspezifisch. Quantitativer Ausdruck dafür ist ein gutes Diskriminanzmodell, welches auf der Grundlage berechneter Diskriminanzwerte mit hoher Trefferquote die Territorien jeweils der korrekten Vogelart zuordnet.

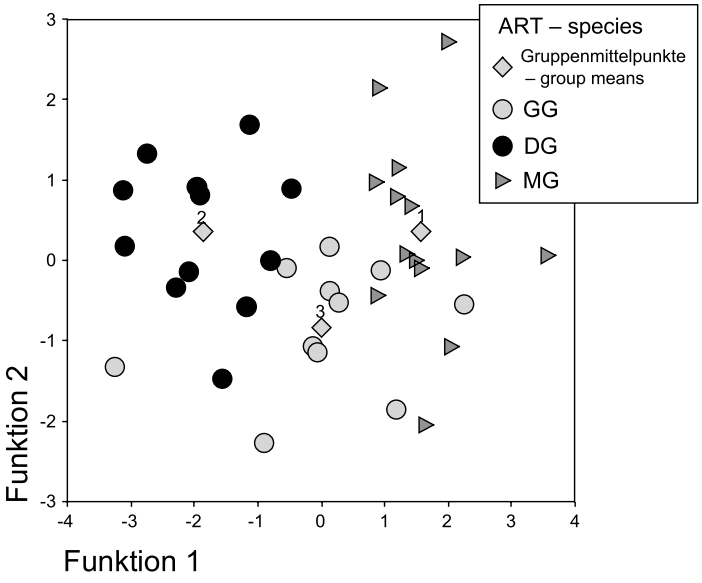
Vorgehensweise (vgl. Output 2): In diesem Diskriminanzmodell ergibt sich die zu überprüfende Gruppenzugehörigkeit der Territorien aus der Artzugehörigkeit. Bei drei Vogelarten (und folglich drei Gruppen) werden zwei Diskriminanzfunktionen erstellt. Die erste DF hat einen Eigenwert von 2,246 und ist für 87,5 % der Trennkraft des Diskriminanzmodells verantwortlich. Dagegen gehen gerade mal 12,5 % der Trennung bei einem Eigenwert von 0,32 auf die zweite DF zurück. Daraus kann zwar geschlossen werden, dass die erste DF einen wesentlich besseren „Job" macht als die zweite DF, aber nicht, wie gut das Diskriminanzmodell insgesamt ist, denn die Gesamt-Trennkraft eines Modells liegt immer bei 100 % (egal wie schlecht oder gut das Modell ist, s. oben). Hier hilft der quadrierte kanonische Korrelationskoeffizient weiter. Er liegt für die erste DF bei dem Wert 0,692. Das bedeutet, dass 69,2 % der durch diese DF hervorgerufenen kanonischen Varianz der Diskriminanzwerte durch Gruppenunterschiede erklärt werden kann. Für

Eigenwerte – eigenvalues

Funkt.	Eigenw.	% der Var.	kumul.%	quadr. kan. Korr.
1	2,246	87,5	87,5	0,692
2	0,320	12,5	100	0,242

Wilks' Lambda
bei Test der Funktionen 1-2 –
when testing functions 1-2:
0,233

Streudiagramm der Gruppen –
scatter plot of groups



Klassifizierungsergebnisse –
classifcation results

in Klammern Angaben in % – in brackets %-values

a) direkt – direct

Gruppenzughörigkeit – group membership tatsächlich – real	vorhergesagt – predicted MG	DG	GG
MG 14 Fälle – cases	**12 (85,7)**	0 (0,0)	2 (14,3)
DG 12 Fälle – cases	0 (0,0)	**11 (91,7)**	1 (8,3)
GG 11 Fälle – cases	2 (18,2)	1 (9,1)	**8 (72,7)**

Trefferquote: 83,8% – correct classification rate

b) kreuzvalidiert – cross validated

Gruppenzughörigkeit – group membership tatsächlich – real	vorhergesagt – predicted MG	DG	GG
MG 14 Fälle – cases	**11 (78,6)**	0 (0,0)	3 (21,4)
DG 12 Fälle – cases	0 (0,0)	**8 (66,7)**	4 (33,3)
GG 11 Fälle – cases	4 (36,4)	3 (27,3)	**4 (36,4)**

Trefferquote: 62,2% – correct classification rate

Output 2: Ergebnisse der Diskriminanzanalyse, Fallbeispiel Teil 2. MG = Mönchsgrasmücke, DG = Dorngrasmücke, GG = Gartengrasmücke. – Results of Discriminant Analysis, Empirical Example Part 2.

die zweite DF liegt dieser Wert nur bei 24,2 %. Wilks' Lamda (als das Gegenstück zum quadrierten kanon. Dikriminanzkoeffizienten) liegt bei Berücksichtigung beider DF bei 0,233 und ist, wie gefordert, relativ klein. Diese Ergebnisse deuten insgesamt auf ein mäßig gutes Diskriminanzmodell hin („könnte besser sein, aber auch schlechter!"). Weiteren Aufschluss über die Qualität des Modells geben die Klassifizierungsergebnisse.

