Multivariate Datenanalyse mittels Biplots - Analyse am Beispiel der *benthos* Daten

Gerhard Welzl

Inhaltsverzeichnis

1 Multivariate Analyse mittels Biplots

Die Ausgangsdaten für eine multivariate Analyse lassen sich meist in Form einer rechteckigen Anordnung (*Tafel-Daten*) darstellen. Eine derartige Datenmatrix enthält Informationen über zwei unterschiedliche Entitäten: Die Reihen dieser Matrix korrespondieren mit den Stichproben- oder Beobachtungs- Einheiten, die Spalten mit den Variablen, deren Messwerte für diese Einheiten erhoben wurden. Ziel der multivariaten Analyse ist es, Strukturen in den Beobachtungseinheiten zu erkennen, Zusammenhänge zwischen den Variablen zu beschreiben sowie eine gemeinsame Interpretation von Variablen und Beobachtungen zu ermöglichen. Schon bei kleineren Tafeln, insbesonders aber bei einer größeren Variablenzahl ist es kaum möglich, direkt auf der Basis der Zahlen Strukturen zu erkennen. *The humain brain has not evolved to interpret tables of numbers in this way. We are much better in absorbing pictorial information.*

Eine Visualisierung derartiger Tafel-Daten kann durch die Konstruktion sog. **Biplots** erreicht werden. Biplots stellen eine Verallgemeinerung von Scatterplots für zwei Variable auf mehrere Variable dar. Eine exakte Definition eines Biplots beruht algebraisch auf der Zerlegung einer Zielmatrix (Ausgangsdatenmatrix) in eine linke und eine rechte Matrix mit jeweils zwei Spalten. Im zweidimensionalen Biplot werden dann diese beiden Matrizen gemeinsam als *scores* (bezogen auf die Reihen der Matrix) bzw. als *loadings* (bezogen auf die Spalten=Variable) dargestellt. Für das Verständnis ist eine geometrische Interpretation entscheidend: Durch Projektion auf die Achsen des Biplots sind die ursprünglichen Daten reproduzierbar. Tabelle 1 zeigt ein sehr einfaches Beispiel (Datenmatrix mit fünf Reihen und vier Spalten, Seite 19 in [1]).

	y1	y2	y3	y4
x1	8	2	2	-6
x2	5	0	3	-4
x3	-2	-3	3	1
x4	2	3	-3	-1
x5	4	6	-6	-2

Tabelle 1: Zielmatrix

Diese Daten können als Biplot dargestellt werden (Abbildung 1). Die Reihen der Ausgangsmatrix sind als Punkte (x1, x2, x3, x4, x5) , die Spalten als Achsen dargestellt. Durch Projektion eines Punktes (z.B. x1) auf eine Achse (z.B. y1) ergibt sich der Wert in der Daten-Tafel (z.B. 8). Einschränkend muss erwähnt werden, dass diese exakte Reproduktion der Rohdaten nur für sehr einfache Beispiele möglich ist. Bei realen Daten wird es nötig sein, die exakten Werte durch in gewissen Sinn optimale Approximationen zu ersetzen. Die Prinzipien der Biplot-Konstruktion bleiben aber erhalten.

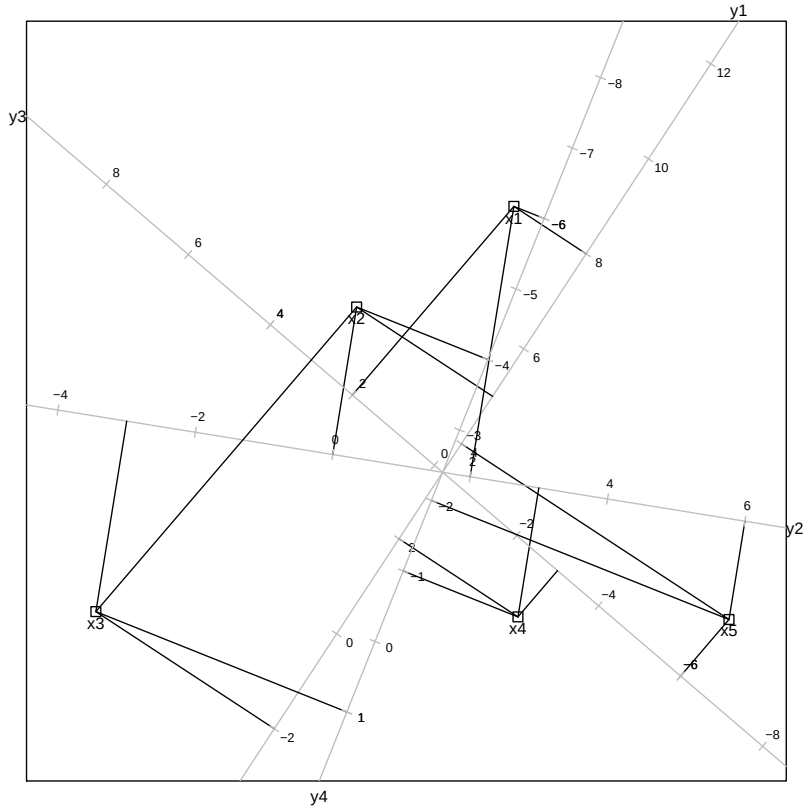

Abbildung 1: Biplot

Im folgenden sollen Methoden beschrieben werden, bei denen keine zusätzlichen Informationen zu Reihen oder Spalten der Datenmatrix bei der Konstruktion von Biplots berücksichtigt werden (*unsupervised methods*). Außerdem erfolgt eine Beschränkung auf diskrete Merkmale, hier insbesondere sog. Zähldaten, wie sie z.B. bei der Analyse von Abundanzen in ökologischen Studien oder bei der Auswertung von Daten aus Sequenzanalysen anfallen. Als geeignete Methoden zur Auswertung derartiger Daten werden die Korrespondenzanalyse (z.B. die R-packages **ca** (siehe auch [1]) und **cabipl** (siehe [2]) oder Verfahren der Multidimensionale Skalierung (z.B. integriert in das *mothur* Projekt) genannt. Das Problem bei dieser methodenorientierten Vorgehensweise ist, dass es entweder zu einer blinden Anwendung vorhandener Software kommt oder dass es bei der exzessiven Anzahl von Varianten - darunter etliche Methoden, die so gut wie das Gleiche machen - dem Benutzer sehr erschwert wird eine geeignete Auswahl zu treffen.

Die am meisten benutzte Variante der Korrespondenzanalyse dürfte die *Canonical correlation approximation* sein; ein wesentlicher Grund dafür ist sicherlich, dass diese Variante die Grundeinstellung für Biplots im R-package **ca** ist (`map = "(symmetric)"`. Das gleiche Biplot kann mit `cabipl` unter Verwendung

der Variante `ca.variant = "Corr"` erzeugt werden. Eine Bewertung dieser Methode findet sich in [2] (Seiten 295 und 301): *This gives approximations to the row and column chi-sqared distance, but the relationships between the two sets of points seem to have no simple interpretation. Furthermore, their inner product seem to be of little interest. Nevertheless, this seem to be the most commonly occuring form of correspondence analysis... So care has to be taken not to use inner product interpretation for such diagrams..* Die Grundeinstellung bei `cabipl` basiert auf *Approximation to Pearsons chi-squared*. Diese Variante steht im package **ca** überhaupt nicht zur Verfügung.

Diese Beobachtungen scheinen erneut gegen eine - oft propagierte - *automatische* Anwendung von Statistikprogrammen zu sprechen. Die Probleme bei der Auswahl geeigneter Korrepondenzanalysemethoden sollten von einer Methodenorientierung wegführen hin zu einer problemorientierten Vorgehensweise. Von entscheidender Bedeutung bei dieser Auswahl ist dabei das Konzept der **Distanzen** zwischen den Stichproben (Reihen der Datenmatrix). Für stetige bzw. diskrete Variable sind jeweils unterschiedliche Distanzmatrizen in Betracht zu ziehen. Während stetige Variable häufig auf der Basis der standardisierten Euklidschen Distanz analysiert werden, ist dieses Abstandsmaß für Zähldaten weniger geeignet, da hiermit kleine Abstände zwischen Stichproben mit einer Vielzahl von gemeinsamen Nullen erzeugt werden.

Bei der Analyse von Zähldaten ist außerdem zu berücksichtigen, ob das Profil der Gesamtabundanzen bezüglich der Stichproben (Summe über die Reihen der Datenmatrix) einen Einfluss auf das Distanzmaß haben soll oder nicht. Im ersten Fall ist die Verwendung von absoluten Abzundanzen, im zweiten Fall die Verwendung von relativen Abundanzen naheliegend. Bei klassischen ökologischen Studien wird häufig eine feste Zahl von Proben pro Stichprobeneinheit untersucht, wodurch Unterschiede zwischen den beiden Vorgehensweisen zu vernachlässigen sind.

Für die Analyse der Struktur von Artengemeinschaften mittels Abundanzdaten wurden von Ökologen mehrere Distanzmaße vorgeschlagen:

basierend auf relativen Abundanzen

- Chi-square distance
- Hellinger distance
- Bhattacharya distance
- Bray-Curtis dissimilarity
- Yue-Clayton dissimilarity

basierend auf absoluten Abundanzen

- Pearson residuals transformation
- Contingency ratio
- Bray-Curtis dissimilarity

Es wäre vorteilhaft, wenn es einen allgemeinen Zugang zur Analyse von Tafel-Daten auf der Basis all dieser Distanzmaße gäbe, der auch eine einheitliche Software umfassen würde. In [3] wird ein derartiges vereinheitlichtes Konzept zur Visualisierung vorgeschlagen. Allerdings beschränkt sich diese Verallgemeinerung auf Transformationen des *Pearson contingency coefficient*. Damit können nur einige der Distanzen aus obiger Liste reproduziert werden.

Ein weiteres allgemeines Konzept beruht auf der Konstruktion einer *Weighted Euclidean Distance* [4]. Hierbei werden für beliebige Distanzmaße Gewichte für die Variablen derart geschätzt, dass die vorgegebene Distanzmatrix möglichst gut approximiert wird. Dies geschieht unter Verwendung des *majorization algorithm*, wodurch auch ein Bezug zur multidimensionalen Skalierung erkennbar ist. Mit Hilfe dieser Gewichte können dann mittels einer gewichteten Hauptkomponentenanalyse (wPCA) *Weighted Euclidean Biplots* erzeugt werden, wobei **ein** Programm zur Konstruktion von Standard-Biplots für alle möglichen Distanzen genügt (z.B. eine leicht modifizierte Form von `PCAbipl` aus dem **UBbipl** package).

2 Daten

Die *benthos* Daten (Analyse in [1]) sind erhoben worden, um die Auswirkungen der Ölgewinnung auf das maritime Leben zu untersuchen. Die Datenmatrix besteht aus 13 Reihen, zugeordnet 13 Probestellen am Meeresboden der Nordsee in der Nähe eines Ölfeldes. An jeder Probestelle wurden die benthischen Organismen identifiziert, gezählt und die Ergebnisse zu einer ökologischen Abundanz-Matrix zusammengefasst (Anzahl der Spalten = Anzahl der Arten = 92). Zwei der Probestellen (R40 und R42) dienen als Kontrolle; sie sind weit enfernt vom Ölfeld und werden als unbelastet angesehen. Während die Gesamtabundanzen pro Probestelle zwischen 516 und 1331 variieren, ist die Zahl der Gesamtabundanzen pro Art sehr unterschiedlich (3 - 2732). Häufig werden vor der statistischen Analyse Organismen mit kleiner Gesamtabundanz (etwa ≤ 10) ausgeschlossen; die nachfogenden Auswertungen wurden jedoch mit den Gesamtdaten durchgeführt.

Abbildung 2 zeigt einen kleinen Teil dieser Abundanzmatrix in Form einer *heatmap*.

