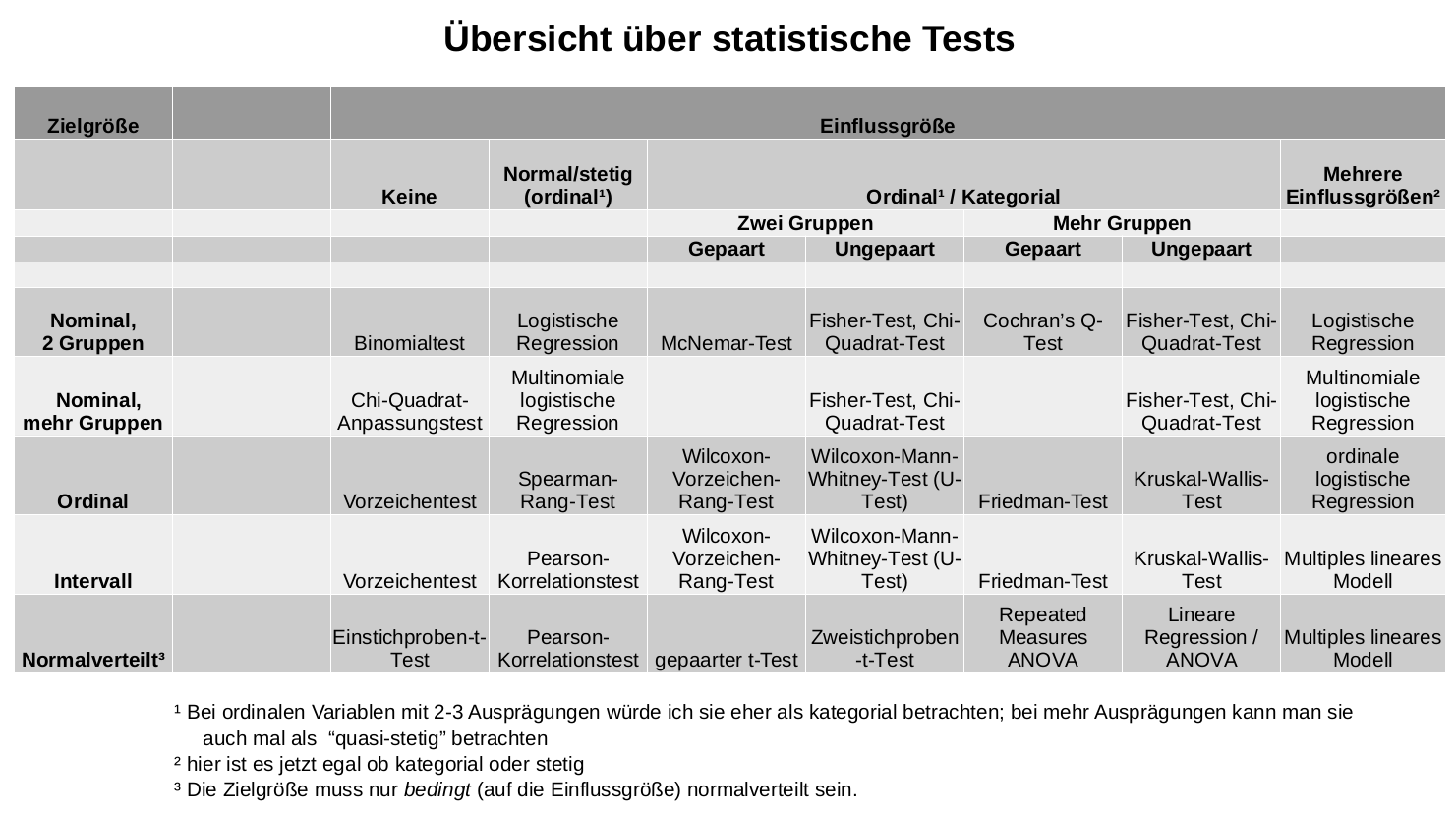
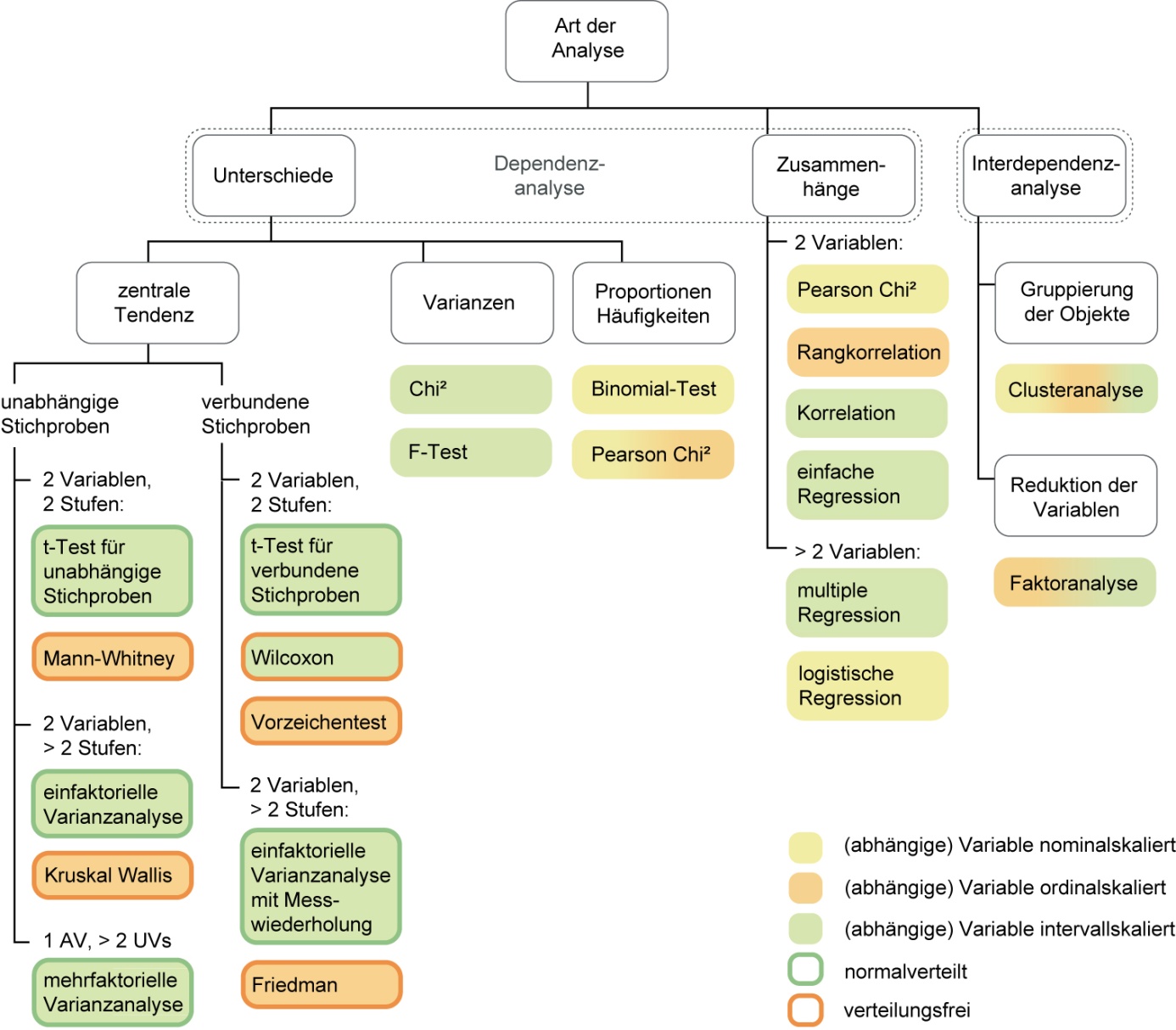
**Auswahl statistischer Testverfahren zur**