Bei der direkten Klassifizierung der Territorien aufgrund der Diskriminanzwerte ergibt sich eine Trefferquote von 83,8 %. Insgesamt 6 Territorien wurden falsch klassifiziert. Beispielsweise wurden zwei Territorien der Mönchsgrasmücke aufgrund der vom Modell berechneten Diskriminanzwerte der Gartengrasmücke zugeordnet. Da die direkte Klassifizierung jedoch im Allgemeinen zu Ergebnissen führt, die zu positiv ausfallen (s. oben), sollte man sich besser an die Ergebnisse der kreuzvalidierten gruppierten Fälle halten. Bei diesem strengeren Verfahren werden 14 Territorien falsch klassifiziert. Dieses entspricht einer Trefferquote von nur noch 62,2 %. Diese Quote ist nicht per se bei einem Drei-Gruppen-Modell abzulehnen (vgl. unterschiedliche Kriterien zur Einschätzung von Klassifizierungsergebnissen im vorausgehenden Kapitel). Bei der oben genannten Fragestellung dieser Teiluntersuchung sollten dann aber alle drei Vogelarten eine vergleichbare Trefferquote aufweisen. Es wird hier jedoch bei genauerer Betrachtung der Klassifizierungstabelle deutlich, dass insbesondere die Gartengrasmücke für Konfusion sorgt und das Ergebnis negativ beeinflusst. Die fehlklassifizierten Territorien der beiden anderen Arten werden nämlich ausschließlich der Gartengrasmücke zugeordnet und lediglich 4 Territorien (36,4 %) aller Territorien der Gartengrasmücke werden korrekt klassifiziert. Einen guten optischen Eindruck über die intra- und interspezifische Ähnlichkeit der Territorien vermittelt das Streudiagramm, in dem die Diskriminanzwerte der beiden DF für jedes Territorium gegeneinander aufgetragen sind. Es ist also keine strenge interspezifische Trennung der Habitatnische für alle drei Arten nachweisbar (zumindest nicht auf der Grundlage der hier gewählten Vegetationsstruktur-Variablen). Die Gartengrasmücke liegt offensichtlich in ihrer Habitatwahl zwischen der Mönchsgrasmücke und der Dorngrasmücke, die sich untereinander gut abgrenzen lassen, und zeigt große Überschneidungen mit beiden Arten.

Fazit: Eine Gruppierung der Vogelterritorien auf der Grundlage der Artzugehörigkeit ist somit nicht zufriedenstellend. Eine abschließende Clusteranalyse soll deshalb klären, ob sich im Datenmaterial eine bessere Gruppierung finden lässt.

Grenzen der Diskriminanzanalyse
Wie schon bei der HKA gilt auch für die DA, dass nur intrinsische Abhängigkeiten für die Definition der kanonischen Achsen berücksichtigt werden können. Auch hier wird die große Bedeutung einer gut durchdachten Variablenauswahl durch den Bearbeiter deutlich. Außerdem ist die Interpretierbarkeit der DF bei Multikollinearität zwischen den Ursprungsvariablen auch bei einem guten Modell stark eingeschränkt. (McGarigal et al. 2000: 180).

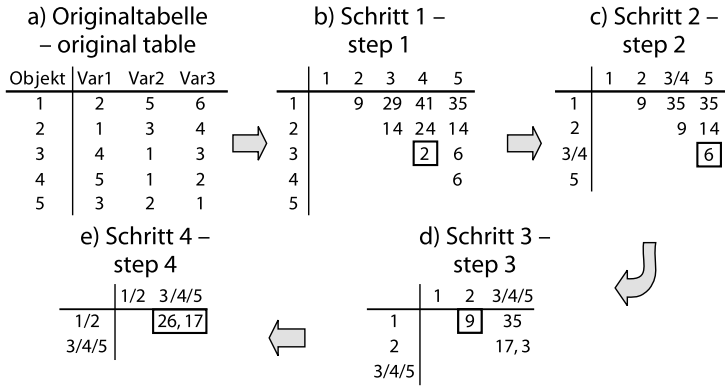
3.3. Clusteranalyse

Ziel der Analyse
Das Ziel einer Clusteranalyse (CA) besteht in der Gruppierung von Objekten aufgrund vorgegebener Merkmalsvariablen. Im Gegensatz zur Diskriminanzanalyse ist die Gruppenzugehörigkeit der Objekte also a priori nicht bekannt und soll erst durch die CA herausgefunden werden. Unter dem Begriff CA werden eine ganze Reihe ähnlicher Verfahren zusammengefasst, deren Vorstellung an dieser Stelle zu weit führen würde (vgl. McGarigal et al. 2000: 90). Die Variante mit der wohl größten Bedeutung für die ökologische Forschung ist die hierarchische, agglomerative CA. Bei den agglomerativen Verfahren sind zu Beginn der Analyse alle Objekte getrennt und vereinigen sich schrittweise zu höheren Klassen. Dieses geschieht auf der Grundlage der zwischen den Objekten berechneten Distanz bzw. Ähnlichkeit bzgl. ihrer Variablenwerte. Das Vereinigungsniveau wird bei der hierarchischen CA verwendet, um die Beziehungen innerhalb und zwischen den einzelnen Gruppen zu quantifizieren. Dargestellt wird diese Hierarchie meistens in baumartigen Diagrammen (sog. „Dendrogrammen").

Wie entstehen hierarchische Cluster?
Zur Veranschaulichung der Arbeitsweise einer CA diene eine einfache, hypothetische Datenmatrix (Tab. 2a), in der 5 Objekte (z. B. fünf verschiedene Untersuchungsflächen) durch drei Variablen (z. B. Häufigkeit von drei Vogelarten) charakterisiert werden. Im ersten Schritt wird auf der Grundlage einer vom Bearbeiter auszuwählenden Methode die Distanz der einzelnen Objekte untereinander berechnet. Aus didaktischen Gründen wird hier als sehr einfaches Distanzmaß die sog. „Quadrierte Euklidische Distanz" (QED) verwendet. Sie entspricht der Summe der quadrierten Differenzen zwischen den einzelnen Wertepaaren zweier Objekte für jede Variable. Die Distanz zwischen den Objekten 1 und 2 errechnet sich beispielsweise wie folgt:

$$QED_{1,2} = (2-1)^2+(5-3)^2+(6-4)^2 = 1+4+4 = 9.$$

Die erste abgeleitete Matrix mit den Distanzwerten für alle Objektpaare zeigt Tab. 2b. Der kleinste Wert (d. h. die geringste Distanz und somit größte Ähnlichkeit) wurde für die Paarung Objekt 3/Objekt 4 berechnet. Diese beiden Objekte werden somit als erste der 5 untersuchten Objekte zu einem „Minicluster" vereinigt, und zwar auf dem berechneten Distanzniveau von 2. Vor der nächsten Vereinigung ist die Frage zu klären,