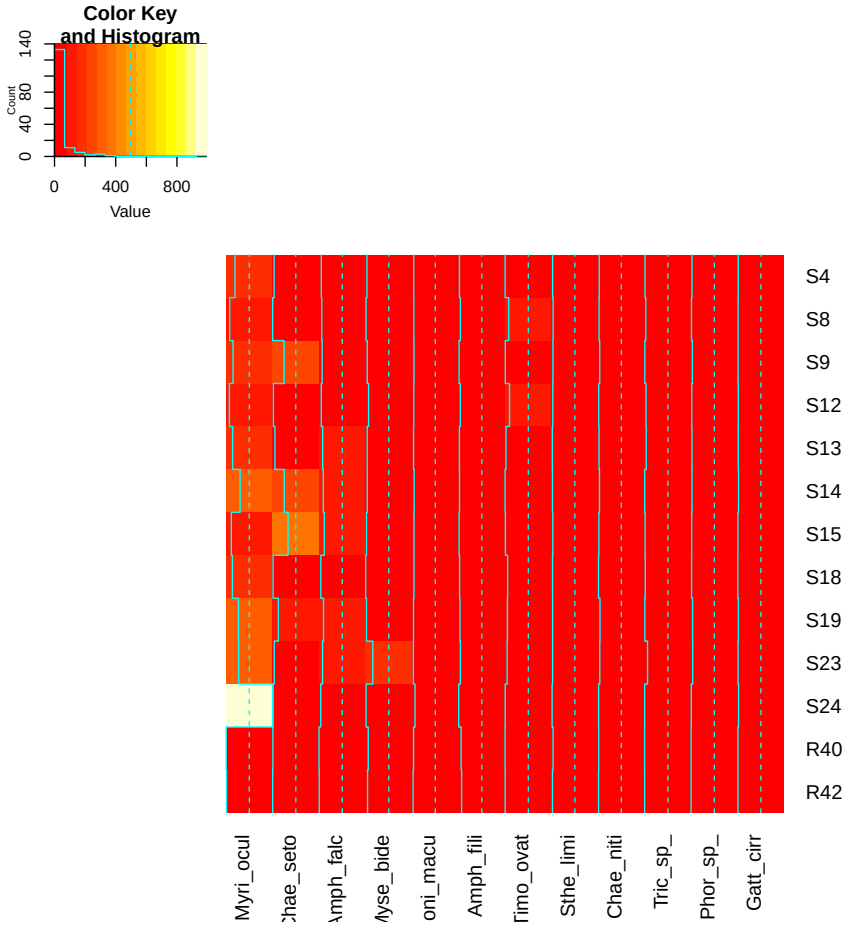

Abbildung 2: heatmap *benthos* Daten (Teil)

In Abbildung 2 ist ein Effekt zu erkennen, der öfter bei derartigen ökologischen Daten auftritt: eine Art tritt an einer Probestelle besonders häufig auf. *Myri_ocul* wurde 992 mal an der Probestelle S24 gezählt; die zweithöchste Abundanz liegt bei 302. Diese *Ausreisser*situation wird auch durch die Dichtefunktion in Abbildung 3 veranschaulicht.

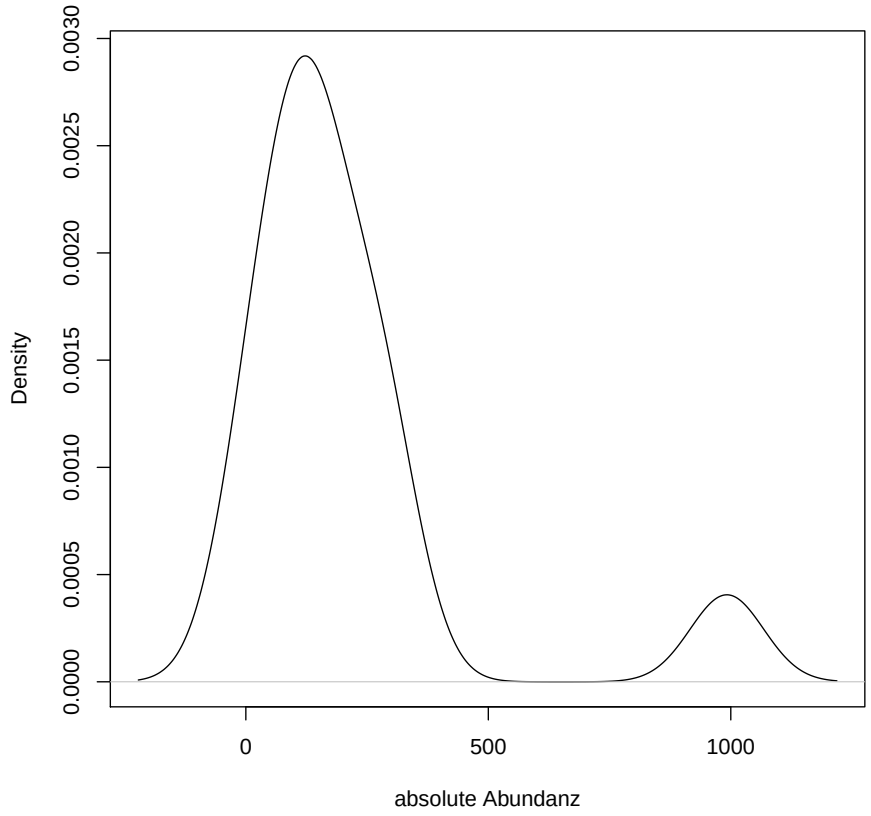

Abbildung 3: Dichtefunktion Myrio_ocul

3 Analyse der relativen Abundanzen

Tabelle 2 zeigt einen Ausschnitt der relativen Abundanzen für die *benthos* Daten.

	Myri_ocul	Chae_seto	Amph_falc	Myse_bide	Goni_macu	Amph_fili	Timo_ovat
S4	0.32	0.06	0.08	0.05	0.06	0.03	0.00
S8	0.14	0.01	0.10	0.02	0.07	0.07	0.14
S9	0.18	0.30	0.08	0.04	0.05	0.01	0.01
S12	0.11	0.03	0.07	0.10	0.06	0.06	0.14
S13	0.22	0.08	0.12	0.05	0.05	0.03	0.03
S14	0.29	0.24	0.09	0.04	0.04	0.02	0.00
S15	0.13	0.38	0.13	0.02	0.05	0.01	0.01
S18	0.26	0.02	0.07	0.01	0.08	0.04	0.11
S19	0.30	0.14	0.11	0.02	0.03	0.03	0.06
S23	0.28	0.04	0.08	0.16	0.03	0.04	0.04
S24	0.75	0.01	0.03	0.01	0.05	0.00	0.00
R40	0.01	0.02	0.00	0.16	0.09	0.15	0.00
R42	0.04	0.01	0.02	0.14	0.07	0.21	0.01

Tabelle 2: Relative Abundanzen *benthos* Daten (Teil)

3.1 Chi-square distance

Ein Biplot, das auf einer Approximation der Chi-Quadrat Distanzen beruht, kann sowohl mit dem package **UBbipl** (Programm `cabipl, ca.variant = "Chisq2Rows"`), als auch mit dem package **ca** (`map = "rowgreen"`) erzeugt werden (siehe Abbildung 4). Dieses Ergebnis ist der Abbildung **Exhibit 8.4** (Seite 87 in [1]) sehr ähnlich (hier wurden absolute Abundanzen benutzt). Die zugehörige Interpretation gilt auch für Abbildung 4: *Rare species are pulled towards the centre. The species vectors now show which ones are important to interpret, because their length now reflect the contributions of the species to the solution. The figure shows which are the important species that separate out the unpolluted sites R40 and R42 to the right, while species Chaetesona setosa is generally found at pollutes sites, particulary S15 which is close to the oilfield. There is a very high abundance of Myriochele oculata at site S24 which is not related to the polluton gradient.*

Die *Herunterstufung* seltener Arten kann direkt aus Abbildung 4 abgelesen werden, da die Größe der Dreiecke in der Abbildung proportional der Gesamtabundanz der einzelnen Arten ist und die kleineren Dreiecke meist näher zum Nullpunkt liegen.

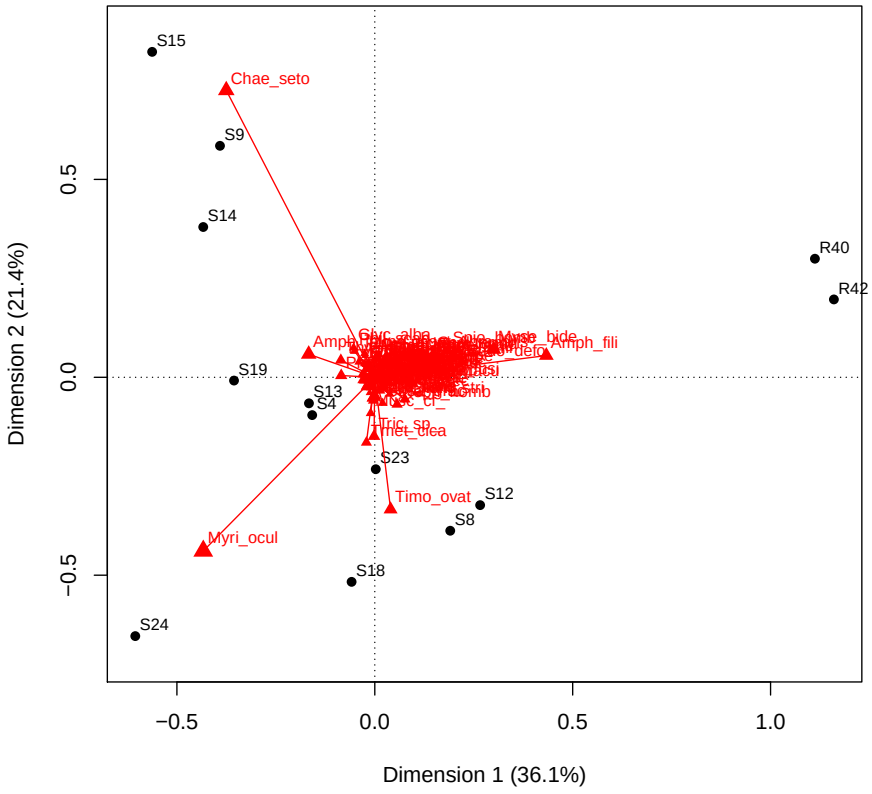

Abbildung 4: Biplot (CA): Chi-square distance

Dieser Effekt kann durch die separate Darstellung der Ladungen verdeutlicht werden, wobei in Abbildung 5 die Ladungen für seltene Arten (Gesamtabundanz ≤ 10) rot markiert sind.

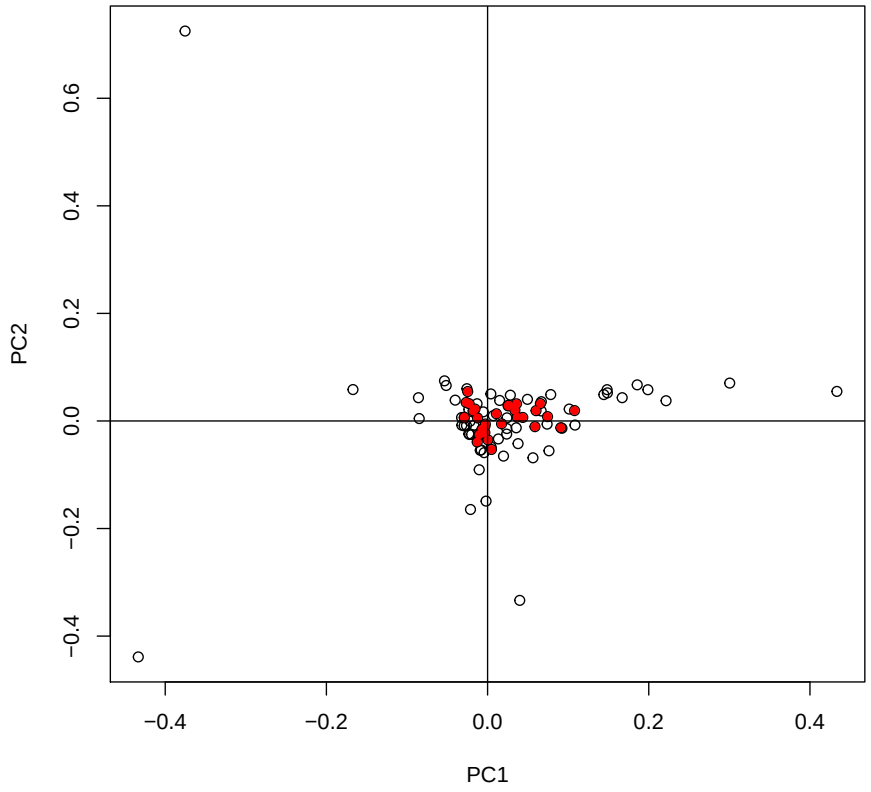

Abbildung 5: Biplot Ladungen: Chi-square distance (seltene Arten rot)