­­­­­­­

S.2/3

1.2 Uni- und multivariate Daten

**Univariat** bedeutet hier, dass für jede Stichprobeneinheit (Objekt) nur eine abhängige Variable zur Verfügung steht. Ein Beispiel für ein univariates Problem wäre, eine Eigenschaft einer Art (z. B. deren Abundanz, Gewicht, Länge etc.) zu einer Umweltvariablen, sagen wir zum pH-Wert des Bodens, in Beziehung zu setzen. Die Eigenschaft der Art ist dabei die abhängige Variable, die Umweltvariable die beeinflussende unabhängige oder erklärende Variable. In einer weiteren Studie untersuchen wir nun das Verhalten der Art nicht nur im Hinblick auf den pH-Wert, sondern beziehen auch noch die Leitfähigkeit mit ein. Beides sind Fälle für die Regressionsanalyse und gehören damit zur univariaten Statistik. Ein weiteres, noch einfacheres Beispiel ist das Gewicht von erwachse-nen Rennmäusen aus 2 verschiedenen Populationen. Die typische Frage ist dann, ob sich die Populationen signifikant unterscheiden; ein Test für Mittelwertvergleiche ist hier die Methode der Wahl. Wurden zusätzlich auch noch die Körperlängen gemessen, können diese in einem separaten Schritt ebenfalls durch Mittelwertvergleiche untersucht werden. Ein dritter Schritt könnte dann ein Test darauf sein, ob Körperlänge und Gewicht zusammenhängen (eine ziemlich triviale Frage): Dafür bieten sich simple Korrelationen an. Die eingangs genannten Beispiele sind aber anders, wie das Problem der osteuropäischen Wälder zeigt. Zu jedem Objekt, d. h. zu jeder Vegetationsaufnahme, gibt es viele abhängige Variablen, denn die Vegetationskundler haben Daten über die Häufigkeit aller vorkommenden Gefäßpflanzenarten erhoben; hinzu kommen noch die abiotischen Informationen. Das Problem ist multivariat. So etwas lässt sich mit univariater Statistik nicht mehr problemlos fassen.

**Multivariate** Daten enthalten in der Regel viele Zufälligkeiten (statistisches Rauschen). So sind z. B. nicht in jeder Vegetationsaufnahme auf stickstoffreichem Boden Brennnesseln enthalten, aber die eine oder eben auch andere Art aus einer Gruppe stickstoffliebender Pflanzen wird immer vorkommen. Es ist also nicht das Vorkommen einer einzelnen Art von Interesse, vielmehr ist das Vorkommen von Artengruppen wichtig. Darüberhinaus ist am Anfang meist nicht bekannt, welche abiotische Komponente (wenn die richtige überhaupt gemessen wurde) das Vegetationsmuster erklärt. Meist stehen dabei nicht einzelne Variablen, sondern eine Kombination aus verschiedenen Variablen in Beziehung zum Vegetationsmuster. Solche Zusammenhänge in multivariaten Daten kann die univariate Statistik (etwa durch alle möglichen wechselseitigen Korrelationen zwischen den Arten) schlecht fassen. Ein wichtiger Grund hierfür ist, dass es problematisch und aufwändig wäre, viele wechselseitige Tests durchzuführen, da solche Tests Wahrscheinlichkeiten prüfen. Das übliche Signifikanzniveau von p < 0.05 bedeutet, dass ein Zusammenhang mit 95%iger Sicherheit nicht zufällig ist. Anders ausgedrückt, wir akzeptieren, dass in 5 % der Fälle z. B. eine Korrelation als signifikant erkannt wird, obwohl sie reiner Zufall ist. Würden wir in unserem Waldbeispiel jede Art mit jeder anderen korrelieren, kämen wir leicht auf einige 100 Kombinationen, genau genommen auf (n²-n)/2, wenn n die Zahl der Arten ist. Bei dem üblichen Signifikanzniveau von p = 0.05 würden wir also im Mittel 0.05 \* (n²-n)/2 signifikante Korrelationen erwarten. Biologisch gesehen sind diese aber vermutlich bedeutungslos, sie sind ein statistisches Artefakt als Folge der vielen Berechnungen (statistical fishing). Hier braucht es andere, eben multivariate Rechenverfahren.