Tab. 2: Agglomerationsschritte zur Clusteranalyse. Distanzmaß: Quadrierte Euklidische Distanz; Agglomerationsverfahren: Average Linkage zwischen den Gruppen (siehe Text und Abb. 4a). – Steps of agglomeration of Cluster Analysis. Distance measure: Squared Euclidean Distance; Fusion strategy: Average Linkage between groups (see text and Fig.4a).

wie die Distanz zwischen einem Cluster und einem Einzelobjekt bzw. zwischen zwei Clustern berechnet werden kann. Hier gibt es mehrere alternative, als Agglomerationsverfahren bezeichnete Methoden, aus denen vom Bearbeiter das geeignete Verfahren auszuwählen ist. Die wichtigste (und in der ökologischen Forschung wahrscheinlich auch einzig plausible) Methode ist das sog. „average Linkage zwischen den Gruppen". Dabei wird der Mittelwert aus den Distanzen aller möglichen Objektpaarungen zwischen zwei Clustern berechnet und als Repräsentant für die Distanz zwischen diesen beiden Clustern verwendet. Die Distanz zwischen Objekt 1 und dem Cluster 3/4 errechnet sich z. B. aus dem Mittelwert der Distanzen zwischen Objekt 1/Objekt 3 und Objekt 1/Objekt 4, also (29+41)/2 = 35. Aus der zweiten abgeleiteten Matrix (Tab. 2c) wird deutlich, dass die geringste Distanz nicht zwischen zwei Einzelobjekten, sondern zwischen Cluster 3/4 und Objekt 5 auftritt, welche deshalb in einem zweiten Schritt auf dem Niveau 6 vereinigt werden. Anschließend werden nach dem oben beschriebenen Agglomerationsverfahren die Distanzen zwischen den noch nicht vereinigten Objekten 1 bzw. 2 und dem neuen Cluster3/4/5 jeweils neu berechnet (Tab. 2d). Die geringste Distanz dieser dritten abgeleiteten Matrix besteht zwischen Objekt 1 und Objekt 2, die deshalb auf dem Niveau 9 zu einem Zweier-Cluster zusammengefasst werden. In einem letzten Schritt (Tab. 2e) werden Cluster3/4/5 und Cluster1/2 bei einem Distanzwert von 26,167 vereinigt. Dieser Distanzwert berechnet sich aus (29+41+35+14+24+14)/6 (vgl. Tab. 2b). Die ursprüngliche Matrix ist nun vollständig zusammengefasst.

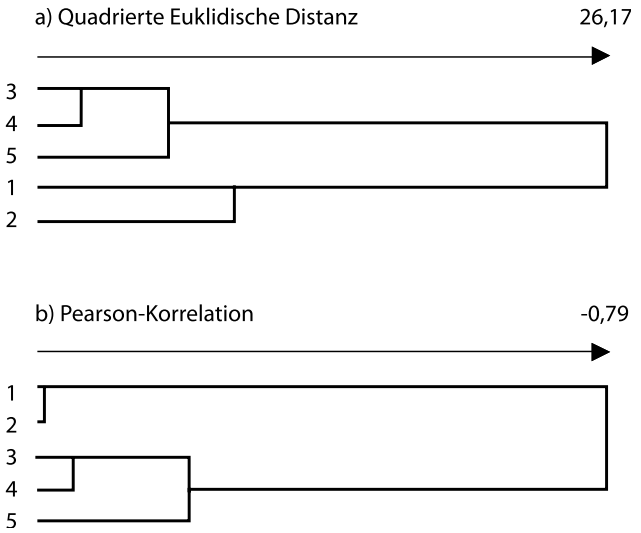


Abb. 4: Dendrogramme zur Clusteranalyse (vgl. Tab. 2a) auf der Grundlage verschiedener Maße. a) quadrierte Euklidische Distanz, b) Pearson-Korrelation. – Dendrogrammes to Cluster Analysis (cf. Tab. 2a) based on different measures. a) squared Euclidean Distance, b) Pearson Correlation

Eine graphische Darstellung dieser Agglomerationsschritte zeigt Abb. 4a in Form eines Dendrogramms. Dieses Dendrogramm ist von links nach rechts zu lesen (Es gibt auch andere Darstellungen, z. B. von unten nach oben) und zeigt durch die vertikalen Linien die Vereinigungen der Objekte an. Zu Beginn der Analyse sind alle Objekte getrennt, am Ende sind alle Objekte vereinigt. Je früher sich zwei Objekte vereinigen, desto ähnlicher sind sie in Bezug auf die verwendeten Variablen. Das Vereinigungsniveau ist aus der horizontalen Skala ablesbar. Für jede CA gibt es eine Vielzahl äquivalenter graphischer Umsetzungsmöglichkeiten, deren inhaltliche Aussage identisch ist und aus denen das Computerprogramm eine (i.d.R. sehr anschauliche) Version auswählt. So könnte die vertikale Abfolge der Objekte im Dendrogramm in Abb. 4a vielfach abgeändert und die Agglomerationsschritte trotzdem noch dargestellt werden. Z. B. könnte Objekt 5 auch oberhalb des Clusters 3/4 angeschlossen werden, oder Cluster 1/2 könnte im Dendrogramm auch oberhalb des Clusters 3/4/5 liegen. Leider kommt es bei manchen Statistikprogrammen zu nicht unerheblichen Verzerrungen bei der graphischen Ausgabe der Dendrogramme. Es wird dann zwar die Agglomerationsreihenfolge, aber nicht das Distanzniveau der Vereinigungen korrekt dargestellt. Gewissheit bekommt der Anwender über den Abgleich mit der sog. Zuordnungsübersicht, in der die Agglomerationsschritte mit den errechneten Werten schriftlich aufgeführt werden.

Es wird deutlich, dass eine Clusteranalyse von zwei Dingen wesentlich beeinflusst wird: Zum einen von der Wahl eines bestimmten Distanz- oder Ähnlichkeitsmaßes und zum anderen von der Wahl eines Agglomerationsverfahrens.