Es kann gezeigt werden, dass das Biplot zur Korrespondenzanalyse (Approximation der Chi-Quadrat Distanzen) auch mittels einer modifizierten Form der Hauptkomponentenanalyse (PCA) erzeugt werden kann. So wie bei stetigen Variablen (*Standardized Euclidean distance*) eine Gewichtung durch die Wurzel aus der inversen Standardabeichung (Skalierung) erfolgt, können andere Gewichte benutzt werden um andere Distanzmaße abzubilden. Wenn mit den relativen Abundanzen eine gewichtete Hauptkomponentenanlyse unter Verwendung der Wurzel der inversen mittleren relativen Abundanzen der Arten als Gewichte durchgeführt wird, so ist das Standard-Biplot der PCA (Abbildung 6) inhaltlich identisch mit dem Biplot der Korrespondenzanalyse (Abbildung 4)

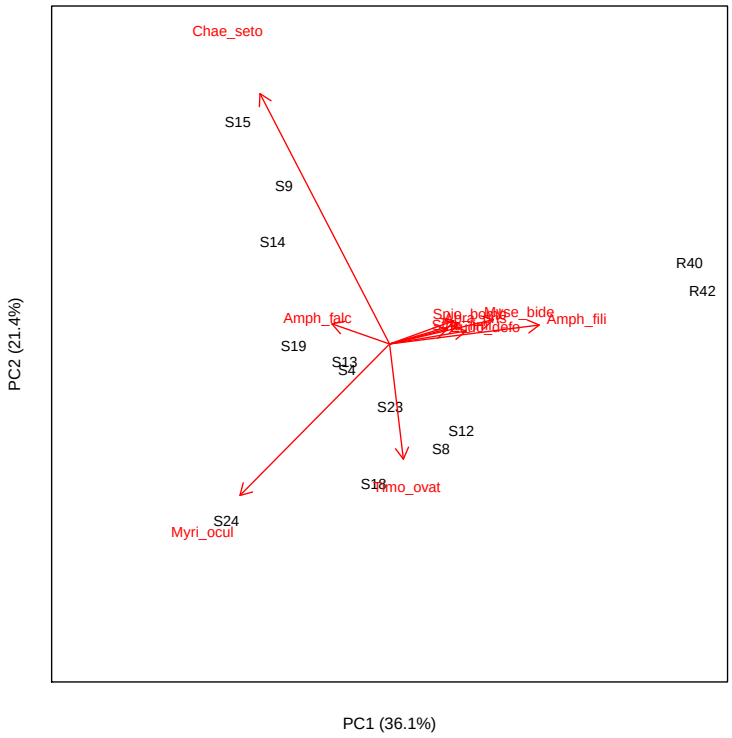

Abbildung 6: Biplot (wPCA) : Chi-square distance / 10 ausgewählte Variable

Der Vorteil der Verwendung von Biplot auf der Basis einer Hauptkomponentenanalyse liegt auch darin, dass alle bekannten Eigenschaften und Maße der PCA benutzt werden können, wie z.B die Maße für die generelle Güte der Anpassung eines Biplots: *quality* und für die Güte der Approximation der Variablen: *adequacy* (siehe Seite 87 in [2]). Während *quality* definiert ist als die Summe der erklärten Varianzen der beiden Biplot-Komponenten ist *adequacy* gleich der quadrierten Länge der Biplot Vektoren. Wenn die Anzahl der Variablen groß ist, wird *adequacy* häufig benutzt, um Variable mit der besten Approximation im Biplot auszuwählen. Abbildung 6 zeigt nur die 10 Biplot Vektoren mit den besten *adequacies*. Die *quality* ist $0.361 + 0.214 = 0.575$

Ein weiterer Vorteil liegt darin, dass auch Biplots mit kalibrierten Achsen erzeugt werden können. Abbildung 7 ist prinzipiell identisch mit Abbildung 6, hat aber kalibrierte Achsen. Die Kalibrierung bezieht sich auf relative Abundanzen. Aus Gründen der Übersichtlichkeit ist eine Beschränkung auf einige wenige Achsen sinnvoll. Die approximierten Werte der Variablen können als Projektionen auf die kalibrierten Achsen abgelesen werden (in Abbildung 7 sind das beispielsweise die Werte für die Probestelle S15: 0.18 (*Myri_ocul*), 0.36 (*Chae_seto*) und 0.01 (*Amph_fili*)). Das Maß für die Güte der Approximation der einzelnen Va-

riablen (*adequacies*) ist bei dieser Darstellung den Variablennamen hinzugefügt. Für diese Darstellungsform sind jeweils die drei *besten* Variablen ausgewählt. Bei konkreten Datenanalysen sollten auch inhaltliche Überlegungen diese Auswahl mitbestimmen.

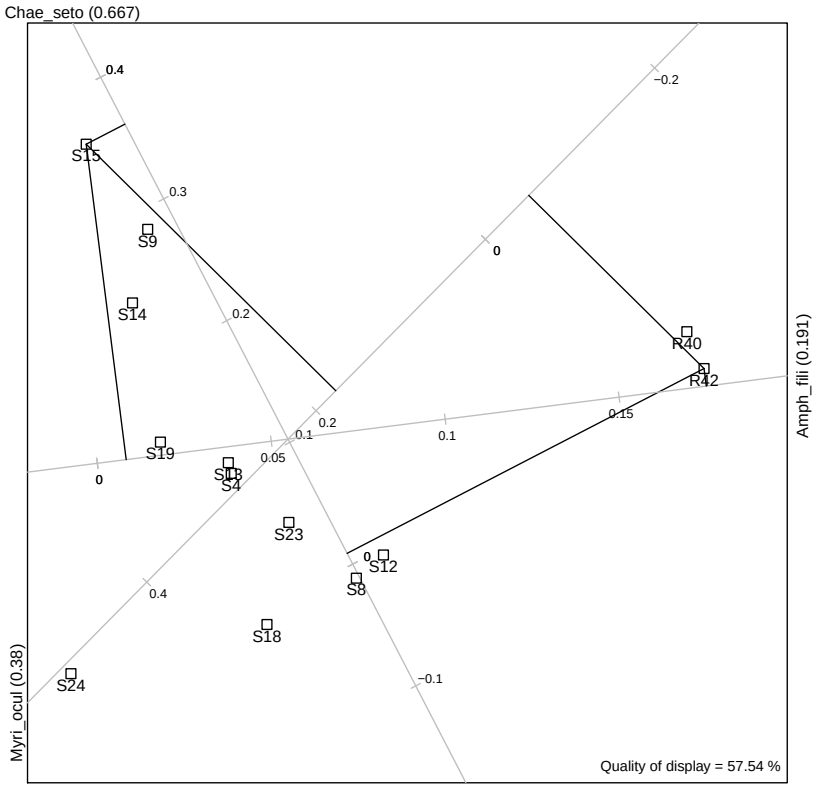

Abbildung 7: Biplot (wPCA) : Chi-square distance / 3 Achsen kalibriert, Probestellen S15 und R42 mit Projektionen

3.2 Hellinger distance

Ein Biplot, das auf einer Approximation der Hellinger Distanzen beruht, kann mit einer Haupkomponentenanalyse durchgeführt weren, nachdem die relativen Abundanzen einer Wurzeltransformation unterzogen wurden (Abbildung 8).

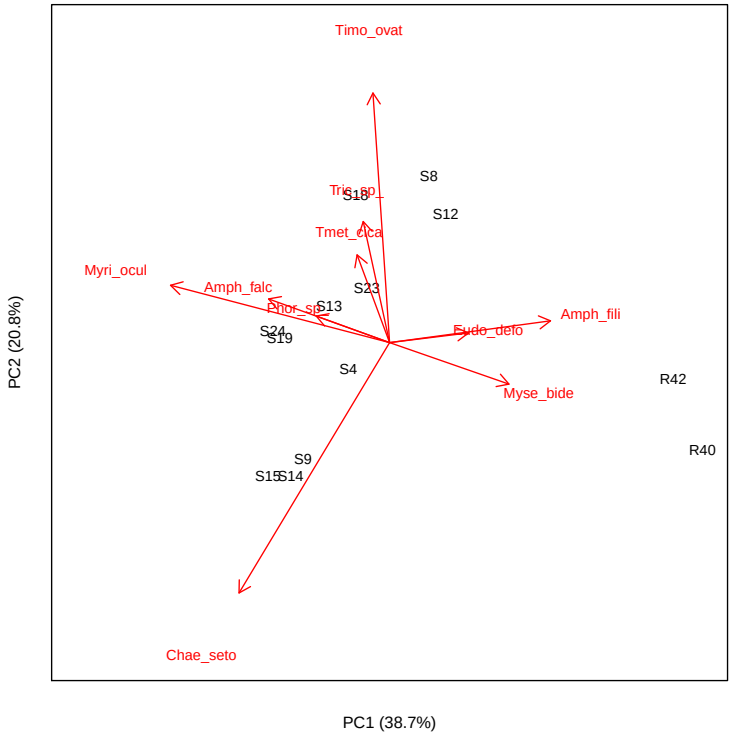

Abbildung 8: Biplot (PCA) : Hellinger distance / 10 ausgewählte Variable

Es ist möglich, die Kalibrierung der Achsen so zu gestalten, dass die relativen Abundanzen angezeigt werden. Dabei ist zu beachten, dass eine *nichtlineare* Skale entsteht. In Abbildung 9 sind die Werte für die Probestelle S15 0.29 (*Myri_ocul*), 0.26 (*Chae_seto*) und 0.01 (*Timo_ovat*) abzulesen. Im Vergleich zu Abbildung 7 besteht die größte Änderung in der Abwertung = Angleichung des Ausreissers für die Variable *Myri_ocul* an der Probestelle S24 (von 0.50 auf 0.32).

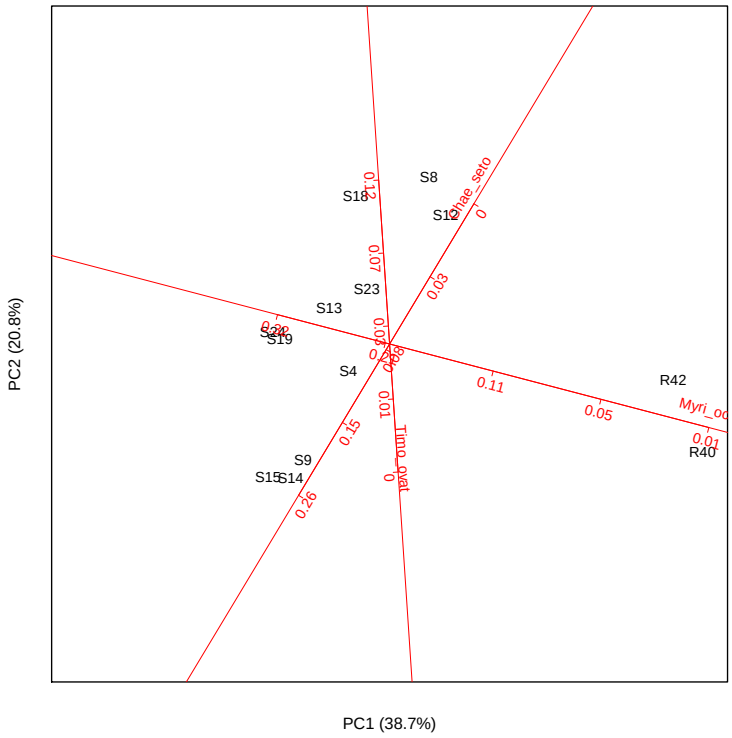

Abbildung 9: Biplot (PCA) : Hellinger distance / 3 Achsen kalibriert

Auch bei dieser Methode erhalten seltene Arten eher geringe Ladungen (siehe Abbildung 10)