S.4

Methoden zur Berechnung

Die hier vorgestellten Methoden sind ausnahmslos rechenaufwändig, so dass die Nutzung geeigneter Software unumgänglich ist. Wir haben für die hier vorgestellten Analysen weitgehend die unter Ökologen verbreitete Standardsoftware benutzt. So wurden die meisten Ordinationen mit CANOCO (Version 4.5, ter Braak u. Smilauer 2002) gerechnet. Für einige weniger verbreitete Ordinationsverfahren, Permutationsverfahren und Clusteranalysen haben wir PC-ORD (Version 4.25, McCune u. Mefford 1999) verwendet, Dendrogramme wurden zum Teil auch mit MVSP (Version 3.12f, Kovach 1995) erstellt. Für die univariate Statistik haben wir ebenfalls Standardsoftware benutzt wie SPSS (Version 12.0, SPSSInc. 2003) und das sehr flexible Paket R (R Development Core Team 2004). Zu letzterem gibt es auch spezielle Erweiterungen für multivariate Fragestellungen (**VEGAN, Oksanen et al. 2006**), mit denen sich praktisch jede in diesem Buch besprochene Methode berechnen lässt.

S.57ff

5.1 Dimensionsreduktion als Analysestrategie

Als Ordinationen werden eine ganze Fülle von meist komplexen mathematischen Verfahren bezeichnet, denen gemeinsam ist, dass ein Hauptergebnis die grafische Darstellung der Daten in einem Koordinatensystem ist. Diese Darstellungen sind meist einfache Streudiagramme. Nehmen wir wieder den oben erwähnten multivariaten Datensatz, bestehend aus 3 Vegetationsaufnahmen im vordersten Bereich einer Salzwiese (Tabelle 4.3). Wir haben gesehen, dass sich die Ähnlichkeit der Aufnahmen untereinander sehr leicht grafisch darstellen lässt, indem wir sie in ein Koordinatenkreuz mit den Arten als Achsen eintragen (Abb. 4.1). Mit anderen Worten, wir bilden die Aufnahmen im Artenraum ab. Genauso schlicht ist eine Grafik, die die Lage der Arten im Aufnahmeraum abbildet (Abb. 5.1). Dieser ist wegen der 3 vorhandenen Aufnahmen jetzt 3dimensional, also komplizierter, aber durchaus noch vorstellbar. Schwierig wird es erst, wenn es mehr als 3 Arten bzw. Aufnahmen sind. Für Mathematiker sind solche mehrdimensionalen Räume unproblematisch, für den Laien wird ihre Darstellung aber leicht zu abstrakt. Wichtig an diesem Beispiel ist aber, dass sich multivariate Datensätze als n-dimensionale Hyperräume verstehen lassen. Werden z. B. Barberfallenproben im Hinblick auf ihre Artenzusammensetzung analysiert, so hat dieser Hyperraum so viele Achsen, wie es Arten in den Fallen gibt. Wenn umgekehrt Arten in ihrer Ähnlichkeit betrachtet werden sollen, dann hat der Hyperraum so viele Achsen, wie es Objekte (also Fallen) gibt. Generell spricht man von R-Analyse, wenn Objekte (z. B. Barberfallen) im Raum der ihnen zugeordneten Variablen (z. B. Arten) dargestellt werden. Werden Variablen im Objektraum betrachtet, spricht man von Q-Analyse. Oft werden Q- und R-Analysen auch als 2 Betrachtungsweisen des gleichen Problems beschrieben. Dies stimmt aber nur bedingt, denn Variablen und Objekte sind insofern unterschiedlich, als dass in einem Datensatz zwar eine Art in einer Falle fehlen kann, in der Regel aber nicht umgekehrt eine Falle nur in einer Art fehlt (aus diesem Grunde sind für Q- und R-Analysen auch unterschiedliche Ähnlichkeitsmaße geeignet, s. Kap. 4).

S.76

6.3 DCA (Detrended Correspondence Analysis)

Hill u. Gauch (1980) haben die beiden Hauptprobleme erkannt und eine Modifikation der Korrespondenzanalyse entwickelt. Mittels detrending (dt: entzerren) werden die hier beschriebenen Fehler korrigiert (wenn auch mathematisch nicht gerade elegant). Als Ergebnis ist die DCA entstanden, die im Moment die meist verwendete Ordinationsmethode bei der indirekten Gradientenanalyse sein dürfte. Für das detrending gibt es mehrere Herangehensweisen, die wichtigste ist aber das detrending by segments.