1) Auswirkungen der Wahl eines Distanz- oder Ähnlichkeitsmaßes: Der Bearbeiter hat die Wahl zwischen verschiedenen Distanzmaßen (z. B. Euklidische Distanz, Quadrierte Euklidische Distanz, Manhattan-Distanz)

und – sozusagen als Gegenstück – sog. Ähnlichkeitsmaßen (z. B. Pearson-Korrelation, Kosinus). Ähnlichkeitsmaße messen die Übereinstimmung der Objekte; Distanzmaße messen die Unähnlichkeit. Wäre in unserem Beispiel statt der QED ein Ähnlichkeitsmaß bei der CA verwendet worden, würde die Paarung mit dem höchsten Wert zuerst vereinigt werden. Die Ergebnisse einer CA können, je nach verwendetem Distanz- oder Ähnlichkeitsmaß, für den gleichen Datensatz vollkommen voneinander abweichen! Zur Illustration diene das Dendrogramm in Abb. 4b, das statt der QED die Pearson-Korrelation verwendet hat. Nach dieser Methode sind die Objekte 1 und 2 am ähnlichsten.

Wann sollte welches Maß verwendet werden? Es gibt zwar keine verbindlichen Regeln, aber folgende Entscheidungshilfe: Abb. 5 zeigt die hypothetischen Profilverläufe von drei Objekten auf der Grundlage von vier Variablen. Wenn der absolute Abstand der Variablenwerte entscheidend für die Ähnlichkeit von zwei Objekten ist, sollte ein Distanzmaß verwendet werden. Ein Distanzmaß klassifiziert deshalb Objekt 1 und Objekt 2 als ähnlich. Wenn dagegen der primäre Ähnlichkeitsaspekt im Gleichlauf zweier Variablenprofile liegt (vgl. Objekt 1 und Objekt 3), dann ist ein Ähnlichkeitsmaß geeigneter (Backhaus et al. 1994: 277). Zwei Beispiele aus der Feldornithologie sollen demonstrieren, auf welche Weise argumentiert werden könnte:

Angenommen, dass die Fußmorphologie von verschiedenen röhrichtbewohnenden Singvögeln mithilfe einer Clusteranalyse untersucht werden soll. In diesem Fall sind wohl eher die absoluten Abstände der Variablenwerte (Zehenlängen, Lauflänge, Krallenlängen usw.) unterschiedlicher Individuen oder Arten entscheidend, weil diese Werte direkt mit der umgreifbaren Halmdicke oder Halmanzahl von verschiedenen Röhrichttypen zusammenhängen könnten. Wenn dieser Ähnlichkeitsaspekt betont werden soll (vgl. Objekt 1 und Objekt 2 in Abb. 5), wäre ein Distanzmaß angemessener. Dagegen wird in unserem Grasmücken-Fallbeispiel, bei dem die Vegetationsstruktur in den Territorien der einzelnen Vogelindividuen verglichen wird, den absoluten Variablenwerten weniger Bedeutung beigemessen, denn die absoluten Werte der Strukturvariablen werden stark von der Größe eines Territoriums beeinflusst. Ein großes Territorium hat generell höhere Variablenwerte (z. B. Vegetationsmenge in einer Vegetationsschicht) als ein kleines Territorium mit identischer Vegetationsstruktur. Wenn der Bearbeiter sich für einen Ansatz entscheidet, bei dem kleine und große Territorien als ähnlich klassifiziert werden sollen, wenn sie eine ähnliche Vegetationsstruktur aufweisen, spricht das für die Verwendung eines Ähnlichkeitsmaßes (vgl. Objekt 1 und Objekt 3 in Abb. 5).

Generell (und insbesondere bei der Verwendung von Distanzmaßen) wird auch hier wieder die wichtige Funktion der z-Standardisierung der Variablen (s. Kap. 2.2) deutlich. Ohne eine solche Transformierung

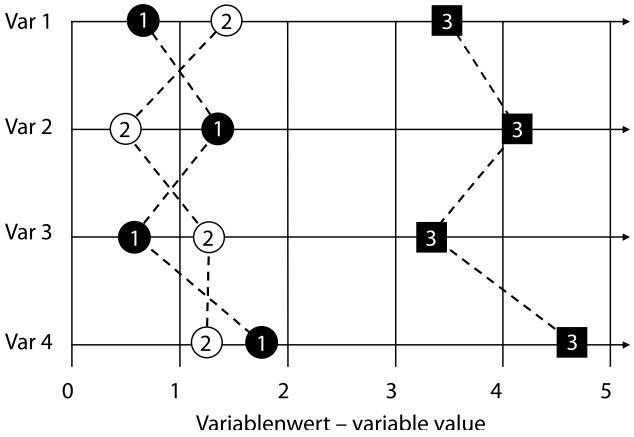


Abb. 5: Hypothetische Profilverläufe von drei Objekten auf der Grundlage von 4 Variablen (Var). Distanzmaße klassifizieren wegen der kleineren absoluten Distanzen der Variablenwerte Objekt 1 und Objekt 2, Ähnlichkeitsmaße dagegen wegen des ähnlichen relativen Profilverlaufs Objekt 1 und Objekt 3 als untereinander ähnlicher. – Hypothetic profiles of three objects based on four variables (Var). Owing to smaller absolute distances of values, distance measures classify object 1 and object 2 to be more similar to each other. By contrast similarity measures classify object 1 and object 3 more similar, because of the similar relative course of the profile of variables.

der Variablenwerte würden die Cluster nur von den Variablen bestimmt, die den größten absoluten Wertebereich abdecken.

2) Auswirkungen unterschiedlicher Agglomerationsverfahren: Auch die Wahl der Agglomerationsmethode beeinflusst die Gruppenbildung. Neben dem hier verwendeten „average Linkage zwischen den Gruppen" gibt es auch die Optionen, die Distanz des entferntesten Objektpaares zweier Cluster oder des nächstgelegenen Objektpaares als Repräsentanten für die Distanz dieser beiden Cluster zu nutzen. Diese beiden als „Complete Linkage" (oder „Furthest Neighbour") bzw. „Single Linkage" (oder „Nearest Neighbour") bezeichneten Agglomerationsverfahren führen aber zu Verzerrungen bei der Abbildung der Gruppenstrukturen der Originalmatrix (sog. „space-distorting clustering strategies", vgl. McGarigal et al. 2000: 105). Unterschiede zwischen Objekten werden überbetont bzw. abgeschwächt. Diese Verfahren sollten deshalb i.A. für ökologische Untersuchungen nicht verwendet werden, wenn nicht besondere inhaltliche Überlegungen dafür sprechen. Das „Average Linkage" wird dagegen auch als „space-conserving" bezeichnet, weil die multidimensionale Struktur der Originalmatrix im Verlauf der Agglomeration besser abgebildet wird.

