Überflüssige Information? Zum Verständnis moderner Deutung archäologischer Zusammensetzungsdaten mit transformationsbasierter Redundanzanalyse (tb-RDA)

Georg Roth

Zusammenfassung

Zu den häufigsten multivariaten Datensätzen in den Archäologien gehören Kreuztabellen, die in den Zeilen Inventare (Fälle) beschreiben, indem sie in den Spalten die Anzahlen verschiedener Typen bei diesen Fällen vermerken. Wie lassen sich die Ähnlichkeiten und Unterschiede bei den Anteilen der Typen zwischen den Inventaren deuten? Wenn bei den Inventaren die möglichen Ursachen auch bekannt und mit Merkmalen beschrieben sind, bietet die transformationsbasierte Redundanzanalyse (tb-RDA) – eine Standardmethode der Ökologie u. a. für Naturschutzfragen – eine erprobte, objektive Methode, die historisch verallgemeinerbare Deutungen erlaubt. Die tb-RDA liefert Antworten für den Auftrag: „Ordne mir die Fälle nach Ähnlichkeit, schließe dabei Zufallseffekte aus und zeige mir dann v. a. die auf die Ursachen zurückgehenden Unterschiede". Der Beitrag erläutert auf zwei Ebenen, einer sprachlich einfachen metaphorischen und einer terminologisch präzisen, die Funktionsweise und Deutungsmöglichkeiten einer tb-RDA. Eingangs wird geklärt, wie sich neuerdings Anzahlen sinnvoll mit Hauptkomponentenanalyse (PCA) und RDA bearbeiten lassen. Es folgen Erklärungen der Einzelschritte der RDA (Regression, PCA) sowie Erläuterungen und Deutungsregeln zur höchst informationsreichen Ergebnisgrafik (Triplot). Alle Schritte sind mit erklärenden schematischen Grafiken illustriert. Eingeflochten sind Hinweise auf interessante Varianten der tb-RDA. Geschlossen wird mit Hinweisen zu Software-Implementierungen und epistemologischen Überlegungen zu den ganz neuen Deutungsmöglichkeiten, welche diese Methode ermöglicht.

Schlüsselwörter: *Analyse von Typentabellen, Typen-Abundanz-Paradox, ursachenbasierte Ähnlichkeitsanalyse, Zusammensetzungsanalyse, Redundanzanalyse (RDA), Methodenverständnis der RDA*

Summary

One of the most common kinds of multivariate data in archaeology are crosstabulations of assemblages by types with counts of different types in columns. How can one interpret similarities and differences in the proportions of types between the assemblages? If attributes assumed to represent causal forces are also known for the assemblages, transformation-based redundancy analysis (tb-RDA) offers a weil understood multivariate statistical tool to answer such questions. This canonical ordination method allows to test, if assumed causes can be considered random noise or represent real general historical relations. Formulated as instructions RDA solves the task of: "Arrange the assemblages according to similarity, exclude random effects and only show me those differences that are related to non-random causes". The article explains the functionality and possible interpretations of tb-RDA. To start with the old problem arising when count data is submitted to PCA (and RDA) is solved by introducing Chord and Hellinger transformations for count data. Consequently all steps of an RDA (linear regression and PCA) are explained and interpretation rules for its graphical result, the triplot, are offered. The article closes with considerations on software and epistemological issues arising from the new possibilities for interpretation enabled by RDA.

Keywords: *multivariate analysis, type abundance paradox, transformations for multivariate count data, canonical ordination, redundancy analysis (RDA), principles of method*

1. Einführung

Redundanz bezeichnet im allgemeinen Sprachgebrauch meist Überfluss oder mehrfach Vorhandenes, auf Vorrat gehaltenes, kurz, Überflüssiges. Redundanzanalyse (RDA) als Namen für eine multivariate, kausal deutende statistische Methode zu verwenden, scheint zunächst unglücklich. Ursprünglich wurde das hier zu erörternde Verfahren denn auch von seinem Entdecker, dem indischen Statistiker Calyampudi Rao, 1964 als ‚Hauptkomponentenanalyse mit instrumentellen Variablen' (engl. *principal components of intrumental variables*, heute abgekürzt als PCAIV; Rao 1964) bezeichnet. PCAIV ist die heute noch gängige Bezeichnung im französischen Sprachraum (Lebreton *et al.* 1991). Erst in der 1977 durch Arnold Van den Wollenberg unabhängig von Rao gemachten Wiederentdeckung der Methode, welche in der weitverbreiteten Zeitschrift ‚Psychometrika' veröffentlicht wurde, erschien der Name Redundanzanalyse (Legendre *et al.* 2011, 630). Dabei bezog sich Van den Wollenberg auf den 1968 von Douglas Stewart und William Love eingeführten, sog. Nicht-symmetrischen Index der Redundanz (engl. *nonsymmetric index of redundancy*), welcher schlicht synonym für das Konzept erklärter Streuung (engl. *explained variance*; Gittins 1985, 40) in der Regressionsanalyse bei multivariaten Abhängigen ist. Wenn ein Phänomen ein anderes als seine Auswirkung hervorbringt, dann ist die im verursachten Phänomen ausgedrückte Information zum Teil schon im verursachenden Phänomen enthalten. Formuliert mit Blick auf eine Datentabelle bedeutet dies,

die Information über die Ähnlichkeit der Fälle, welche in einem Bündel die Fälle beschreibender, verursachter Merkmale gemeinsam enthalten ist, ist redundant, wenn es kausal wirkende Merkmale gibt, die Reaktionen beim Merkmalsbündel verursachen. Kurz gesagt, die Redundanzanalyse erlaubt es, die Ursache oder die Ursachen eines mittels vieler Merkmale beschriebenen Phänomens zu ergründen (Lambert *et al.* 1988, 288). Die Beschreibung besteht aus vielen Merkmalen, weshalb die RDA zu den multivariaten Methoden gehört. Weil sie die Fälle nach der Ähnlichkeit bei ihren vielen Merkmalen anordnet, gehört sie zum Teilfeld der Ordinationen (engl. *ordination*; Legendre und Legendre 2012, 425). Da zugleich die Ursache-Wirkung-Beziehung in einer rechnerisch möglichst einfachen (= kanonischen) Form ausgedrückt wird, gehört sie dort zur Unterabteilung der kanonischen Ordinationen (Legendre und Legendre 2012, 625). Schließlich wird sie auch als ‚asymmetrisch' bezeichnet (Legendre *et al.* 2011, 269), denn die zwei beteiligten Daten(-teil-)tabellen gehen nicht rechnerisch gleichberechtigt in die Analyse ein. Die der verursachenden Merkmale hat größere Bedeutung für das Ergebnis, die andere wird stets ‚in Abhängigkeit' verrechnet. Andere, eng verwandte, asymmetrische kanonische Ordinationsmethoden sind die Kanonische Korrespondenzanalyse (abgek. CCA; ter Braak 1986) und die distanzbasierte Redundanzanalyse (abgek. db-RDA; Legendre und Anderson 1999).