S.65

Korrespondenzanalyse

6.1 Das Prinzip

Die Korrespondenzanalyse (CA = correspondence analysis) wurde speziell für ökologische Fragestellungen von Hill (1973) eingeführt, der sie damals als reciprocal averaging (= RA) bezeichnete. Es handelt sich hier um eine Erweiterung der Weighted-averaging-Technik, die eigentlich aus der univariaten direkten Gradientenanalyse stammt (s. Kap. 2.10). Mittlerweile ist die Korrespondenzanalyse bzw. ihre Weiterentwicklung DCA(= detrended correspondence analysis) eines der am weitesten verbreiteten Ordinationsverfahren der indirekten Gradientenanalyse in der Ökologie. Die Korrespondenzanalyse geht von der Annahme aus, dass sich eine große Zahl der Arten unimodal und nicht linear entlang der wichtigsten Umweltgradienten verhalten, d. h. die Arten besitzen ein Optimum irgendwo entlang dieser Gradienten. Abb. 6.1 zeigt die Vorkommenswahrscheinlichkeit einer Auswahl von Pflanzenarten entlang eines Bodenfeuchtegradienten, hier als mittlerer Grundwasserstand MWS dargestellt. Die meisten Artreaktionen zeigen eine Glockenkurve; Art 4 bevorzugt demnach feuchtere Bereiche als z. B. Art 7. Nur an den Enden des Gradienten, im sehr trockenen und sehr nassen Bereich, verhalten sich die Arten linear wie Art 5 und Art 2. Die Korrespondenzanalyse trägt genau dieser unimodalen Artantwort Rechnung. Bevor wir auf das Prinzip dieses Verfahrens zu sprechen kommen, möchten wir die grundlegenden Schritte zunächst in Bezug zu einem Umweltgradienten erläutern. Wenn sich eine Art unimodal entlang eines Gradienten verhält, tritt die Art am häufigsten in der Nähe ihres Optimums auf. Eine sinnvolle Schätzung dieses Optimums ist daher gegeben, wenn der Durchschnitt der Umweltvariablenwerte über alle Objekte, welche die Art enthalten, unter Einbeziehung der Häufigkeit der Art berechnet wird. Diese Schätzung ist das gewichtete Mittel GM, das schon in Kap. 2.10 beschrieben wurde (Gl. 2.22, Tabelle 2.14). Das gewichtete Mittel beschreibt dabei die Lage des Optimums der Art im Umweltgradienten genauer als der einfache, ungewichtete Mittelwert.

S.81

NMDS

Unter anderem hiermit hängt auch der Vorwurf zusammen, dass CA und DCA bestimmte Muster in den Daten nur unzureichend wiedergeben. Minchin (1987) stellte beim Vergleich von DCA und NMDS (Kap. 12.3) fest, dass letztere Methode für eine Reihe von Datensätzen besser interpretierbare Ergebnisse erzielt. So können unter bestimmten Bedingungen die Aufnahmen an einem Ende des Gradienten bei der DCA zu einer „Zunge“ gestaucht sein. Ausführliche Simulationen zeigen, dass Korrespondenzanalysen überhaupt nur dann sichere Ergebnisse liefern, wenn es einen klaren

Hauptgradienten gibt und weitere Gradienten demgegenüber relativ kurz sind (van Groenewoud 1992). Das hat auch damit zu tun, dass zu einer vernünftigen Abschätzung der gewichteten Mittelwerte (weighted averaging) schlicht eine gewisse Gradientenlänge nötig ist. Es gibt allerdings nicht selten diesen einen Hauptgradient, selbst wenn dieser häufig einen Komplex parallel variierender Faktoren darstellt (z. B. Höhengradient). Das sei nur ein Einblick, Genaueres zu den Unzulänglichkeiten ist z. B. bei McCune et al. (2002) und Legendre u. Legendre (1998) nachzulesen.