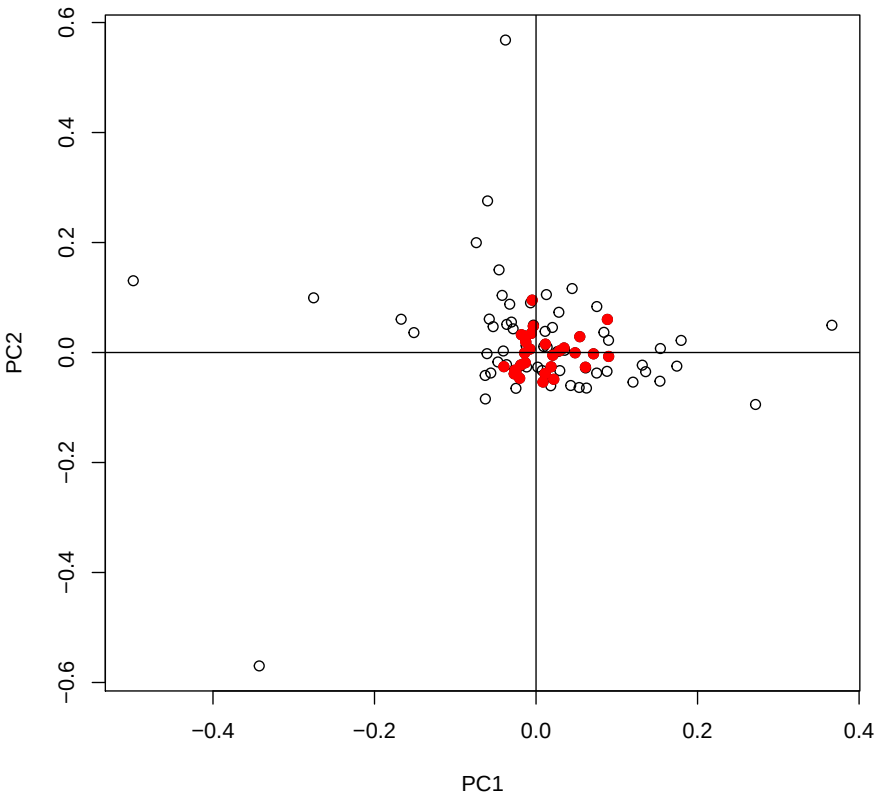

Abbildung 10: Biplot Ladungen: Hellinger distance (seltene Arten rot)

3.3 Bhattacharya (arccos) distance

Ein Biplot auf der Basis der Bhattacharya (arccos) distance kann nicht durch Varianten der Korrespondenzanalyse oder durch Transformationen der Abundanzen konstruiert werden. Üblicherweise wird in diesem Fall die Anwendung der *Multidimensionalen Skalierung* vorgeschlagen. Wie in Kapitel 1 erläutert wird, kann aber auch für jedes beliebige Distanzmaß ein *Weigthed Euclidean Biplot* erzeugt werden. Dazu ist es nötig, in einem ersten Schritt Gewichte zu schätzen, die eine möglichst gute Approximation an die vorgegebenen Distanzen erzeugen. Da dabei ein Optimierungsverfahren (*majorization algorithm*) verwendet wird, sollte die Güte dieser Approximation stets überprüft werden. Abbildung 11 zeigt die gute Übereinstimmung von Original- und approximierten Distanzen.

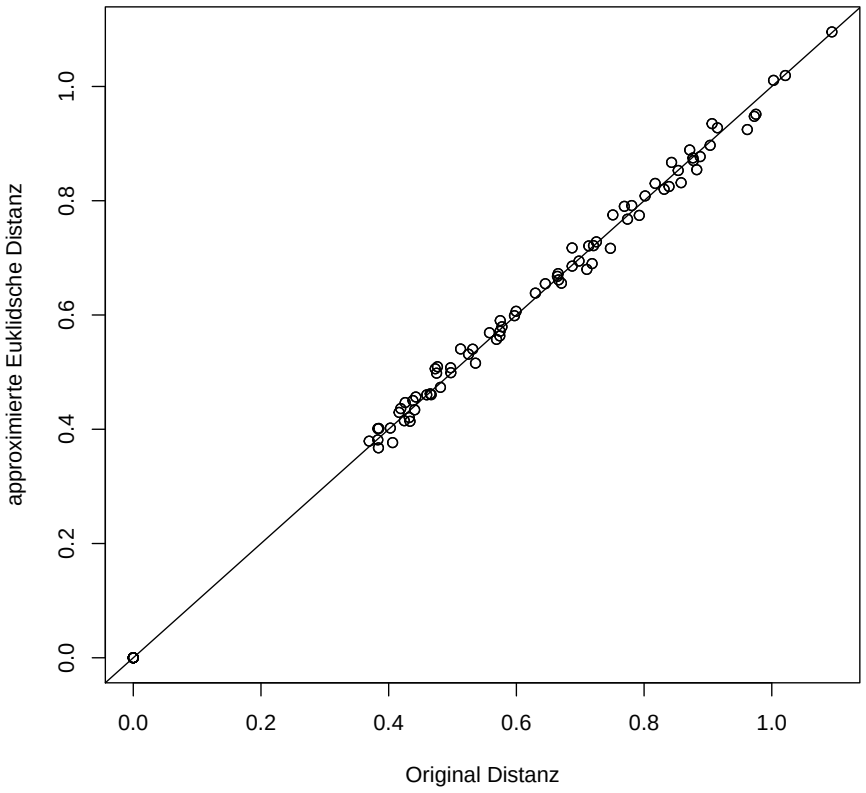

Abbildung 11: Vergleich von Original- und approximierter Distanz Matrix

In einem zweiten Schritt kann dann mit diesen Gewichten eine gewichtete Hauptkomponentenanlyse (wPCA) durchgeführt werden. Abbildung 12 zeigt das daraus resultierende Biplot, wobei nur die 10 *besten* Variablen dargestellt werden. Auch bei Verwendung dieser Variante wird offensichtlich der Ausreisser-Einfluss reduziert (Probestelle S24 liegt in der Nähe anderer Probestellen).

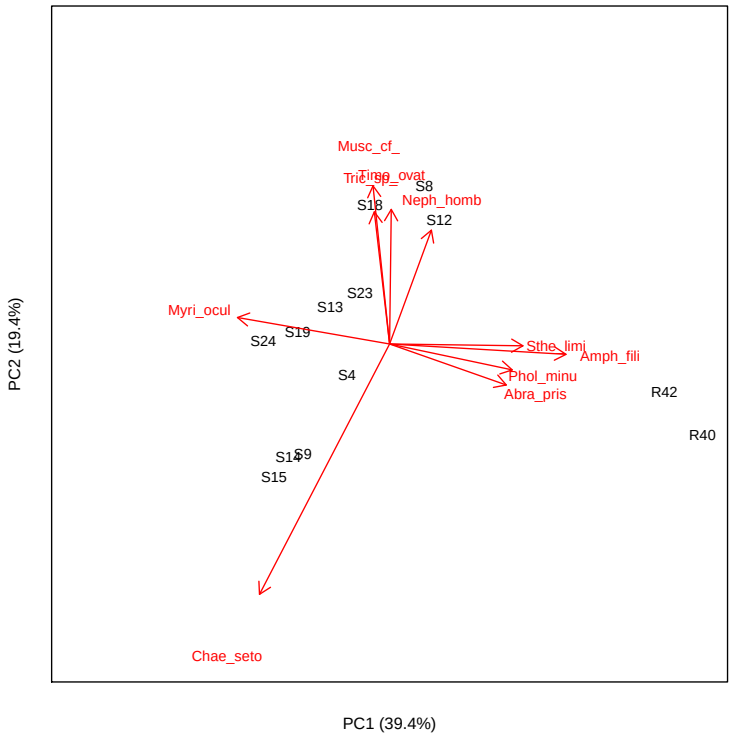

Abbildung 12: Biplot (wPCA) : Bhattacharya (arccos) distance / 10 ausgewählte Variable

Abbildung 13 zeigt drei Achsen mit Kalibrierung. Für die Probestelle S15 gilt: relative Abundanzen 0.27 (*Chae_seto*), 0.01 (*Amph_fili*) und 0 (*Musc_cf_*) (negative Werte nicht sinnvoll aber durch Approximation möglich). Durch die Darstellung kalibrierter Achsen sind Arten mit geringer relativer Abundanz direkt erkennbar.

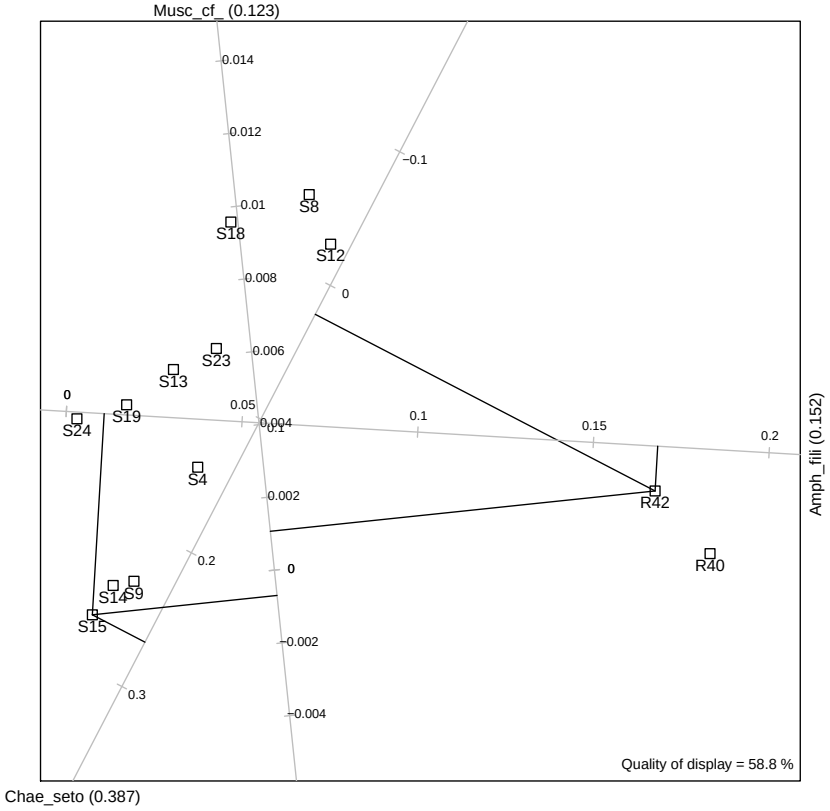

Abbildung 13: Biplot (wPCA) : Bhattacharya (arccos) distance / 3 Achsen kalibriert

Auch Abbildung 14 lässt erkennen, das bei dieser Methode seltene Arten auch höhere Gewichte erhalten können.

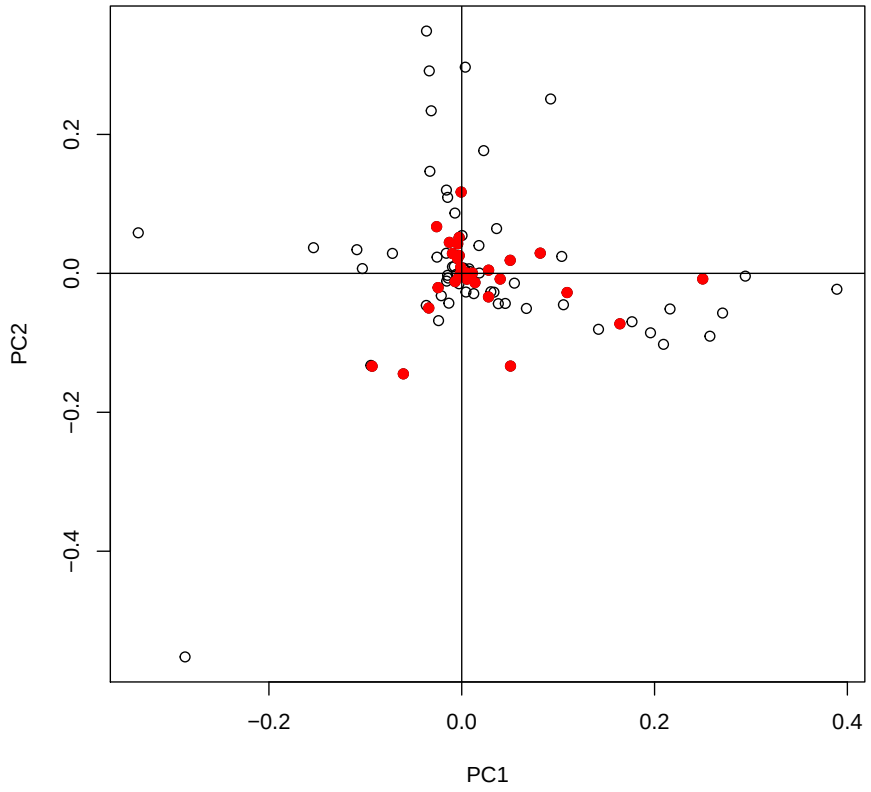

Abbildung 14: Biplot Ladungen: Bhattacharya (arccos) distance (seltene Arten rot)

3.4 Bray-Curtis dissimilarity

Zur Analyse der Daten von ökologischen Studien wird häufig zwischen den Distanz-Alternativen *Chi-square distance* und *Bray-Curtis dissimilarity* eine Entscheidung getroffen. Allerdings ist das vorrangige Anwendungsgebiet für Bray-Curtis basierte Analysen **absolute** Abundanzen, aber auch Anwendungen auf relative Abundanzen sind möglich (siehe [5]).