Die Funktionsweise der RDA sei hier im Hinblick auf ihren Einsatz für archäologische Fragestellungen teils mit umschreibenden, metaphorisch gehaltenen Erklärungen zum besseren Verständnis und teils mit technischen Bezeichnungen zur Anknüpfung an die Methodenlehrbücher dargestellt. Der technische Teil lehnt sich dabei eng an die Darstellung im Lehrbuch von Pierre und Louis Legendre zur Numerischen Ökologie an (Legendre und Legendre 2012, Kap. 11.1). Dort finden sich auch die zugrunde gelegten matrixalgebraischen Definitionen in präziser Ausführung. Die hier immer wieder für Erklärungen angeführten Passagen mit einem geometrischen Verständnis der beiden RDA-Teile, der linearen Regression und der Hauptkomponentenanalyse, sind stark inspiriert von Darstellungen in einem Band von Thomas Wickens (Wickens 1995). Dieses Werk bietet einen guten Einstieg in das allgemeine Prinzipienverständnis multivariater Methoden.

Gehört die Darstellung eines statistischen Verfahrens in ein archäologisches Buch? Für die datenbasierten Archäologien des 21. Jahrhunderts sind Verständnis und Anwendungssicherheit von datenauswertenden Verfahren ebenso Teil der Praxis wie das Verständnis und die sachgerechte Anlage eines Profilschnittes. Allerdings nehmen selbst neue Methodenlehrbücher wenig Rücksicht auf geisteswissenschaftliche Leserinnen und Leser. Warum aber dann eine Methodenerörterung als Beitrag zu einer Festschrift? Im Wintersemester 2017/18 führte der Gefeierte mit dem Autor zusammen die m. W. erste umfassende archäologische Lehrveranstaltung deutscher Sprache zum Thema zeitgenössischer multivariater Datenanalyse durch. Dabei hatten wir unsere liebe Not, den Studierenden und Promovierenden erklärende Text in deutscher Sprache zu bieten. Um dies für die RDA zu bessern, sei dem geduldigen Lehrer, weitsichtigen Wissenschaftler und hochgeschätzten Vorgesetzten deshalb hier in Dankbarkeit die, m. W. in deutscher Sprache so noch nicht vorhandene, zum pragmatischen Verständnis beitragende und als Vorbereitung zum praktischen Einsatz gedachte, RDA-Methodenskizze zugeeignet. Die Darstellung changiert dabei zwischen geradezu umgangssprachlicher Ausdrucksweise und technischen Beschreibungen, denn es geht ja nicht um das Vorexerzieren des eigenen Könnens, sondern um ein möglichst gut verständliches Informationsangebot.

Bei der Erörterung von rechnerischen Methoden heißt es häufig, es müsse alles am konkreten Beispiel erläutert werden. Allerdings entstand so nicht selten das Missverständnis, eine Methode sei an Daten zu bestimmten Ausschnitten der Welt gebunden. Hier sei deshalb einmal der umgekehrte Weg gegangen, der vom Abstrakten zum Konkreten weisend erklärt. Denn der allgemeine gedankliche Schritt von den Voraussetzungen und Funktionsweisen einer Methode zu ihrer Anwendung zeigt erst, für welche verschiedene Fragestellungen sie eingesetzt werden kann. Und wer diesen Schritt bewältigt, kann sich seines Verständnisses viel sicherer sein, als der, der einen Datensatz durch einen gleichartigen ersetzend, eine Analyse quasi in Kopie nachbildet.

2. Allgemeine Funktionsweise und Begriffe

Wozu ließe sich denn eine RDA nun einsetzen? Die unkonkrete, aber richtige Antwort lautet: zu erstaunlich Vielem. Das Einsatzfeld lässt sich am besten als umgangssprachliche Handlungsanweisung an die als personifiziert gedachte Methode formulieren: „RDA, ordne mir meine Fälle nach Ähnlichkeit an, und zeige mir dabei vor allem den Teil der Ähnlichkeit, welcher auf die von mir vermutete Ursache zurückgeht". Sofort wird klar, die RDA ist im Zeitalter der (multivariaten) Datensätze eine unverzichtbare Methode der analytisch-deutenden Archäologie. Aber die Methode bietet noch mehr. Wieder ausgedrückt als direkte Ansprache: „Prüfe auch, ob die vermutete Ursache eine vom Zufall unterscheidbare Auswirkung hat, messe die Stärke der Auswirkung und zeige mir eine zusammenfassende Ergebnisgrafik". Die Begriffe ‚Fälle', ‚Datentabelle', ‚Ähnlichkeit', ‚Prüfen' und ‚Auswirkungsstärke' etc. werden gleich noch genauer erläutert. Aber es lässt sich schon sagen, die RDA bietet eine dreiteilige Antwort, wenn nach der Ursache eines sich in vielen Merkmalen auswirkenden Kausalphänomens gefragt wird. Sie prüft, ob die Auswirkung der Ursache auch ein Zufallseffekt sein könnte, enthält also einen statistischen Test. Wenn dieser bei einer RDA-Anwendung den Zufall nicht ausschließen kann, sollte sie bei diesen Daten nicht ausgeführt werden. Ist aber der Zufall durch den Test vernünftigerweise auszuschließen, wird es spannend.

Der nächste Ergebnisteil der RDA bietet nämlich schon ein Maß für die Stärke der Kausalbeziehung, das sog. kanonische R^2 – lies: R-Quadrat. Dies ist ein dem multiplen Determinationskoeffizienten (R^2) eng verwandtes Maß. Am informationsreichsten und anschaulichsten jedoch ist eine detailreiche Ergebnisgrafik, der sog. Triplot, welcher die Ähnlichkeitsanordnung der Fälle, die Kausalreaktion der verursachten Merkmale (abgek. VM) und ihrer beider (Fälle und VM) Verbindung zu dem oder den Kausalmerkmal(en) (abgek. KM) zeigt. Diese beiden Abkürzungen, VM und KM, sind für das weitere Textverständnis grundlegend. Der Name Triplot basiert darauf, dass drei Entitäten, Fälle, VM und KM gemeinsam in ihren deutbaren Beziehungen dargestellt werden. Die Zahlengrundlagen des Triplots werden natürlich auch ausgegeben. Bei Interesse ist für die einzelnen VM messbar, wie stark sie kausal vom KM beeinflusst wurden. Die Worte ‚Merkmal' und ‚Variable' sind im Folgenden synonym.

Nach diesen zwar einfach verständlichen, aber doch eher vagen Ausführungen ist es an der Zeit, zunächst die einzelnen Begriffe genauer mit Inhalt zu füllen. Was ist also gemeint mit Fällen, Ähnlichkeiten etc. und welche Form sollte der Datensatz haben?