Inhaltliche Interpretation

Aufgrund der hierarchischen Struktur eines Dendrogramms gibt es keine eindeutige Vorschrift, auf welcher Ebene die Cluster als relevant anzusehen sind. So können in einem Dendrogramm beispielsweise 6 kleine Cluster erkennbar sein, die sich im weiteren Verlauf zu

drei größeren Clustern vereinigen, von denen jeder aus zwei der gerade erwähnten, kleineren Cluster besteht. Ob nun aus einem solchen Dendrogramm ein Drei-Gruppen-Modell oder ein Sechs-Gruppen-Modell abzuleiten ist, kann nur der Bearbeiter selbst aus inhaltlichen Erwägungen heraus entscheiden. Dafür wird die Zusammensetzung der einzelnen Cluster analysiert. Wenn z. B. die Territorien unterschiedlicher Vogelarten einer Clusteranalyse unterzogen wurden, kann ausgezählt werden, aus welchen Vogelarten sich ein Cluster zusammensetzt. Außerdem kann untersucht werden, wie sich eine bestimmte Vogelart anteilsmäßig auf die verschiedenen Cluster verteilt.

Ein selten durchgeführter, aber sehr aussagekräftiger Ansatzpunkt für eine inhaltliche Interpretation eines Dendrogramms ist die Klärung der Frage, welche Merkmalskombination hinter den einzelnen abgegrenzten Clustern überhaupt steht. Derartige Informationen sind aus dem Dendrogramm selbst nicht ableitbar und gehören auch nicht unmittelbar in den Bereich der CA. Zu diesem Zweck werden die Variablenprofile aller klassifizierten Objekte gemeinsam in einem Diagramm dargestellt (z. B. nach dem Schema von Abb. 5), welches zusätzlich die Clusterzugehörigkeit jedes Objektes erkennen lässt. Jeder Cluster sollte dann mit einem Grundmuster eines Variablenprofils in Verbindung gebracht und somit auch inhaltlich interpretiert werden können.

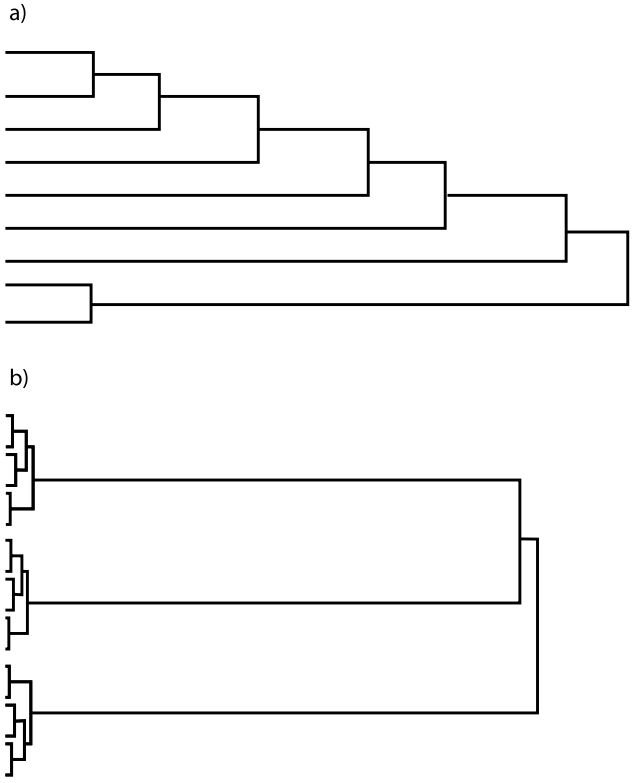


Abb. 6: Ungünstige Dendrogrammtypen aufgrund unausgewogener Gruppenstrukturen. a) gering ausgeprägte Gruppenunterschiede („Treppenbildung"), b) übermäßig stark ausgeprägte Gruppenunterschiede. – Unfavourable types of dendrogrammes owing to unbalanced group structures. a) minor group differences („stair step sequence"), b) excessive group differences.

Indikatoren für die Qualität der Clusteranalyse
Es gibt keine quantitativen Indikatoren für die Qualität einer CA, wie sie für die beiden anderen behandelten Verfahren angeboten werden. Das einzige Kriterium ist die Interpretierbarkeit des aus der CA hervorgegangenen Dendrogramms. Wie sollte ein „gelungenes" Dendrogramm aussehen?

Es sollten mehrere Cluster nebeneinander bestehen und deutlich gegeneinander abgrenzbar sein. Unerwünscht ist eine sog. „Treppenbildung" im Dendrogramm, bei der schrittweise jeweils ein Objekt an einen bestehenden Cluster angehängt wird (Abb. 6a). Eine solche Struktur spiegelt eine schwach ausgeprägte oder fehlende Gruppenstruktur im Datenmaterial wider und wird beispielsweise auch durch das Single-Linkage-Agglomerationsverfahren (s. oben) gefördert. Auch eine übertrieben deutliche Clusterstruktur, bei der sich alle Objekte eines Clusters sofort vereinigen und dann nur noch am anderen Ende des Dendrogramms mit anderen ähnlichen Clustern vereinigt werden (Abb. 6b), spiegelt eine unausgewogene Gruppenstruktur wider. In einem solchen Fall sind die Unterschiede im Datenmaterial so stark, dass eine Clusteranalyse wohl gar nicht nötig gewesen wäre. Außerdem fördert das Complete-Linkage-Agglomerationsverfahren derartige Strukturen, weil es Unterschiede im Datenmaterial überbetont.