Eine rangbasierte multivariate Analyse muss versuchen, die Beziehungen zwischen Objekten hinsichtlich vieler Variablen abzubilden. Als Ausgangspunkt bietet sich die Berechnung einer Unähnlichkeits- oder Distanzmatrix an, aus der dann leicht eine Rangfolge der paarweisen Abstände

zwischen den Objekten abgeleitet werden kann. Die Nichtmetrische Multidimensionale Skalierung (NMDS) bildet diese Rangfolge in einem wenigdimensionalen Raum ab. Wie bei den metrischen multidimensionalen Skalierungen (PCoA) auch, wird hier eine Projektion gesucht, die die Beziehungen in dem ursprünglichen Datensatz möglichst unverzerrt wieder-gibt. Allerdings wird bei der NMDS nur auf Monotonie in den Beziehungen geachtet, nicht auf die absolute Größe der Distanzen. Sie benutzt dafür also eben nicht direkt die Werte aus der Dreiecksmatrix, sondern nur die

relative Reihenfolge der Objekte bzw. Aufnahmen in Bezug auf den gewählten Unähnlichkeits- oder Distanzkoeffizienten. Die NMDS ist also ein rangbasiertes Verfahren. Damit hat sie die gleiche Flexibilität wie die PCoA, ist aber robuster. Wegen des einfacheren Modells gilt die NMDS sogar als relativ unempfindlich gegen fehlende Werte in der Matrix und 12.3 NichtmetrischeMultidimensionale Skalierung 143 kann daher bei lückigen Datensätzen hilfreich sein (Legendre u. Legendre 1998). Theoretisch sollte die NMDS also die ideale generelle Ordinationsmethode sein und wird entsprechend auch von einigen Autoren als die vielversprechendste Alternative zu den klassischen Verfahren propagiert (Brehm u. Fiedler 2004; McCune et al. 2002; Minchin 1987). So einfach die Idee ist, so schwierig ist aber ihre Umsetzung. NMDS ist die aufwändigste der hier vorgestellten indirekten Ordinationsmethoden, und wir beschränken uns nur auf die Erklärung des Prinzips. In der Regel wird auch in der NMDS ein iterativer Algorithmus genutzt (Kruskal 1964). Die ersten Schritte ähneln der PCoA; wir können die Daten evtl. transformieren und müssen dann ein Unähnlichkeits- oder Distanzmaß wählen. Wegen ihrer für ökologische Fragestellungen günstigen Eigenschaften wird oft die Bray-Curtis-Unähnlichkeit benutzt (Faith et al. 1987). Dass sie semimetrische Eigenschaften hat, stört hier nicht, denn bei den weiteren Schritten wird ja nicht mit den direkten Distanzen, sondern nur mit deren Rängen gerechnet. Ähnlich wie bei Spearman-Rangkorrelationen treten zwar auch hier häufig Bindungen auf (Kap. 2.3), denn die Bray-Curtis-Unähnlichkeit liefert ja Werte von 1, wenn die Objekte keine Arten gemeinsam haben. NMDS kann damit aber umgehen (Details: Legendre u. Legendre 1998). Gelegentlich wird stattdessen auch die Euklidische Distanz empfohlen (Backhaus et al. 2003; Kenkel u. Orlocci 1986), diese sollte aber, wie in Kap. 4.3 gezeigt, bei ökologischen Fragestellungen nur mit Vorsicht angewandt werden. Das Prinzip der Projektion ist nun, die Objekte (i. d. R.) zunächst zufällig in einem z. B. 3dimensionalen Ordinationsraum zu verteilen und dann zu vergleichen, wie gut oder schlecht diese neue Anordnung mit der Rangstruktur in der ursprünglichen Dreiecksmatrix übereinstimmt. Dazu be-rechnen wir die Distanzen der Objekte im neuen (z. B. 3dimensionalen) Ordinationsraum. Da dieser euklidisch ist, nutzen wir dafür den Satz des Pythagoras, also die Euklidische Distanz. Wenn wir jetzt die Objektpaare entsprechend der Reihenfolge ihrer Abstände in der ursprünglichen Dreiecksmatrix auftragen und in die nächste Spalte die entsprechenden Distanzen im Ordinationsraum schreiben, können wir die Güte der Ordination beurteilen. Im Idealfall sollten paarweise Abstände, die in der ursprünglichen Matrix relativ klein waren (also einen niedrigen Rang hatten), auch im neuen Ordinationsraum einen kleinen Wert haben. Anders ausgedrückt, die Werte sollten monoton steigen. Diese Idee verdeutlicht ein Shepard-Diagramm. Dabei werden die ursprünglichen Abstände zwischen den Objekten (hier auf Basis der Bray-Curtis-Dreiecksmatrix) gegen die neuen Distanzen im Ordinationsraum aufgetragen (Abb. 12.4).