Zur Konstruktion eines Biplots ist wieder die Schätzung von optimalen Gewichtskoeffizienten erforderlich. Abbildung Fig. 15 zeigt die erreichte gute Anpassung.

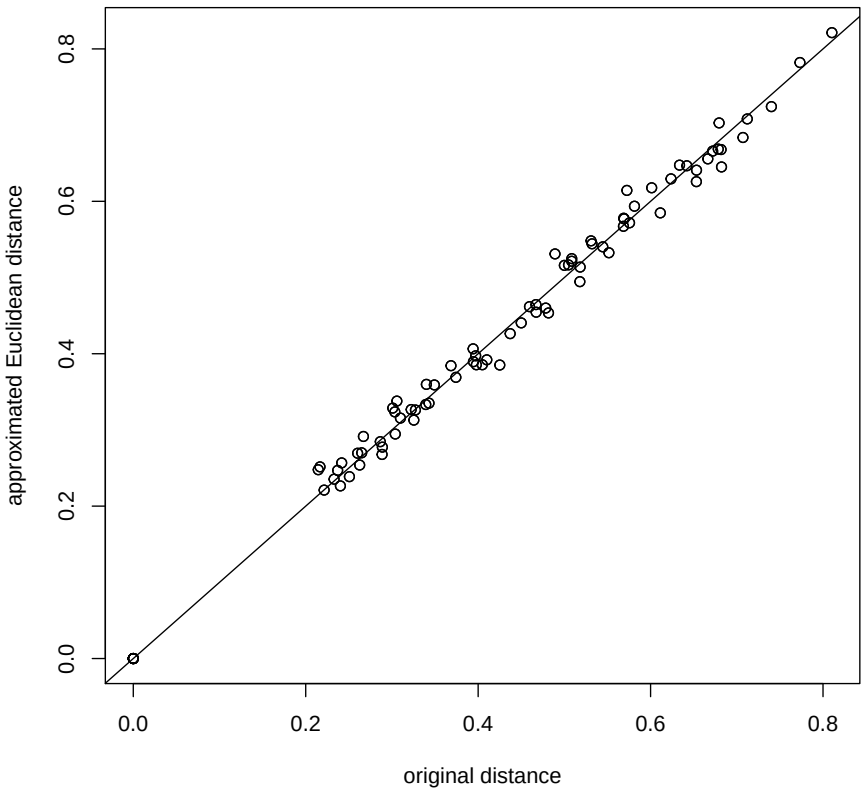

Abbildung 15: Vergleich von Original- und approximierter Distanz Matrix

Unter Verwendung dieser Gewichte kann eine gewichtete Hauptkomponentenanlyse (wPCA) durchgeführt werden. Abbildung 16 zeigt das zugehörige Biplot, wobei wieder nur die 10 *besten* Variablen dargestellt werden.

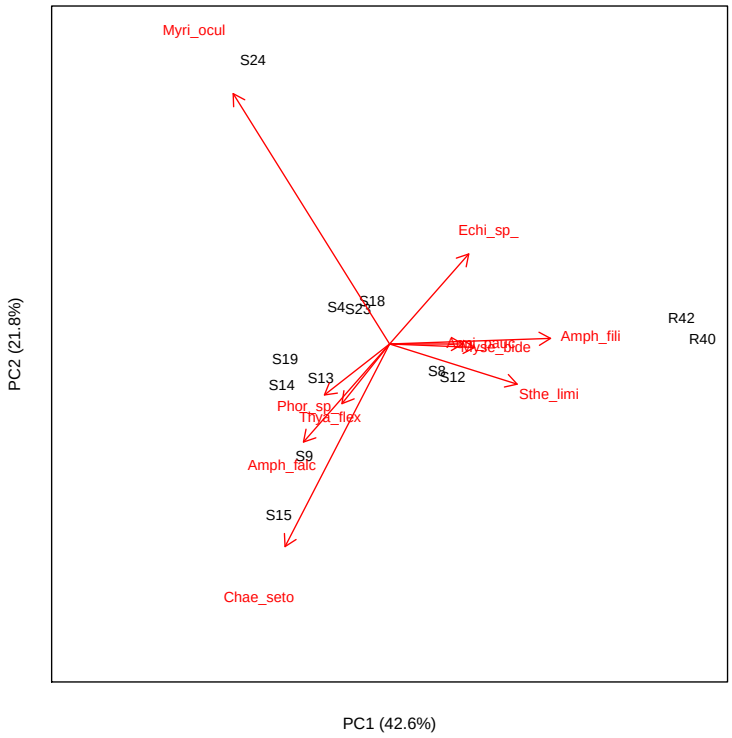

Abbildung 16: Biplot (wPCA): Bray-Curtis dissimilarity / 10 ausgewählte Variable

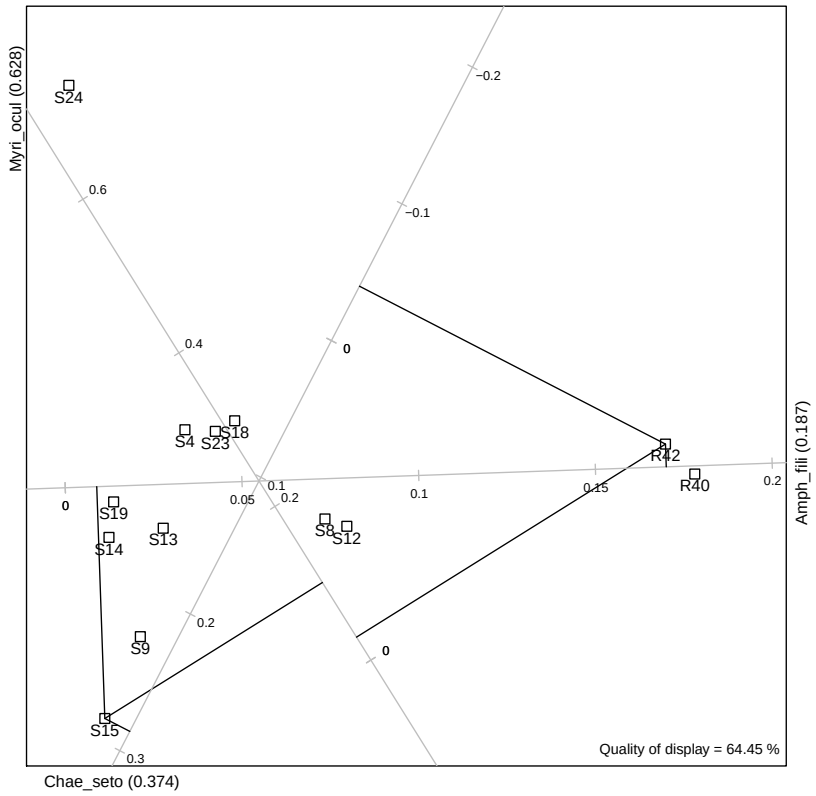

Abbildung 17: Biplot (wPCA): Bray-Curtis dissimilarity / 3 Achsen kalibriert

In Abbildung 17 sind die Werte für die Probestelle S15 0.10 (*Myri_ocul*), 0.29 (*Chae_seto*) und 0.01 (*Amph_fili*) abzulesen.

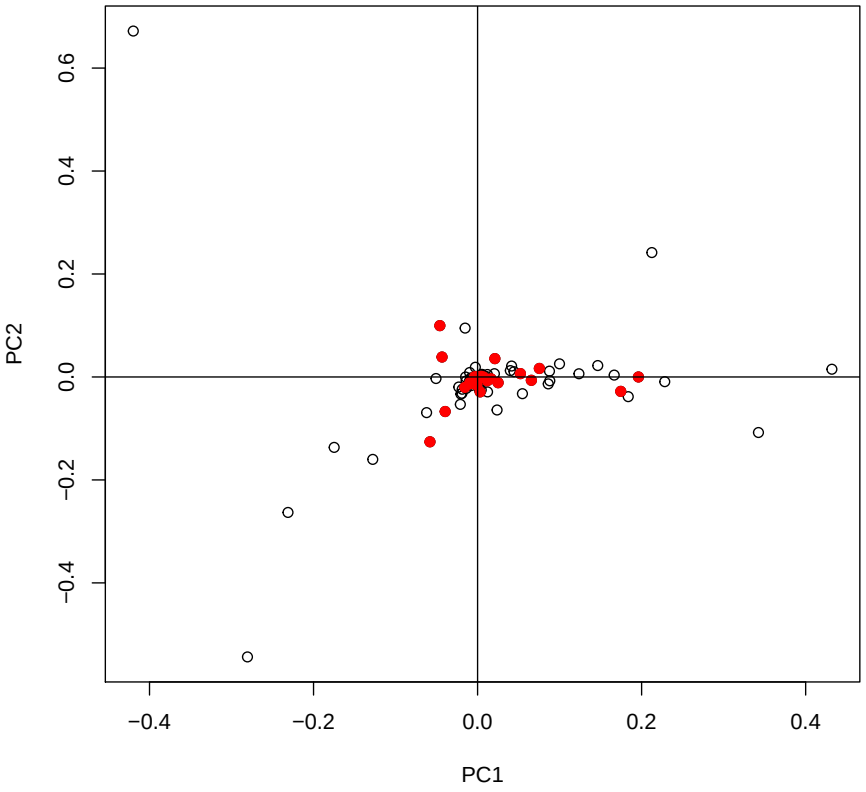

Abbildung 18: Biplot Ladungen: Bray-Curtis dissimilarity (seltene Arten rot)

3.5 Yue-Clayton dissimilarity

Distanzen (bzw. dissimilarities) nach Yue-Clayton werden vor allem bei der multidimensionalen Skalierung im *mothur* Projekt verwendet.

Zur Konstruktion eines Biplots ist wieder die Schätzung von optimalen Gewichtskoeffizienten erforderlich. Abbildung Fig. 19 zeigt die erreichte zufriedenstellende Anpassung.