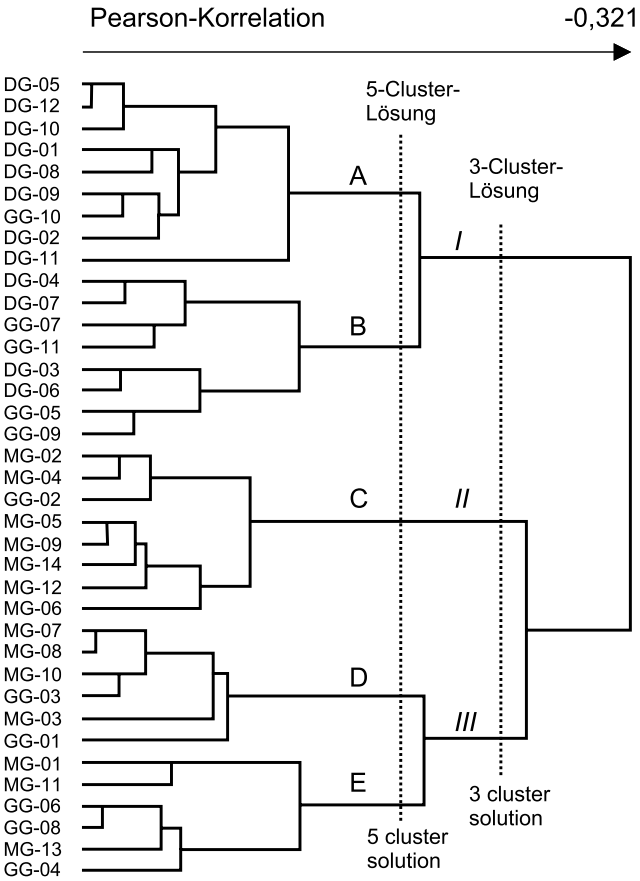
Fallbeispiel Teil 3, Clusteranalyse
Ausgangssituation: Die Diskriminanzanalyse der Territorien der drei untersuchten Vogelarten hat gezeigt, dass sich die Territorien auf der Grundlage ihrer Vegetationsstrukturen nicht zufriedenstellend interspezifisch trennen lassen (vgl. Fallbeispiel, Teil 2). Insbesondere die schlechten Klassifizierungsergebnisse der Gartengrasmücke deuteten auf Überschneidungen dieser Art mit den beiden anderen Vogelarten hin.

Ziel: Mit Hilfe der CA soll deshalb auf der Grundlage derselben Vegetationsstrukturvariablen getestet werden, wie ähnlich die Territorien in Bezug auf die Vegetationsstruktur untereinander sind.

Vorgehensweise (vgl. Output 3): Da in diesem Fall das Variablenprofil die strukturelle Zusammensetzung eines Territoriums kennzeichnet (wogegen die Absolutwerte der Variablenwerte eher von der auch intraspezifisch stark schwankenden Größe der Territorien abhängen), wird in dieser CA ein Ähnlichkeitsmaß (Pearson-Korrelation) verwendet (vgl. Abb. 5). Das Dendrogramm in Output 3 lässt deutlich 5 Cluster erkennen (Cluster A bis E). Alternativ sind auch drei Cluster abgrenzbar (I,II,III). Im folgenden wird die Fünf-Cluster-Lösung interpretiert, weil sie einen Kompromiss darstellt

zwischen der sehr groben Drei-Cluster-Lösung und einer noch höheren Clusteranzahl, die sich aus dem Dendrogramm ebenfalls herauslesen ließe, aber für „Außenstehende" nicht mehr nachvollziehbar wäre.

Cluster A und Cluster C stehen für Vegetationsstrukturen, die fast ausschließlich von Dorn- bzw. Mönchsgrasmücke bewohnt werden. In den Clustern B, D und E kommt jeweils die Gartengrasmücke hinzu. In keinem Cluster sind die Territorien von Mönchs- und Dorngrasmücke gemeinsam vertreten. Daraus kann gefolgert werden, dass sich beide Arten so deutlich in ihrer Habitatwahl unterscheiden, dass keine Überschneidungen zwischen ihnen auftreten. Die Gartengrasmücke zeigt dagegen mit beiden Arten partielle Ähnlichkeiten in der Habitatwahl und es gibt – im Gegensatz zu Mönchs- und Dorngrasmücke – keine Kombination der Vegetationsstruktur, die exklusiv nur von ihr ausgewählt wird. Dieser Eindruck wurde bereits in der Diskriminanzanalyse gewonnen. DA und CA kommen also zu ähnlichen Ergebnissen. Eine inhaltliche Analyse des für jeden Cluster typischen Variablenprofils (ohne Abb.) macht deutlich, dass hinter Cluster A die „offensten" (von der Dorngrasmücke präferiert) und hinter Cluster C die „geschlossensten" Vegetationstypen (von der Mönchsgrasmücke präferiert) entlang eines Vegetationsgradienten stehen. Diese unscharfen Begriffe können hier ohne genauere Kenntnis der Methoden (vgl. Elle 2003) nicht präzisiert werden. Die Cluster B, D und E repräsentieren unterschiedliche, bzgl. der Geschlossenheit intermediäre Vegetationstypen. Diese werden von der Gartengrasmücke bevorzugt. Außerdem dringen auch Mönchs- und Dorngrasmücke von beiden Enden des Vegetationsgradienten in diese Strukturen ein und treffen dort zwar nicht gegenseitig aufeinander, aber teilen diesen Habitattyp mit der Gartengrasmücke.



Output 3: Dendrogramm der Clusteranalyse, Fallbeipiel Teil 3. MG = Mönchsgrasmücke, DG = Dorngrasmücke, GG = Gartengrasmücke. – Dendrogramme of Cluster Analysis, Empirical Example Part 3.

Fazit: Zwei der drei untersuchten Vogelarten sind durch strenge Habitatsonderung entlang eines Vegetationsgradienten gekennzeichnet. Die Habitatpräferenzen einer dritten Art können als intermediär bezeichnet werden und vermitteln zwischen den Präferenzen der beiden anderen Vogelarten.