Mit ‚Fall' ist ganz allgemein irgendein Objekt gemeint, das beschrieben wird. Beschrieben wird es mit Merkmalen, auch Variablen oder – selten – Deskriptoren geheißen; das sind die oben als VM bezeichneten. In einem Datensatz einer Datentabelle steht eine Zeile für einen Fall und seine Werte bei den jeweiligen Merkmalen (VM) stehen in einem Block von Spalten. In einem weiteren Block von Spalten stehen die Werte der Fälle bei den Kausalmerkmalen (KM). Man könnte sich eine solche Datentabelle als aus zwei Blöcken bzw. Teiltabellen zusammengesetzt vorstellen. In der einen stehen die Werte der VM der Fälle, in der anderen die Werte der KM bei den Fällen.

Für eine klassische RDA müssen die VM kontinuierliche ratioskalierte Variablen sein. D. h. die Variablenausprägungen liegen als Zahlenwerte, u. U. mit Dezimalstellen, vor, die i. d. R. Maßeinheiten besitzen (Stevens 1946). Ratio bedeutet Verhältnis oder Bruch. Bei einer kontinuierlichen Ratiovariable ist also das Vielfache oder der Bruchteil eines Zahlenwertes eine sinnvolle Information. Leicht erkennbar schließt dies im engeren Sinn Anzahlen aus, denn das 2,3-fache der Anzahl 4, der Wert 9,2, existiert nicht als Anzahl. Das mag noch verschmerzbar sein und mit Rundung zu lösen. Das sich dahinter verbergende, bei Hauptkomponentenanalyse (engl. *principal component analysis*, PCA) und RDA das Prinzip von Anzahlen tatsächlich verletzende Problem wird gleich noch diskutiert. Die hier erläuterte Variante, die transformationsbasierte RDA (tbRDA) löst dieses Problem, wie gleich gezeigt wird.

Wie wird bei solchen Beschreibungen die Ähnlichkeit mit nur einer Zahl ausgedrückt? Bevor zwei Fälle, also zwei Zeilen einer Datentabelle, verglichen werden können, ist bei Ratiovariablen mit unterschiedlichen Maßeinheiten eine Vorverarbeitung nötig (Legendre und Legendre 2012, 152). Diese garantiert, dass solche Merkmale gemeinsam verwertet werden können. Dazu wird von den Werten einer Spalte der Mittelwert der Spalte abgezogen und das Ergebnis durch die Standardabweichung der Spalte geteilt. Dieses Vorgehen heißt Z-Transformation oder Normalisierung. Liegen aber alle Ratiovariablen in den gleichen Maßeinheiten vor, kann die Z-Transformation übersprungen werden. Nach der Transformation haben die Merkmale keine Maßeinheit mehr und sind deshalb zusammen verrechenbar. Auch liegen alle Zahlenwerte transformierter Merkmale ungefähr im gleichen Größenbereich. Im Vorgriff sei betont: Sind die VM als nach Typen bzw. Kategorien ausgezählte Objektanzahlen, also multivariate Anzahldaten, ist eine Z-Transformation unangemessen, denn das eigentliche Problem löst eine andere Transformation (s. u.).

Man denke sich nun die Merkmalswerte (VM) eines Falles – transformiert oder nicht – als seine Koordinatenangaben. Ein Fall ist so als ein Punkt in einem Raum vorstellbar, der so viele Dimensionen (Koordinatenachsen) besitzt, wie die Datentabelle der VM Spalten. Wenn zwei Fälle ähnliche Werte haben, liegen diese beiden Punkte demnach nahe beieinander. Dass man sich einen Raum mit z. B. 15 Dimensionen nicht vorstellen kann, macht nichts. Auch in einem solchen Raum gilt, der Abstand zwischen zwei Punkten ist mit dem Pythagorassatz, den wir alle seit der 7. Klasse kennen, berechenbar. Dieser Abstand heißt euklidische Distanz – es gibt auch andere Distanzberechnungen (Legendre und Legendre 2012, Kap. 7). Kurz und vereinfacht gesagt, ein Distanzmaß ist der Ausdruck des Unterschiedes zwischen zwei mit vielen Merkmalen beschriebenen Fällen vermittels einer einzigen Maßzahl (*ibid.*). Für alle Arten von Merkmalen gibt es geeignete Distanzmaße (*ibid.*). Eine RDA bemisst die Ähnlichkeit zwischen Fällen mit der euklidischen Distanz ihrer u. U. transformierten Merkmalswerte. Man sagt, sie ‚erhält' die euklidischen Distanzen der Fälle (a. a. O., 630). Ein Wert nahe Null bedeutet, die Fallpunkte liegen nahe beieinander, sie sind sich ähnlich. Größere Werte bedeuten, die Fälle sind sich eher unähnlich.

3. Das Typen-Abundanz-Paradox und seine Lösung (tb-RDA)

Nicht nur in den Archäologien wurde lange Zeit übersehen, dass Methoden, die wie die RDA oder die ihr verwandte PCA eigentlich Situationen benötigen, bei der die Distanzberechnung allein mit dem Pythagorassatz Sinn macht. Bei multivariaten Anzahldaten – also, wenn die VM alle ausgezählte unterschiedliche Kategorien sind und die Merkmalswerte alle ganze Zahlen bzw. Anzahlen – ist die euklidische Distanz ungeeignet. Das ist zunächst enttäuschend, sind doch Anzahldaten mitunter die spannendsten multivariaten Daten in den

Archäologien. Nur ein paar Beispiele seien genannt: Münzhorte, die nach Münztypen ausgezählt sind, Keramikinventare, die nach Gefäßform- oder Gefäßverzierungstypen ausgezählt sind, Silexinventare, die nach Rohmaterialarten oder Werkzeugtypen ausgezählt sind, Bronzehorte, die nach Gerätetypen ausgezählt sind und natürlich archäozoologische oder archäobotanische Inventare, die nach Lebewesenarten (oder Taxa) ausgezählt sind. Die Fälle sind die einzelnen Horte, Inventare, Assemblagen, kurz, die betrachteten Phänomene. Die Merkmale, genauer die multivariaten VM, sind die Anzahlen bei den einzelnen Typen. Gemeinsam ist ihnen allen, dass Zeilen- und Spaltensummen dieser Anzahldaten Sinn machen. Die Zeilensumme gibt an, wie viele der unterschiedlichen Typen es insgesamt bei einem Fall/Inventar/Hort gibt. Die Spaltensumme vermerkt, wie viele es insgesamt von einem Typen, also aufsummiert über alle Fälle, gibt. Jede Datentabelle, die solchermaßen sinnvolle Randsummen aufweist stellt multivariate Anzahldaten dar, die als VM für eine tb-RDA (s. u.) interessant wären. Kausalauswirkungen auf solche VM dürfen aber nicht direkt mit einer klassischen RDA untersucht werden, bzw. es ist vorher eine ganz bestimmte Art von Datentransformation anzuwenden. Sie dürfen nicht einmal mit einer PCA erkundend geordnet werden, was sogleich begründet wird.