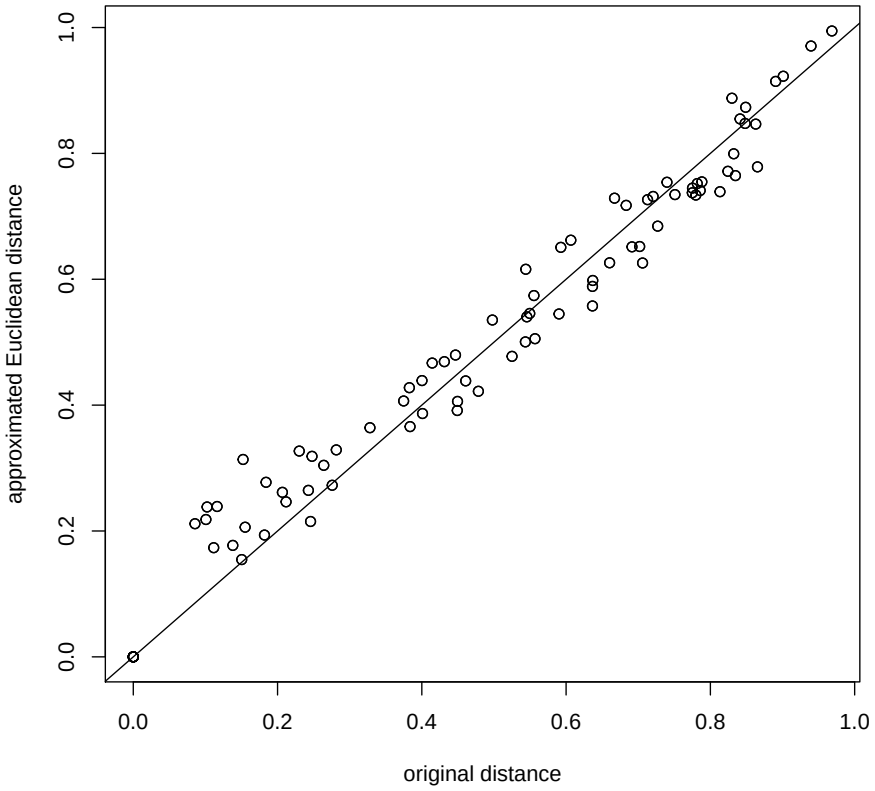

Abbildung 19: Vergleich von Original- und approximierter Distanz Matrix

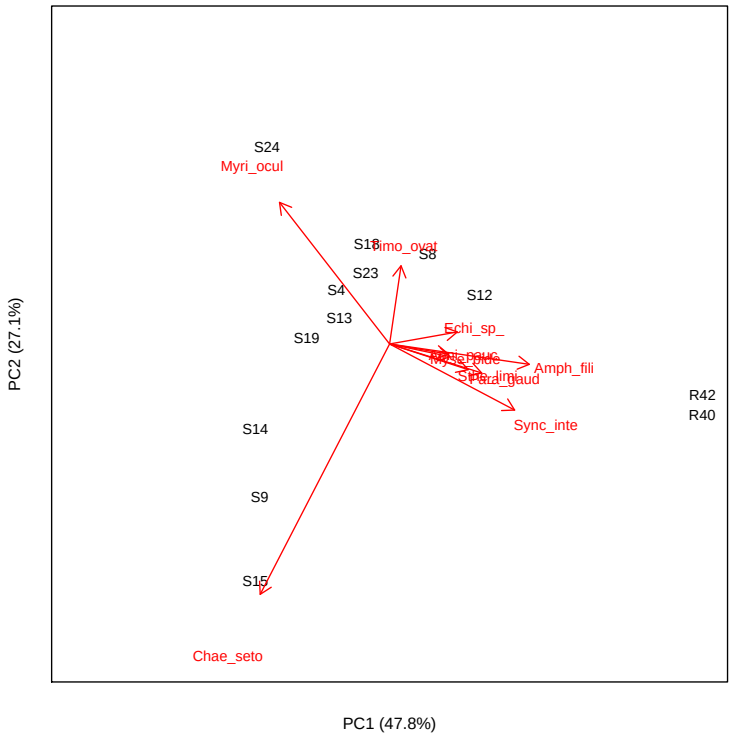

Abbildung 20: Biplot (wPCA): Yue-Clayton dissimilarity / 10 ausgewählte Variable

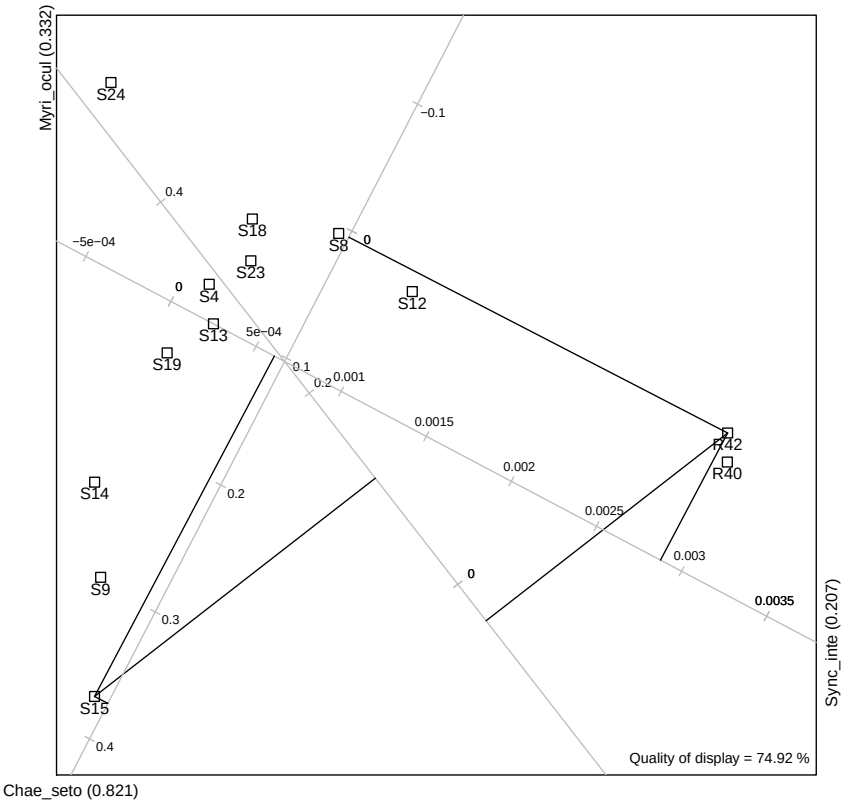

Abbildung 21: Biplot (wPCA): Yue-Clayton dissimilarity / 3 Achsen kalibriert

In Abbildung 21 ergeben sich für die Probestelle S15 geschätzte relative Abundanzen von 0.11 (*Myri_ocul*), 0.37 (*Chae_seto*) und 0.0006 (*Sync_inte*).

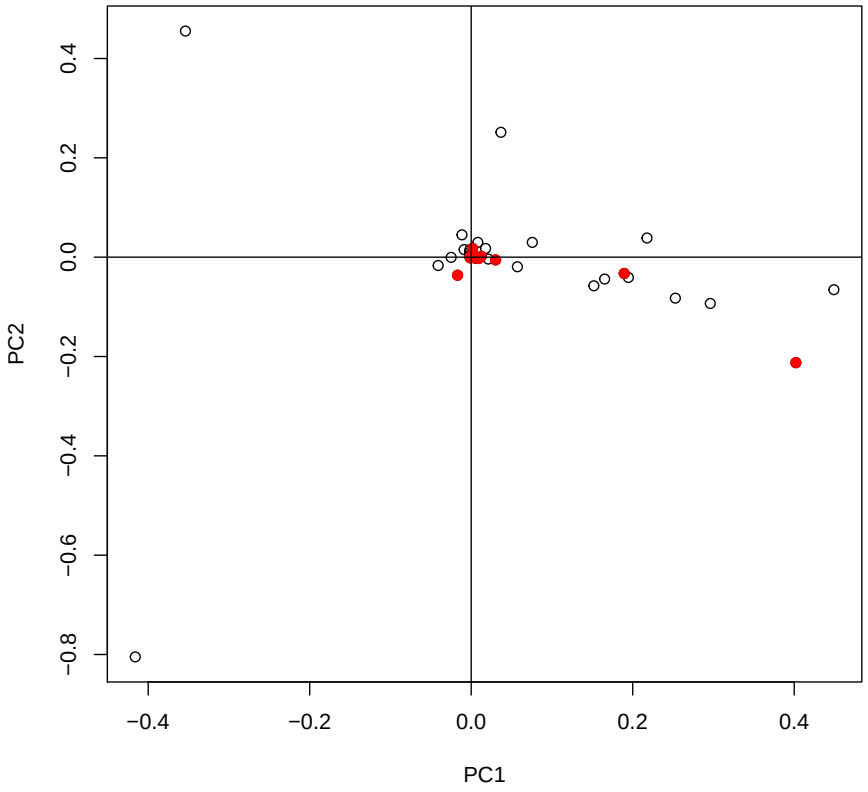

Abbildung 22: Biplot Ladungen: Yue-Clayton dissimilarity (seltene Arten rot)

4 Analyse der absoluten Abundanzen

Die vorliegenden *benthos* Daten legen eine Verwendung von **relativen** Abundanzen nahe. Die Analyse von **absoluten** Abundanzen ist insbesondere dann angezeigt, wenn die *rowprofiles*, d.h. die Abundanzverteilung bezüglich der Stichproben (Summe über die Reihen der Datenmatrix) einen Einfluss auf das Distanzmaß haben soll. Im folgenden werden dennoch Approximationen zu *Pearson residuals*, zu *Contingency ratio* und zu *Bray-Curtis dissimilarity* betrachtet.

4.1 Pearson residuals

Die Verwendung von *pearson residuals* geht auf die ursprüngliche Anwendung der Korrespondenzanalyse auf Kontingenztafeln zurück. Dabei spielt das **Unabhängigkeitsmodell** eine entscheidende Rolle: Die beobachteten Zahlen in einer bestimmten Zelle der Tafel werden mit den Zahlen verglichen, die zu erwarten sind, wenn Reihen und Spalten der Kontingenztafel **unabhängig** sind. Abweichungen von dieser Unabhängigkeit werden mit den *pearson residuals* gemessen.

Biplots zu diesem Modell können nicht mit dem Package **ca** erzeugt werden; unter Verwendung des `cabipl` Programs mit dem Parameter `ca.map = "PearsonResB"` wird das Biplot in Abbildung 23 erzeugt.

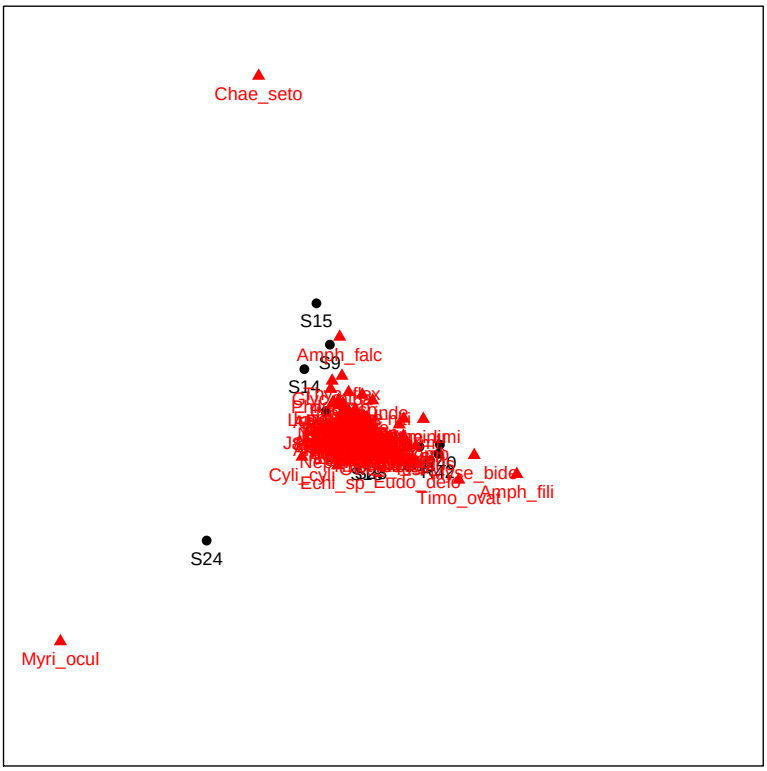

Abbildung 23: Biplot (CA) : Pearson Residuals

Auch hier kann wieder eine *Bevorzugung* von Arten mit höherer Gesamtabundanz festgestellt werden (siehe auch Abbildung 26). Das obige Biplot kann auch mittels einer Hauptkomponentenanalyse mit den *pearson residuals* erzeugt werden (siehe Abbildung 24). Um völlige Übereinstimmung zu erreichen, muss eine PCA-Variante mit `center = FALSE` verwendet werden.

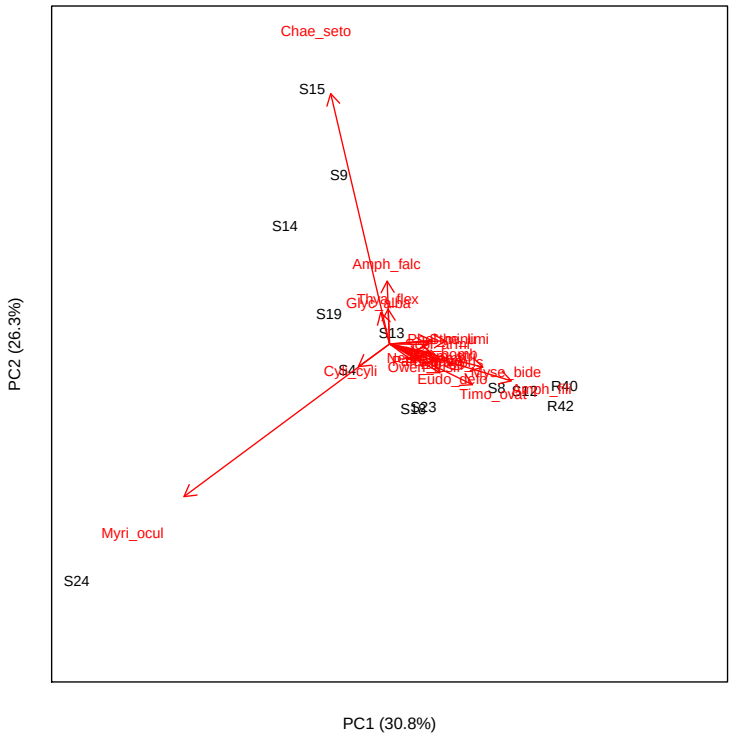

Abbildung 24: Biplot (PCA) : Pearson Residuals / 20 ausgewählte Variable

Bei der kalibrierten Version (Abbildung 25) ist zu beachten, dass sich die Kalibrierung auf die *pearson residuals* bezieht.