Anmerkung des Autors: Eine Möglichkeit, die in einer CA ermittelte Gruppierung zu bekräftigen, ist die „Überprüfung" der Gruppen durch eine Diskriminanzanalyse. Eine abschließende DA der Grasmückenterritorien, bei der nicht die Artzugehörigkeit, sondern die Zugehörigkeit zu den 5 Clustern die Gruppenzugehörigkeit bestimmte (ohne Abb.), führte zu einem sehr „treffsicheren" Diskriminanzmodell, bei dem trotz Kreuzvalidierung 91,9 % aller Territorien korrekt klassifiziert wurden. Lediglich drei Territorien wurden falsch zugeordnet.

Grenzen der Clusteranalyse

Die CA spürt v.a. sphärische Cluster auf (d. h. Cluster, die im dreidimensionalen Merkmalsraum eine im weitesten Sinne kugelförmige oder ellipsoide Gestalt haben). Relevante Muster in einer Datenmatrix müssen jedoch nicht unbedingt diese sphärische Struktur aufweisen. Es gibt viele weitere denkbare Formen (z. B. verschachtelte bogenförmige oder ringförmige Cluster), die zwar graphisch-visuell (z. B. in einem Streudiagramm) leicht wahrnehmbar sind, jedoch durch die Clusteralgorithmen nicht adäquat dargestellt werden können (vgl. McGarigal et al. 2000: 87). Außerdem ist die hierarchische CA besonders empfindlich gegen das Auftreten von Extremwerten. Häufig wird der CA auch aufgrund der vielfältigen Einflussmöglichkeiten, die ein Bearbeiter z. B. durch die Wahl verschiedener Distanzmaße und Agglomerationsverfahren auf die Analyse hat, vorgeworfen, dass sie eher eine „Kunst" als eine Wissenschaft sei. Auch die Zahl der vom Bearbeiter aus einem Dendrogramm herausgelesenen Cluster ist oft sehr subjektiv (McGarigal et al. 2000: 124).

4. Schlussbetrachtung

Die drei multivariaten Verfahren konnten hier nur ansatzweise vorgestellt werden. Vor ihrer Anwendung in

wissenschaftlichen Studien sind eine ganze Reihe von Fragen zu klären, die an dieser Stelle nur stichpunktartig genannt werden können. Jedes statistische Verfahren fußt auf bestimmten Annahmen über die Verteilung bzw. Erhebung der analysierten Daten (Multi-Normalverteilung, Zufalls-Stichprobe, Linearität von Abhängigkeiten usw.), die bei Freilanddaten i.d.R. mehr oder weniger stark verletzt werden. Hierbei ist auch zu berücksichtigen, ob ein multivariates Verfahren zu rein deskriptiven Zwecken oder im Rahmen der Inferenz-Statistik zum Zwecke der Hypothesenprüfung verwendet werden soll. Im letzteren Fall sind die Ansprüche an das Datenmaterial viel höher. Außerdem gibt es Richtlinien bzgl. der Stichprobengröße, Variablenanzahl, Gruppenanzahl, der Behandlung von Extremwerten („outlier") oder fehlenden Werten in einer Datenmatrix („missing values") usw., die vom Bearbeiter eingehalten werden sollten.

Die Wissenschaft ist geteilter Meinung über den Nutzen der multivariaten Statistik für die Ökologie. In den ausgehenden 1960-er und den 70-er Jahren stellten die ersten multivariaten Ansätze in der Ornithologie (z. B. Cody 1968, James 1971) gegenüber den bis dahin vorherrschenden qualitativen, deskriptiven Ansätzen zweifellos einen enormen Fortschritt dar (vgl. Block & Brennan 1993). Nachdem in einer anfänglichen Euphorie in diesen quantitativen Verfahren ein „Allheilmittel" der Ökologie gesehen wurde, setzte sich später mehr und mehr eine differenziertere Sicht der Dinge durch. Eine kritische Übersicht über den Gebrauch und Missbrauch multivariater Statistik in der Ökologie liefern James & McCulloch (1990). Der Einsatz multivariater Verfahren kann immer dann zu einem großen Erkenntnisgewinn führen, wenn im Rahmen einer Untersuchung viele verschiedene Variablen erhoben werden und deren Zusammenspiel evtl. noch gar nicht verstanden wird (z. B. Habitatwahl, Morphologie, Vogelzugforschung, „Community Ecology" usw.). Allerdings ist bei der Interpretation der Ergebnisse zu beachten, dass statistische Zusammenhänge (Korrelationen) noch keinen kausalen Zusammenhang (Ursache-Wirkungs-Beziehung) implizieren.

Neben den bereits genannten Arbeiten vermittelt folgende Auswahl einen Eindruck über die Einsatzmöglichkeiten multivariater Verfahren in der Ornithologie: Smith 1977, Cody 1978, Holmes et al. 1979, Bairlein 1981, Noon 1981, Glück 1983, Rice et al. 1983, Winkler & Leisler 1985, Bairlein et al. 1986, Glück & Gaßmann 1988, Mitschke 1993, Elle 2002, Mezquida 2004. Es gibt kein Patentrezept für eine gute multivariate Analyse. Eine sinnvolle Auswahl der in die Modelle eingehenden Variablen und eine gut durchdachte Datenerhebung sind generell von höchster Wichtigkeit und liegen in der Verantwortung des Bearbeiters. Es ist ein Irrglaube, dass ein schlechter Untersuchungsansatz durch multivariate Statistik „gerettet" werden kann.

5. Zusammenfassung

Diese praxisbezogene Einführung stellt Möglichkeiten und Grenzen des Einsatzes multivariater statistischer Verfahren in der Feldornithologie vor. Hauptkomponentenanalyse, Diskriminanzanalyse und Clusteranalyse gehören zu den wichtigsten multivariaten Verfahren in der ökologischen Forschung. Dieser Artikel liefert die theoretischen Grundlagen und ist gleichzeitig eine Orientierungshilfe für die Anwendung dieser Verfahren. Außerdem werden für jedes Verfahren Indikatoren für die Qualität der Analyse sowie Möglichkeiten der Interpretation diskutiert und anhand eines Fallbeispiels demonstriert.