Die Ursache für diese Untersagung einer Methodenanwendung ist das sog. Anzahldaten-Problem. Dies hat in der Ökologie vor über 40 Jahren der ungarisch-kanadische Wissenschaftler László Orlóci erkannt und als Arten-Abundanz-Paradox bezeichnet (Orlóci 1978, 46). Spricht man von einem Paradox, geht es darum, dass irgendetwas auf überraschende Weise unangemessen ist oder unpassend reagiert. Abundanz ist ein Ausdruck für Anzahl oder Menge. Das Paradox hat also etwas mit Mengen bzw. Summen zu tun. Es sei an einem einfachen Beispiel erläutert. Eine Datentabelle beschreibe für drei Horte die Anzahl von jedem dreier Gerätetypen A, B und C. Hort I umfasse vom Typ A und vom Typ B je 10 Exemplare und 0 vom Typ C. Hort II umfasse vom Typ A und vom Typ B je 1 Exemplar und 0 vom Typ C. Hort III schließlich weist 0 Exemplare vom Typ A auf und jeweils 1 vom Typ B und Typ C. Wenn die ähnliche Zusammensetzung der Horte interessiert, ist klar, dass Hort I und Hort II bei einem sinnvollen Abstandsmaß einen Unterschied von 0 aufweisen sollten, denn sie bestehen zu jeweils 50 % aus den Typen A und B. Beide unterscheiden sich vom Hort III dadurch, dass sie keinen Typ C aufweisen. Es besteht zwar zwischen Hort I und Hort II ein Unterschied bei der Fundmenge, bei Hort I sind es 20 Stücke, bei II nur 2 Stücke, aber das ist kein Zusammensetzungsunterschied. Wird nun die euklidische Distanz zwischen den Paarungen aufgrund dieser Anzahldaten berechnet, entsteht das Paradox, dass die beiden Horte mit exakt gleicher anteiliger Zusammensetzung, nämlich I und II, anstatt des zu erwartenden Distanzwertes 0 als ungleicher dastehen, als etwa die Paarung II und III. Ein Distanzwert von 0 ist die Erwartung an ein Distanzmaß für die Unterschiedsmessung bei exakt gleichen Zusammensetzungen – bei 0 bestehen schlicht keine Unterschiede. Die euklidische Distanz des Paares I/II ist aber 12.73 $(((10-1)^2+(10-1)^2+(0-0)^2)^{0.5} = 12.73)$, die des Paares II/III beträgt überraschenderweise nur 1.41 $(((1-0)^2+(1-1)^2+(0-1)^2)^{0.5} = 1.41)$ und die des Paares I/III 13.49. Das Paradoxe ist also, dass das Distanzmaß ‚euklidische Distanz' bei multivariaten Anzahldaten, die auf ihre Zusammensetzung hin untersucht werden sollen, eine höchst unerwünschte Reaktion zeigt – und zwar systematisch. Auch das zur Lösung schon mal erwogene Quadrieren der euklidischen Distanzen, also eigentlich das Weglassen des letztendlichen Wurzelziehens, löst das Problem nicht, es würde im Gegenteil schlimmer (Distanz I/II =162 und Distanz II/III =2).

Da in den Archäologien bei multivariaten Anzahl-Datensätzen meist die Typen an die Stelle treten, welche bei ökologischen multivariaten Anzahl-Datensätzen die Arten einnehmen – Archäobotanik und Archäozoologie mal kurz ausgenommen –, dürfen wir in den Archäologien dieses Paradox auch als ‚Typen-Abundanz-Paradox' bezeichnen. Und das Paradox ist kein Taschenspielertrick. Es ist ein ernsthaftes Problem für die multivariate Auswertung von Anzahldaten. Und es ist eigentlich leicht verständlich, denn im Pythagorassatz werden die Fall- bzw. Zeilendifferenzen ja ‚entlang der Spalten' summiert, also spielen die Zeilensummen, vermittelt über die Größe der einzelnen Spaltendifferenzen, eine Rolle. Das sollten sie aber nicht, wenn man nur Zusammensetzungen analysieren möchte. Nein, keine voreiligen Schlüsse ziehen, das Ersetzen der Anzahlen mit Prozentwerten ist nicht die Lösung.

Die Lösung des Typen-Abundanz-Paradoxes erfolgt durch einen eleganten und flexiblen Umwandlungsschritt der Anzahldaten (VM), eine andere Art Transformation. Eine RDA, die auf die gleich beschriebene Weise verwandelte Anzahldaten als VM analysiert, heißt denn auch transformationsbasierte RDA (tb-RDA; Legendre und Legendre 2012, 647 f). Nun zur Lösung: In der Ökologie sind seit langem sinnvolle Distanzberechnungen für multivariate Anzahldaten bekannt (a. a. O. Kap. 7), wovon u. a. zwei zur Grundlage der Problemlösung gemacht werden können und hier näher beleuchtet seien. Das ist zum einen die 1967 von Orlóci vorgeschlagene Chorddistanz (Orlóci 1967) und zum anderen die 1995 von Rao so bezeichnete Hellingerdistanz (Rao 1995). Die Chorddistanz ist ein ideales Maß für den Unterschied zweier nur mit Anzahlen beschriebener Inventare, wenn alle Spalten gleichartig zum Zusammensetzungsunterschied betragen sollen. Wenn dagegen manche selteneren Typen für die Unterschiede zwischen den Fällen bedeutsamer sein sollen als sog. Durchläufer, die bei allen Fällen auftreten, dann kann dies mit der Hellingerdistanz optimal erfasst werden. Die Problemlösung beruht nun auf der Einsicht, dass beide Distanzberechnungen in zwei aufeinanderfolgende Teilberechnungen zerlegt werden können (Legendre und Gallagher 2001). Im ersten Schritt wird ein für

das jeweilige Distanzmaß spezifischer Rechenschritt auf die Anzahlen ausgeführt und deren Werte werden dadurch umgewandelt. In einem zweiten Rechenschritt wird dann die Distanz zwischen zwei Fällen dadurch berechnet, dass der Pythagorassatz auf die so umgewandelten Anzahldaten angewendet wird. Der problemlösende und trickreiche Effekt dabei ist, werden die Anzahlen zuerst nur dem jeweiligen ersten Transformationsschritt einer Chord- oder Hellingertransformation unterworfen, also der Pythagoras zunächst weggelassen, ergibt die Analyse dieser so verwandelten Anzahlen in allen multivariaten Methoden, die die euklidische Distanz erhalten direkt aus den Methoden heraus sinnvoll deutbare Ergebnisse. Zu diesen Methoden gehören neben PCA und RDA auch die lineare Diskriminanzanalyse (ein Beispiel für die LDA Hellinger-transformierter Anzahlen bei Maier 2015, 129–132).

Das Typen-Abundanz-Paradox wird also durch Chord- oder Hellingertransformation gelöst. Das durch diese Vorgehensweisen dann quasi indirekt angewendete Distanzmaß ist das zur Transformation genutzte: z. B. zeigt die Ergebnisgrafik einer PCA von Chord-transformierten Anzahlen als Abstand zwischen den Fällen die Chorddistanz. Analoges gilt für die Hellingertransformation. Das bedeutet ganz allgemein, anstatt der für Ausreißer bzw. seltene Typen bochanfälligen Korrespondenzanalyse (CA) und ihrer kanonischen Form, der kanonischen Korrespondenzanalyse (CCA), können solchermaßen problemlösend transformierte Anzahldaten leichter und einfacher mit PCA und RDA analysiert werden. In der Archäobotanik wird die transformationsbasierte PCA (zurtb-PCA: Legendre und Legendre 2012, 462) bereits als erfolgreicher Ersatz für die CA angesehen (Zerl 2019, 47).