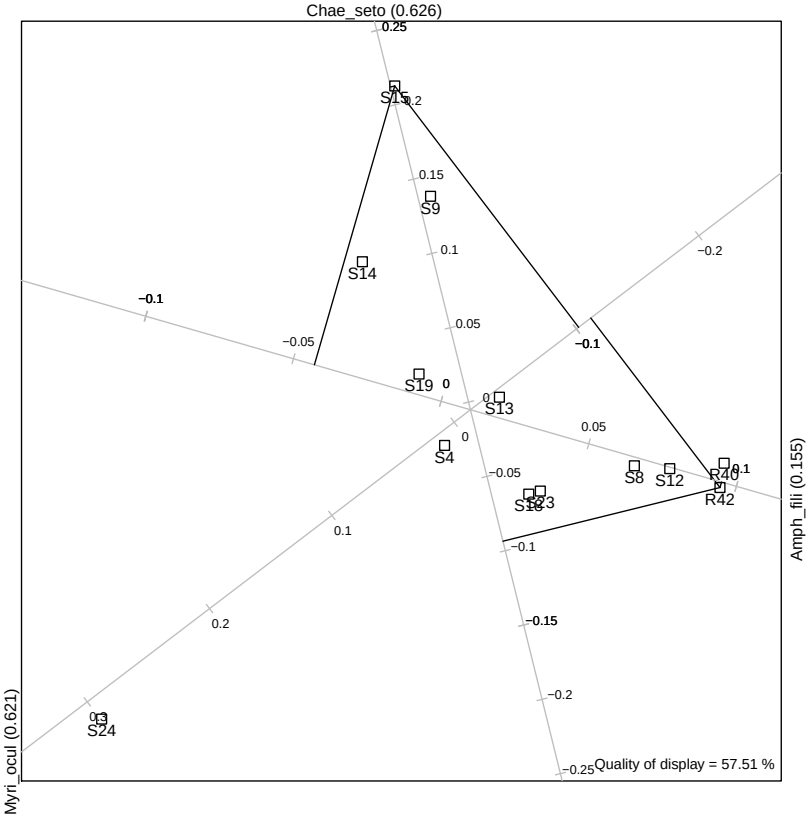

Abbildung 25: Biplot (PCA) : Pearson Residuals / 3 Achsen kalibriert

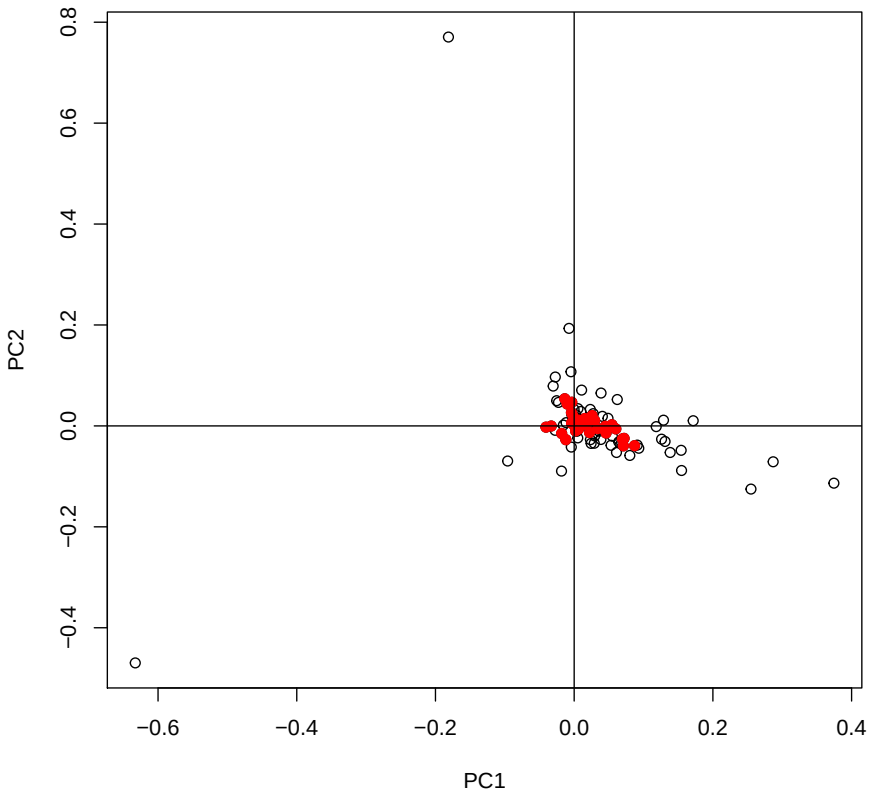

Abbildung 26: Biplot Ladungen: Pearson Residuals (seltene Arten rot)

4.2 Contingency ratio

Ein Biplot, das auf einer Approximation des *Contingeny ratios* beruht, kann sowohl mit dem package **UBbipl** (Programm `cabipl, ca.variant = "ConRatioB"`), als auch mit dem package **ca** (`map = "rowprincipal"`) erzeugt werden (siehe Abbildung 27 für die `rowprincipal` Variante). Dieses Ergebnis ist inhaltlich identisch zur Abbildung **Exhibit 8.3** (Seite 86 in [1]). *The low frequency points often have unusual profiles and lie on the periphery of the map, giving an impression of high importance - for example, in the benthos application a very rare species, occuring in just two or three sites, will have a profile at the outer reaches of the profile space.*

Es ist nicht möglich, dieses Biplot durch eine gewichtete oder ungewichtete Hauptkomponentenanalyse zu konstruieren. Es gibt jedoch eine interessante Eigenschaft, die nur dieser Variante zu eigen ist: die *centroid property*. Diese Eigenschaft besagt, dass die einzelnen Punkte (Probestellen) im Biplot als gewichtetes Mittel aller Variablen-Ladungen interpretiert werden kann, gewichtet mit der relativen Abundanz der Arten. Diese Konstruktion ist vergeichbar der in der Ökologie verwendeten Methode der *Indexkonstruktion*.

Ausserdem gilt: *The centroid property helps with interpreting the biplot because the column points for a row with high weights will be tightly clustered around the corresponding row point* (Seite 293 in [2]).

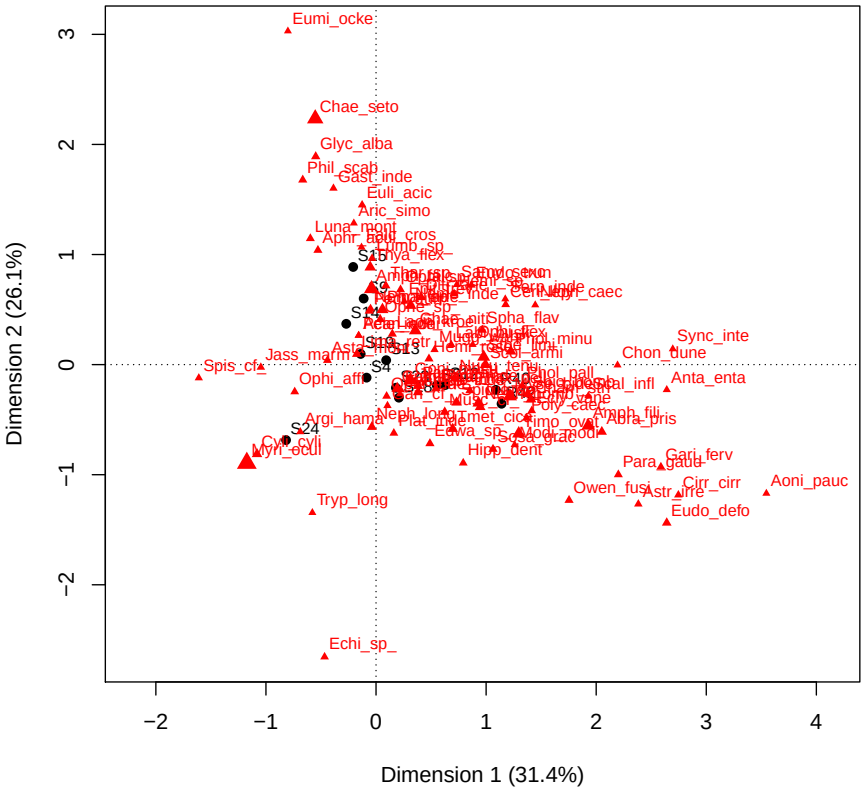

Abbildung 27: Biplot (CA) : contingency ratio

4.3 Bray-Curtis dissimilarity

Die Analyse von absoluten Abundanzdaten wird häufig auf der Basis der *Bray-Curtis dissimilarity* durchgeführt. Zur Konstruktion eines Biplots ist dabei die Schätzung von optimalen Gewichtskoeffizienten erforderlich. Abbildung Fig. 28 zeigt die erreichte gute Anpassung.

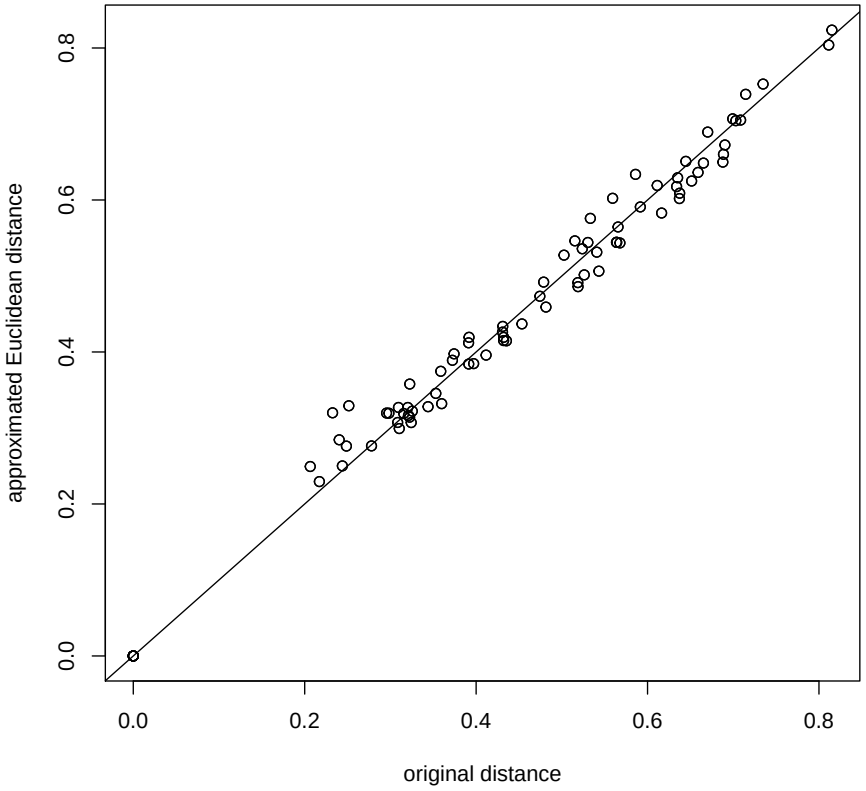

Abbildung 28: Vergleich von Original- und approximierter Distanz Matrix

Unter Verwendung dieser Gewichte kann eine gewichtete Hauptkomponentenanlyse (wPCA) durchgeführt werden. Abbildung 29 zeigt das zugehörige Biplot.