6. Literatur

Ashcroft S & Pereira C 2003: Practical statistics for the biological sciences. Palgrave Macmillan, Basingstoke.

Backhaus K, Erichson B, Plinke W & Weiber R 1994: Multivariate Analysemethoden. Springer, Berlin.

Bahrenberg G, Giese E & Nipper J 1992: Statistische Methoden in der Geographie 2. Multivariate Statistik. Teubner, Stuttgart.

Bairlein F 1981: Ökosystemanalyse der Rastplätze von Zugvögeln: Beschreibung und Deutung der Verteilungsmuster von ziehenden Kleinvögeln in verschiedenen Biotopen der Stationen des „Mettnau-Reit-Illmitz-Programmes. Ökologie der Vögel 3: 7-137.

Bairlein F, Leisler B & Winkler H 1986: Morphologische Aspekte der Habitatwahl von Zugvögeln in einem SW-deutschen Rastgebiet. J. Ornithol. 127: 463-473.

Bärlocher F (1999): Biostatistik. Thieme, Stuttgart.

Bell SS, McCoy ED & Mushinsky HR (eds.) 1991: Habitat structure: the physical arrangement of objects in space. Chapman and Hall, London.

Bibby CJ, Burgess ND & Hill DA 1995: Methoden der Feldornithologie. Neumann Verlag, Radebeul.

Block WM & Brennan LA 1993: The habitat concept in ornithology. Theory and applications. Current Ornithol. 11: 35-91.

Brosius G & Brosius F 1995: SPSS. Base System und Professional Statistics. International Thomson Publishing, Bonn.

Cody ML 1968: On the methods of resource division in grassland bird communities. Am. Nat. 102: 107-147.

Cody ML 1978: Habitat selection and interspecific territoriality among the sylviid warblers of England and Sweden. Ecological Monographs 48: 351-396.

Deichsel G & Trampisch HJ 1985: Clusteranalyse und Diskriminanzanalyse. Fischer, Stuttgart, New York.

Digby PGN & Kempton RA 1987: Multivariate analysis of ecological communities. Chapman & Hall, London.

Elle O 2002: Mikrohabitatwahl und Dispersion als Hinweise auf interspezifische Konkurrenz von Mönchsgrasmücke *Sylvia atricapilla* und Gartengrasmücke *S. borin* in einem Wald-Wiesen-Ökoton. Vogelwelt 123: 9-16.

Elle O 2003: Quantifizierung der integrativen Wirkung von Ökotonen am Beispiel der Habitatwahl der Mönchsgrasmücke und der Dorngrasmücke (*Sylvia atricapilla* und *S. communis*, Sylviidae). J. Ornithol. 144: 271-283.

Fliege G 1986: Einführung in die Statistik für Feldornithologen. Vogelwarte 33: 257-280.

Fowler J & Cohen L (o. D.): Statistics for ornithologists. BTO Guide No 22.
Fox BJ 1979: An objective method of measuring the vegetation structure of animal habitats. Australian Wildlife Research 6: 297-303
Gauch HG 1982: Multivariate analysis in community ecology. Cambridge Univ. Press, Cambridge
Glück E 1983: Nistökologische Sonderung mitteleuropäischer Fringillidenarten im Biotop Streuobstwiese. J. Ornithol. 124: 369-392
Glück E & Gaßmann H 1988: Besiedlung von Hecken unterschiedlicher Struktur durch Vögel und ihre Nutzung als Nistsubstrat. Ökol. Vögel 10: 165-202.
Hardin G 1960: The competitive exclusion principle. Science 131: 1292-1297.
Holmes RT, Bonney RE & Pacala SW 1979: Guild structure of the Hubbard Brook bird community: a multivariate approach. Ecology 60: 512-520.
Hutchinson GE 1965: The niche: an abstractly inhabited hypervolume. In: the ecological theatre and the evolutionary play. Pp 26-78. Yale Univ. Press, New Haven.
James FC 1971: Ordinations of habitat relationships among breeding birds. Wilson Bull. 83: 215-236.
James FC & McCulloch CE 1985: Data analysis and the design of experiments in ornithology. Current Orn. 2: 1-63.
James FC & McCulloch CE 1990: Multivariate analysis in ecology and systematics: panacea or pandora's box? Annual Review of Ecology and Systematics 21: 129-166.
Kesel AB, Junge MM & Nachtigall W 1999: Einführung in die angewandte Statistik für Biowissenschaftler. Birkhäuser, Basel.
Legendre P & L Legendre 1998: Numerical ecology. Elsevier, Amsterdam.
Lorenz RJ 1992: Grundbegriffe der Biometrie. Fischer, Stuttgart.
McGarigal K, Cushman S & Stafford S 2000: Multivariate statistics for wildlife and ecology research. Springer, New York.
Mezquida ET 2004: Nest site selection and nesting success of five species of passerines in a South American open Prosopis woodland. J. Ornithol. 145: 16-22.
Mitschke A 1993: Multivariate Analysen von Brutvogelgemeinschaften im Hamburger Raum. Hamburger avifaun. Beitr. 25: 1-123.
Niemeyer H 1980: Statistische Auswertungsmethoden. In: Berthold P, Bezzel E & Thielcke G: Praktische Vogelkunde: 73-115. Kilda, Greven.
Noon BR 1981: The distribution of an avian guild along a temperate elevational gradient: the importance and expression of competition. Ecological Monographs 51: 105-124.
Odum EP 1983: Grundlagen der Ökologie, Bd.1: Grundlagen. Thieme, Stuttgart.
Rice J, Ohmart RD & Anderson BW 1983: Habitat selection attributes of an avian community: a discriminant analysis investigation. Ecological Monographs 53: 263-290.
Smith KG 1977: Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58: 810-819.
Winkler H & Leisler B 1985: Morphological aspects of habitat selection in birds. In: Cody ML (ed) Habitat selection in birds: 415-434. Academic Press, Orlando.