Diese beiden Transformationen sind außerdem kein numerisches Zauberwerk, sondern von Hand nachrechenbar. Für die Chordtransformation wird so vorgegangen: Alle Anzahlwerte einer Zeile werden quadriert, aufsummiert und die Wurzel dieser Summe gezogen. Durch die so berechnete Zahl werden nun alle Werte der Zeile geteilt. So wird für jede Zeile verfahren. Die Hellingertransformation ist noch einfacher: Alle Anzahlwerte einer Zeile werden aufsummiert und zunächst alle Werte der Zeile durch diese Summe geteilt – ja, das ergibt Anteile und ist wie die Prozentberechnung nur ohne das Multiplizieren mit 100. Von den Anteilen wird die Wurzel gezogen. Wiederum wird so für jede Zeile verfahren. Wenn die so transformierten Werte (kanonischen) Ordinationen unterworfen werden, welche die euklidische Distanz erhalten, ergibt sich die zum jeweiligen Transformationsschritt gehörende Distanz automatisch. Eine tb-RDA der Hellinger-transformierten Daten zeigt z. B. die Unterschiede zwischen den Fällen in Hellingerdistanz und gewichtet so die Unterschiede bei seltener auftretenden Typen etwas stärker. Bei einer Chordtransformation werden die Unterschiede entsprechend als Chorddistanz dargestellt und sind ganz ähnlich wie Unterschiede bei Prozentanteilen zu deuten.

Das Prinzip der Chordtransformation multivariater Anzahldaten hatte eigentlich schon 1975 der deutsch-argentinisch-israelische Ökologe Immanuel Noy-Meir (Noy-Meir *et al.* 1975) herausgearbeitet, aber damals nicht explizit empfohlen, weshalb diese Lösung weitere 25 Jahre schlummerte, bis sie durch Eugene Gallagher und Pierre Legendre von Neuem entdeckt und das Prinzip der Berechnungsaufteilung auch u. a. für die Hellingerdistanz beschrieben wurde (dies. 2001). Durch die Übernahme in sein internationales Lehrbuch hat Pierre Legendre (Legendre und Legendre 2012, Kap. 7.7) diese Behandlung multivariater Anzahldaten schließlich in der Ökologie zum internationalen Standard erhoben. Die Archäologien können hier also eine vielfach überprüfte Problemlösung einfach zu ihrem Vorteil übernehmen.

4. Ursachen-Repräsentation in der RDA

Das Phänomen, dessen Ursache(n) mit einer RDA untersucht werden soll(en), ist also bei der klassischen RDA eine Datentabelle bzw. ein Block aus einer Datentabelle mit ratioskalierten Merkmalen als VM. Bei der tb-RDA ist es eine Datentabelle bzw. ein Block aus einer Datentabelle mit Chord- oder Hellinger-transformierten Anzahldaten als VM. Die mögliche Ursache oder möglichen Ursachen müssen ebenfalls für jeden Fall bekannt sein. Das bedeutet, in der Datentabelle gibt es einen weiteren Block aus einer oder mehreren Spalten (KM), in denen für jeden Fall die Werte des Kausalmerkmals (= eine Spalte) bzw. der Kausalmerkmale (= mehrere Spalten) erfasst wurden. Mit der RDA können durchaus mehrere KM gleichzeitig in ihrer Bedeutung und Auswirkung untersucht werden. In der Ökologie ist dies sogar der Standard, da dort oft viele Umweltmerkmale gleichzeitig in ihrer Auswirkung auf Artenzusammensetzungen untersucht werden (vgl. Borcard *et al.* 2018, Kap. 6.3.). Wenn im Folgenden der Einfachheit halber nur noch im Singular von der Ursache gesprochen wird, sind Situationen mit zahlreichen Ursachen dabei mit gemeint. Zur Datentabelle ist festzuhalten, jeder Fall ist einerseits mit den VM und zugleich mit dem oder den KM beschrieben. Nur zur Klarstellung, die oben beschriebenen Transformationen für VM Anzahldaten dürfen nur auf den Block der Datentabelle mit den Anzahldaten angewendet werden, nicht auf die KM – letzteres macht ja auch keinen Sinn.

Die Ursache kann natürlich durch verschiedenartige Merkmalsarten beschrieben werden: nominalskaliert (Kategorien), ordinalskaliert (Kategorien mit interner Rangfolge), intervall- oder ratioskaliert (allg. zur Merkmalsskalierung: Stevens 1946). Allerdings kann der Teil der RDA, der die Kausalität in Rechenschritten umsetzt, die KM entweder nur als Nominalmerkmal oder als Ratiomerkmal verarbeiten (Borcard *et al.* 2018, 206). Das bedeutet, ordinale KM werden methodenintern technisch zu nominalskalierten berabgestuft, und intervallskalierte KM technisch wie ratioskalierte behandelt werden. Durch die Benutzung nominaler KM kann jede archäologisch sinnvolle Gruppierung – denn Gruppen-

zugehörigkeit ist ein Nominalmerkmal – in einer RDA oder einer tb-RDA als Ursache überprüft, untersucht und dargestellt werden. Bei ordinalen KM aber geht die Information über die Rangfolge zwischen den Kategorien verloren. D. h. ein Relativchronologie-Merkmal, etwa die Stufen der Bronzezeit, kann zwar etwa als Ursache für die Veränderung der Zusammensetzung von Horten verwendet werden, aber die in der Stufeninformation enthaltene Rangfolge lässt sich nicht in der Rechnung erfassen. Weil aber die Methode dezidiert zum Aufdecken von Strukturen gebaut ist, kann sich die Rangfolge der Kategorien im Ergebnis widerspiegeln. Es kann aber auch zu Ergebnissen kommen, bei dem zwar die einzelnen Relativchronologiegruppen gut erfasst und beschrieben werden, aber die Abfolge der Relativchronologiestufen in der Ergebnisgrafik nicht in der korrekten Reihenfolge nebeneinanderliegen. Ein derartiges RDA-Ergebnis lässt sich aber trotzdem sinnvoll in Bezug auf die Gruppenunterschiede deuten, wenn es vom Zufall unterscheidbar ist. Die Behandlung etwa des intervallskalierten KM ‚Datierung', angegeben als (ungefähres) Kalenderjahr, durch die RDA-Rechnung als Ratiomerkmal ist tatsächlich problemlos (vgl. für eine CCA mit einem solchen KM ‚Datierung', Gehlen *et al.* 2020). Für eine konsistente Logik bei der Ergebnisgrafikdeutung sollten nur die Jahreszahlen vor Christus als negative, und die Jahreszahlen nach Christus als positive Zahlenwerte in die Datentabelle eingetragen sein.