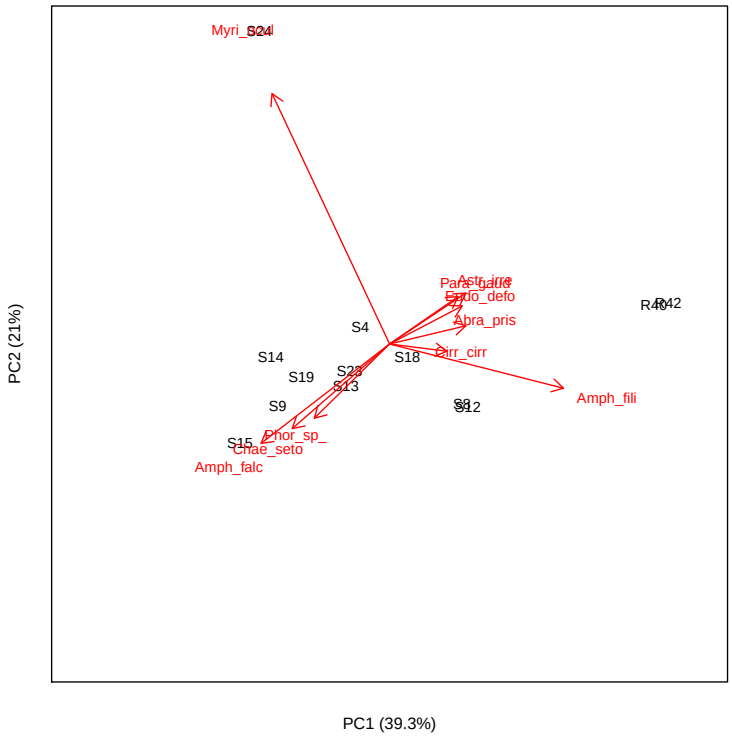

Abbildung 29: Biplot (wPCA) : Bray-Curtis dissimilarity / 10 ausgewählte Variable

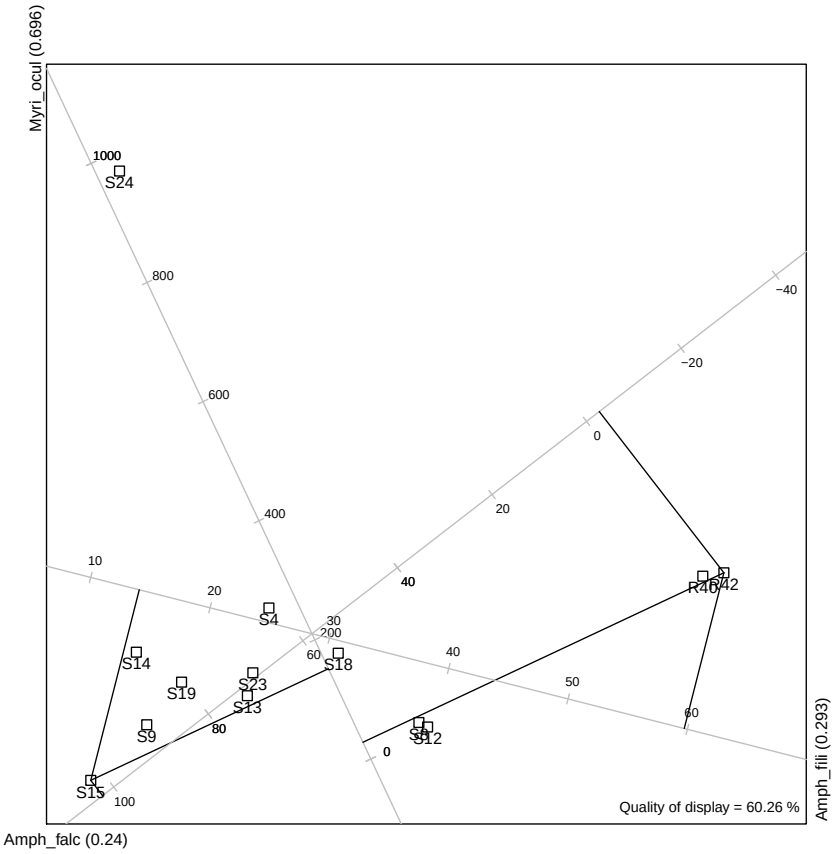

Abbildung 30: Biplot (wPCA) : Bray-Curtis dissimilarity / 3 Achsen kalibriert

Die Kalibrierung bezieht sich auf die absoluten Abundanzen. In Abbildung 30 ergeben sich dabei für die Probestelle S15 fogende approximierte Werte: 151.5 (*Myri_ocul*), 102.4 (*Amph_falc*) und 14.1 (*Amph_fili*). Zu erwähnen ist noch die gute Approximation des *Ausreissers* S24 mit einem approximierten Wert von 971.1 (Originalabundanz 992).

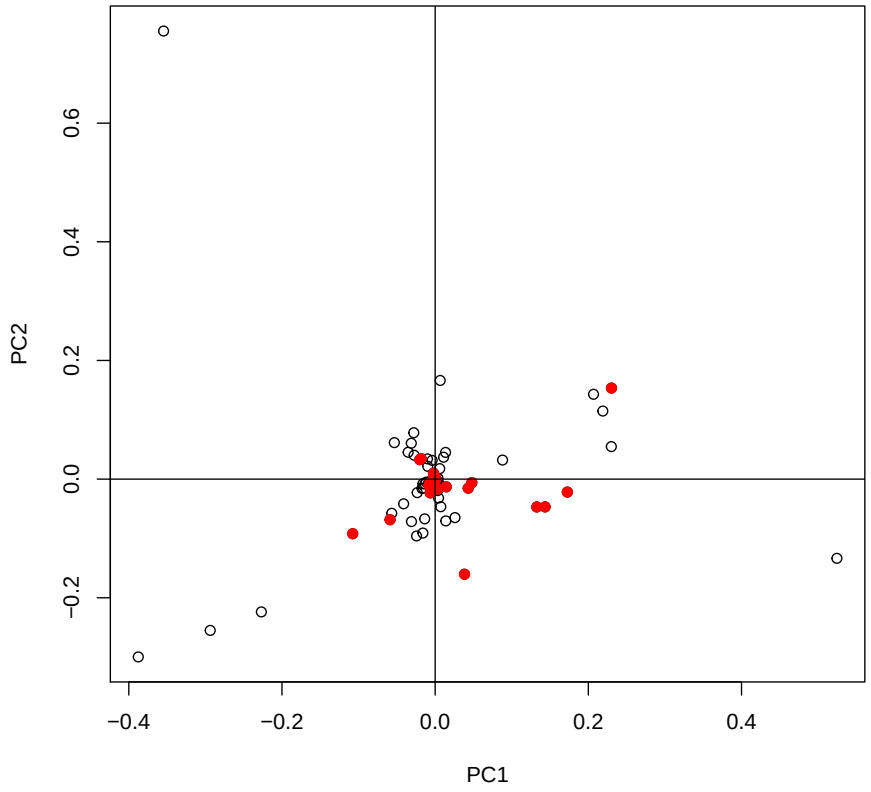

Abbildung 31: Biplot Ladungen: Bray-Curtis dissimilarity (seltene Arten rot)

Aus Abbildung 31 lässt sich erkennen, dass auch seltenere Arten eine höhere Gewichtung erhalten können.

5 Zusammenfassung

Der Vergleich der Ergebnisse mehrere Analysemethoden an **einem** Datensatz hat natürlich nur beschränkte Aussagekraft. Aussagen können gewonnen werden über die technische Durchführbarkeit mit vorhandenen Programmen sowie über die Auswirkung einiger Besonderheiten des Datensatzes.

5.1 Programme

Es konnte gezeigt werden, dass durch die Verwendung von *Weighted Euclidean Biplots* im wesentlichen *ein* Programm zur Konstruktion von Biplots für beliebige Distanzmaße genügt.Dieses Programm basiert auf einer Haupkomponentenanalyse, *preserving all the good properties of dimension-reduction methods that are based on the singular-value decomposition* ([4]).

5.2 Vergleich der Methoden - relative Abundanz

Es wird angenommen, dass sich im *benthos* Datensatz Auswirkungen der Ölgewinnung auf das maritime Leben nachweisen lassen. Zum Nachweis wurden Biplots konstruiert, die auf der Approximation unterschiedlicher Distanzmatrizen beruhen:

- Chi-square distance
- Hellinger distance
- Bhattacharya distance
- Bray-Curtis dissimilarity
- Yue-Clayton dissimilarity

Der Unterschied zwischen Kontroll-Probestellen und Probestellen in der Nähe des Ölfeldes kann mit allen Methoden nachgewiesen werden. Ein größerer Unterschied zwischen den Methoden ergibt sich offensichtlich bei der Behandlung von *Ausreissern*. Während bei *Chi-square distance*, *Bray-Curtis dissimilarity* und *Yue-Clayton dissimilarity* dieser Effekt erhalten bleibt, erfolgt bei *Hellinger distance* und *Bhattacharya distance* eine deutliche Abschwächung.

Diese Besonderheit lässt sich auch aus Abbildung 32 ableiten. Hier werden die sich aus der Biplotkonstruktion für alle Methoden ergebenden approximierten relativen Anbundanzen gegenübergestellt und mit den tatsächlichen verglichen (für *Myri_ocul* und *Chae_seto*).

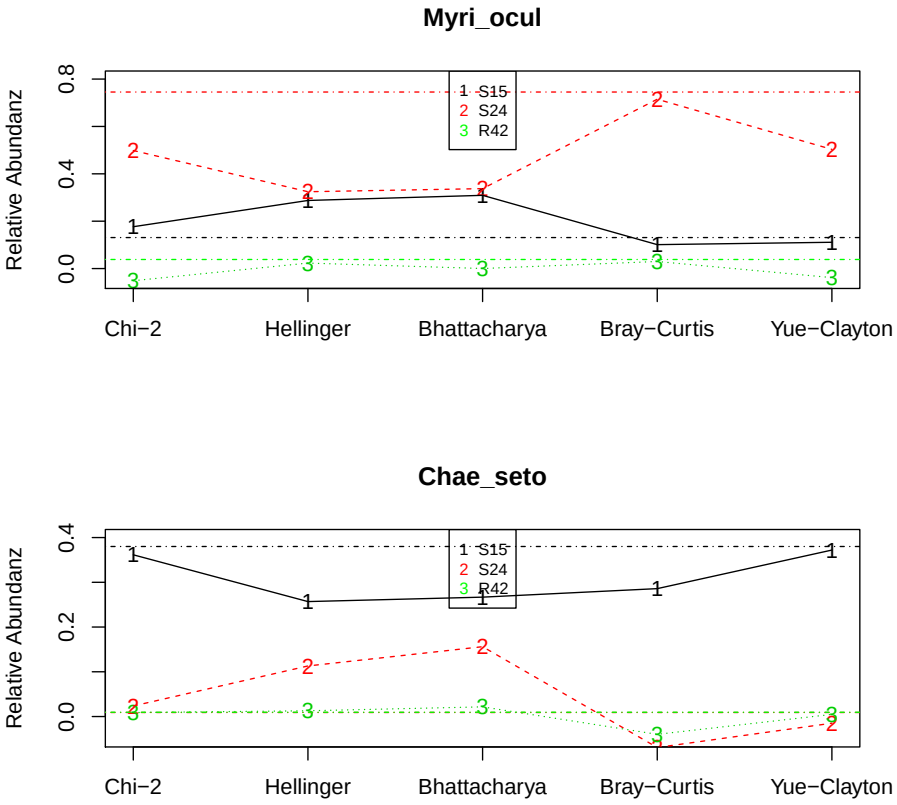

Abbildung 32: Vergleich der approximierten relativen Abundanzen basierend auf verschiedenen Distanzmaßen / gerade Linien = Originalwerte

Es ist aus inhaltliche Überlegungen zu klären, welche *Ausreisser* -Behandlung vorzuziehen ist.

Literatur

[1] Michael Greenacre. *Biplot in Practice*. Bilbao, Madrid: Fundación BBVA (FBBVA), 2009.

[2] John Gower und Sugnet Lubbe. *Understanding Biplots*. Chicester West Sussex: Wiley, 2011.

[3] C.M. Cuadras D. Cuadras. „A Unified Approach for the Multivariate Analysis of Contingency Tables“. In: *Open Journal of Statistics* 5 (2013), S. 223–232. URL: `http://dx.doi.org/10.4236/ojs.2015.5353024`.

[4] Michael Greenacre und Patrick J.F. Groenen. „Weighted Euclidean Biplots“. In: *Barcelona GSE Working Paper Series* n708 (2013), S. 1–20.

[5] Michael Greenacre und Raul Primicerio. *Multivariate Analysis of Ecological Data*. Bilbao, Madrid: Fundación BBVA (FBBVA), 2013.