Die RDA-Varianten können also explizit die Auswirkung der Zeit untersuchen, unbelassen ob diese nun als nominal behandeltes Relativchronologiemerkmal vorliegt, oder als (ungefähre) Jahreszahl. Insbesondere die tb-RDA wird durch ihre Analyse multivariater Anzahldaten als VM zu einem enorm hilfreichen Mittel der historischen Ursachenforschung für eine kausal-deutende Archäologie des 21. Jahrhunderts! So lässt sich etwa eindeutig beweisen, ob, welche, und wie Typen mit der Zeit seltener oder häufiger werden.

Schließlich noch kurz zum technischen Hintergrund: Bei einem ratioskalierten KM kann die RDA die Kausalbeziehung relativ detailliert ausweisen. Hier können etwa bei der tb-RDA graduelle Veränderungen der transformierten Anzahldaten mit schrittweisen Werteänderungen der Ratio-KM verbunden werden, was zu sehr detaillierten Ergebnissen führt. Allerdings kann selbst bei einer solchen vorteilhaften Konstellation die Ursachenauswirkung nur die Form einer ‚Je-Desto-Beziehung' ordentlich erfassen. Oder, statistisch gesprochen, die RDA kann lineare Kausalzusammenhänge erfassen. Gemeint ist mit beiden Ausdrücken ein Kausalzusammenhang etwa in der Art: Je älter ein Hort ist, desto höher ist der Anteil eines bestimmten Gerätetyps. In Anzahldaten auch manchmal enthaltene Information über wachstumsbezogene Anteilsveränderungen, also rechnerisch gesagt exponentielle Effekte, werden nicht als solche, sondern nur mit dem best-angenäherten linearen Kausalzusammenhang erfasst. Bei einem nominalen KM wird die Kausalbeziehung grober erfasst, denn hier gibt es ja nur die Zugehörigkeit zu einer Gruppe als Information zum Unterschied zwischen zwei Fällen und keinen Betrag für den Wertunterschied. Der Effekt des nominalen KM muss also deutlicher als der eines Ratio-KM sein, um beim Testen noch als nicht-zufällig erkannt zu werden. Wie ein nominales KM und ein Ratio-KM rechentechnisch genau behandelt werden, dazu beim Abschnitt über die Regressionsanalyse noch Weiteres.

An dieser Stelle können aber schon zwei Ausblicke eingeflochten werden, die für anspruchsvollere Forschungsfragen interessant sein können. Die Rechentechnik der RDA erlaubt es auch, explizit nach Ähnlichkeit anzuordnen, die nicht von einem (oder mehreren) KM verursacht wurde. Diese Funktionsvariante sei wieder als Handlungsanweisung ausgesprochen: „Partielle RDA, ordne mir die Fälle nach Ähnlichkeit an, aber zeige mir vor allem die Ähnlichkeit, die nichts mit einer bestimmten Ursache zu tun hat, wohl aber mit den anderen Ursachen!" Diese Variante heißt partielle – das ist abteilende – RDA (Legendre und Legendre 2012, 649–652). Abteilen im Sinne von Wegnehmen. Es ist also mit der partiellen RDA beispielsweise möglich, nur die Auswirkung der Kategorie ‚biologisches Alter' auf die Zusammensetzung von merowingerzeitlichen Glasperlengrabbeigaben zu untersuchen, und gleichzeitig explizit alle mit einem relativchronologischen KM verbundene Ähnlichkeit zu entfernen. Gerade das Abteilen von Kausaleffekten findet in den Archäologien als neues Forschungsfragenkonzept noch kaum Beachtung, birgt aber ein enormes Potential für die Verfolgung bisher gar nicht möglicher Fragen!

Der zweite Ausblick nutzt trickreich die Tatsache, dass Rechenwege auch umkehrbar sind, wenn eine Formel einmal entwickelt ist. Für Fälle, bei denen das KM nicht bekannt ist, kann es mithilfe der anderen Fälle und deren RDA-Lösung geschätzt werden (Borcard *et al.* 2018, 232 f)! Ein Beispiel: Für 30 mesolithische Mikrolitheninventare gibt es die nach Typen aufgeschlüsselten Anzahldaten als VM und die Datierung als KM. Die Rechenbeziehung zwischen Zusammensetzung und Datierung kann also anhand der 30 Fälle als Formel erstellt werden. Wenn jetzt zwei weitere Inventare mit bekannten Typenspektren, aber unbekannter Datierung dazukommen, kann die Datierung geschätzt werden, indem der Rechenweg der RDA einfach umgedreht wird (vgl. für eine CCA mit diesem Vorgehen: Gehlen *et al.* 2020). Natürlich ist dies bloß bei klaren Kausalbeziehungen erfolgreich, schließlich würde etwa im Beispiel ein Teil der Formel nur auf einer Zahleninformation (der Datierung) aufbauen. Aber im Prinzip ist ein solches Vorgehen möglich. In der (Paläo-) Ökologie wird es intensiv und erfolgreich zur Rekonstruktion von Umweltmerkmalen anhand von Spektren besonders klimasensitiver Lebewesenarten eingesetzt (Borcard *et al.* 2018, 232). Der Name für diese erweiterte RDA-Variante ist allerdings in den Archäologien

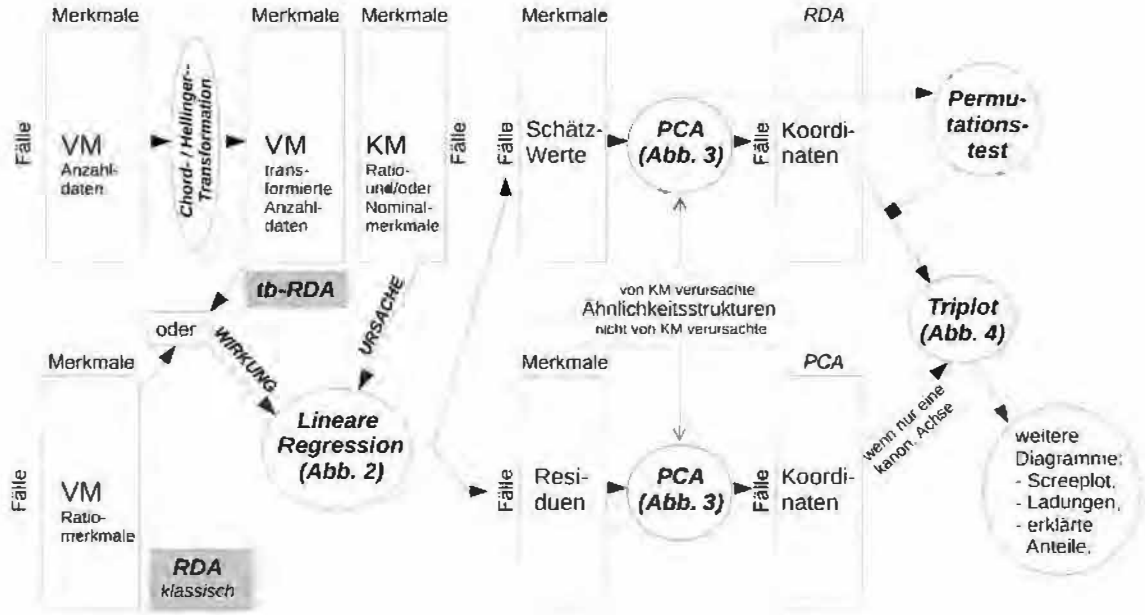


Abb. 1. Flussdiagramm für Ablauf, Schritte, Zwischen- und Endergebnisse einer Redundanzanalyse. In ihrer klassischen Form (Start links unten) untersucht die RDA die Auswirkungen von kausalen Merkmalen (KM) auf die Fallähnlichkeiten bei vielen Ratiomerkmalen (VM), die tb-RDA (Start links oben) dagegen die Auswirkungen dieser Merkmale auf die mittels ausgezählten Typen erfassten Fallzusammensetzungen (transformierte Anzahl-VM). Die verursachte Ähnlichkeit geht in die Schätzwerte aus der linearen Regression (Abb. 2) ein, die nicht kausal ausdrückbaren Werte werden mit den Residuen erfasst. Beide werden einer Hauptkomponentenanalyse (PCA; Abb. 3) unterworfen. Vor der Deutung wird mittels Permutationstest die kausale Wirkung der KM gegen den Zufall abgegrenzt. Schließlich gehen die Fallkoordinaten und die der VM (nicht dargestellt) mit den Beziehungen zu den KM in einen Triplot ein (Abb. 4; nach Fig. 11.1 und 11.2 aus Legendre und Legendre 2012 sowie www.davidzeleny.net/anadat-r/doku.php/en:rda_cca [Stand: 07.05.2021]).

für Missverständnisse anfällig. Denn sie wird im Engl. als ‚*calibration*' bezeichnet, hat aber absolut nichts mit der Kalibration von Radiokarbondaten zu tun und darf nicht damit verwechselt werden – auch wenn, wie im Beispiel oben, ein kalibriertes Radiokarbondatierung als KM benutzt wurde. Damit in Zukunft keine Verwechslungen durch die Nutzung der Bezeichnung ‚*calibration*' aufkommen, sei hier für deutsche archäologische Texte anstatt ‚*calibration using RDA*' die sprechende Bezeichnung ‚RDA basierte Rekonstruktion (RDA-bR)' vorgeschlagen.

5. Dreiteiliger Algorithmus der RDA

Die RDA ist rechentechnisch in zwei Schritte und eine Überprüfung zu untergliedern (Abb. 1). Die Berechnungen gehören beide zu den bestbekannten statistischen Verfahren überhaupt. Denn eine RDA-Berechnung besteht schlicht aus der Aneinanderreihung einer (multiplen) linearen Regression (Schritt a)) und der Hauptkomponentenanalyse der Regressionsschätzwerte (Schritt b)) bzw. -residuen (Borcard *et al.* 2018, 205; Legendre und Legendre 2012, 630). Man könnte sagen, erstere steht für die rechentechnische Umsetzung der Kausalbeziehung, letztere für die – dann ursächlich bedingte – Ähnlichkeitsanordnung. Beide seien hier nochmals kurz in ihrer Funktionsweise als Teile einer RDA zusammengefasst. Der dritte Teil (Prüfung c)) wird bei manchen Software-Umsetzungen getrennt und nur als Option angeboten, dabei steht seine Betrachtung am Anfang der Auseinandersetzung mit einem RDA-Ergebnis. Es handelt sich um die Beurteilung möglicher Zufallseinflüsse auf das Resultat, sprich, einen Test, der nahe legt, ob die RDA gedeutet werden sollte oder ob das Ergebnis zu verwerfen ist.

Die Darstellung hier reproduziert verkürzt und vereinfacht die des ökologischen Standardlehrbuches (Legendre und Legendre 2012, 635–642). Dort sind, wie angemerkt, sämtliche Schritte detailgenau mit ihren Formeln dargestellt.

a) (multiple) lineare Regression (= Ursachenmodellierung)

Für die RDA werden also die VM als sog. Abhängige Merkmale bzw. Regressand(en) einer (multiplen) linearen Regression unterworfen, in der die KM die Rolle der sog. Unabhängigen Merkmale bzw. Regressoren oder Prädiktoren spielen (zur linearen Regression einführend: Fox und Weisberg 2019; vgl. Legendre und Legendre 2012, Kap. 10.3). Dies erfolgt, indem jede Spalte der VM einzeln als Abhängige behandelt wird. Das Prinzip ist anhand einer Beschreibung der linea-

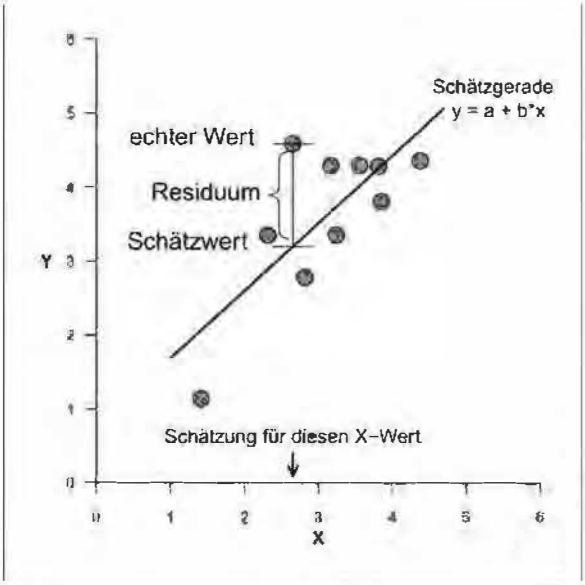


Abb. 2. Prinzip der Schätzung eines verursachten Merkmalswertes Y durch den Wert des verursachenden Merkmals X mithilfe einer linearen Regression. Zu dem mit dem Pfeil bezeichneten Wert des verursachenden Merkmals liefert die zugehörige Y-Position der Schätzgerade den Schätzwert, welcher sich durch den Betrag des Residuums vom echten Wert unterscheidet. Die Schätzgerade wird anhand der Streuung der X- und der Y-Werte ermittelt, aber nur mit den X-Werten ausgedrückt.

ren Einfachregression verständlich (Abb. 2). Gibt es mehr als eine KM wird zwar technisch eine multiple lineare Regression verwendet, zum Verständnis genügt aber bereits der Blick auf die einfache Variante. Die graphische Umsetzung der linearen Regression eines VM und eines KM ähnelt der eines Streudiagrammes zweier Ratiovariablen. Die Fälle erscheinen als Punkte, ihre Merkmalswerte beim KM bilden die X-Koordinaten und ihre Merkmalswerte beim VM die Y-Koordinaten des Streudiagrammes. Die im Diagramm eingezeichnete Regressionsgerade ist so zu lesen, dass sie für jeden X-Wert einen geschätzten Y-Wert anzeigt. Man beachte, dass die eigentliche Kausalbeziehung bzw. die Wahl der vermuteten Ursache *a priori* philosophisch bzw. archäologisch formuliert und begründet werden muss, denn die Regression ist nur die Umsetzung einer Kausalanalyse, sie begründet den Zusammenhang nicht *a priori*, sondern belegt ihn *a posteriori*.

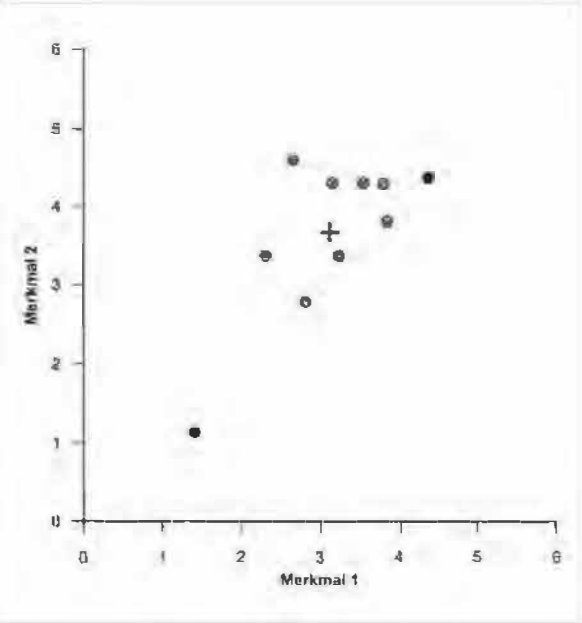
Die Schätzung der Y-Werte baut natürlich auf dem Zusammenhang von VM und KM auf. Allerdings gehen letztendlich nur die X-Werte (KM) der Daten in die Schätzformel (Regressionsgerade) ein. Die Technik dahinter wird verständlich, bedenkt man, dass jede Gerade in 2D allein anhand einer X-Koordinate sowie ihrem Gefälle (ein konstanter Zahlenwert) in Y und dem Y-Startpunkt an der Stelle bei X=0 (weiterer konstanter Zahlenwert) beschrieben werden kann. Liegen die Fallpunkte fast auf einer Geraden, ist der Zusammenhang stark und die Schätzung anhand der Regressionsgeraden gut angepasst. Liegen sie als diffuse Punktwolke vor, die nur einen vagen Zu- oder Abnahmetrend zeigt, ist die Schätzung eher unscharf bzw. ungenau. Wie gut angepasst die Regression also die Fallwerte einer VM schätzt, hängt von der Stärke des Zusammenhangtrends ab. Je stärker dieser ist, desto genauer entspricht der geschätzte Y-Wert eines Falles seinem tatsächlichen Y-Wert. Der Betrag des senkrechten Abstandes (= parallel zur Y-Achse) eines Fallpunktes zur Regressionsgerade ist die graphisch erkennbare Differenz zwischen Schätzwert und echtem Wert. Diese Differenz heißt Residuum (Rest). Es gilt also, Schätzwert plus Residuum gleich echtem Y-Wert. Anders gesagt, die in der Y-Koordinate steckende Information kann zum Teil (Schätzwert) allein anhand der Beziehung (der Regressionsgerade) zwischen KM und VM und der X-Position im Diagramm ausgedrückt werden. Dies ist die sog. erklärte Streuung. Zum Teil (Residuum) allerdings ist die Information über die Y-Position eines Falles nicht so herleitbar. Dies ist die sog. nicht erklärte Streuung oder Residualstreuung. Die Einzelheiten der Regressionen werden i. d. R. nicht in RDA-Ergebnissen ausgeführt.

Wie gesagt wird jedes VM der Daten einzeln einer (multiplen) linearen Regression unterworfen (Legendre und Legendre 2012, 635 f.); multipel ist die Regression, wenn es mehrere KM gibt. Für jeden Fall und jede VM ergeben sich daraus zwei Werte, nämlich der Schätzwert und das Residuum. Schätzwerte und Residuen werden jeweils in einer Tabelle vermerkt, die natürlich die gleichen Ausmaße (Zeilen- und Spaltenzahl) aufweist, wie die Datentabelle der VM. Sind die VM transformierte Anzahldaten, so stehen die Schätzwerte wiederum für transformierte Anzahldaten und eine spätere PCA (s. u.) ‚erhält' die entsprechende multivariate Distanz (Chord- oder Hellinger). Abschließend kann der auf die gesamte Datentabelle der VM bezogene Anteil erklärter Streuung berechnet werden. Dies ist das später mit dem RDA-Ergebnis ausgegebene sog. kanonische R^2 (Legendre und Legendre 2012, 630 f. speziell Formel 11.4), welches heute nur noch selten als ‚bimultivariate Redundanzstatistik' bezeichnet wird (vgl. Borcard *et al.* 2018, 211). Hierin begründet sich wie gesagt der heutige Methodenname. Das R-Quadrat sollte jedoch gerade bei mehr als einem KM nur in seiner korrigierten Form, dem ‚adjustierten kanonischen R^2_{adj}', verwendet werden (zum Grund dafür siehe a. a. O., 212).

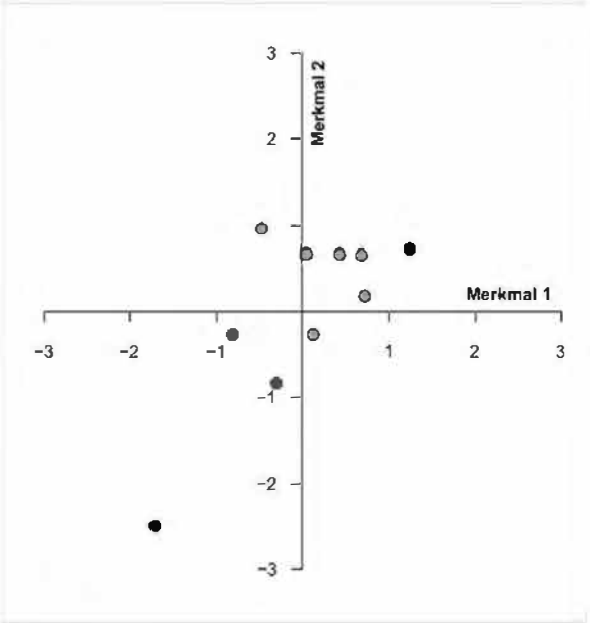
Soweit kann nachvollzogen werden, wie der erste Schritt der RDA abläuft, wenn das KM ratioskaliert ist. Wenn es nominalskaliert ist, wird als Schätzwert für Y der jeweilige Durchschnitt der Y-Werte aller Fälle einer Ausprägung des Nominalmerkmals, also einer ‚Gruppe' verwendet. Gibt es ratio- und nominalskalierte KM, werden die Schätzwerte aus einer Mischung beider Vorgehensweisen erzeugt.

b) Hauptkomponentenanalyse (= Ähnlichkeitsanordnung)

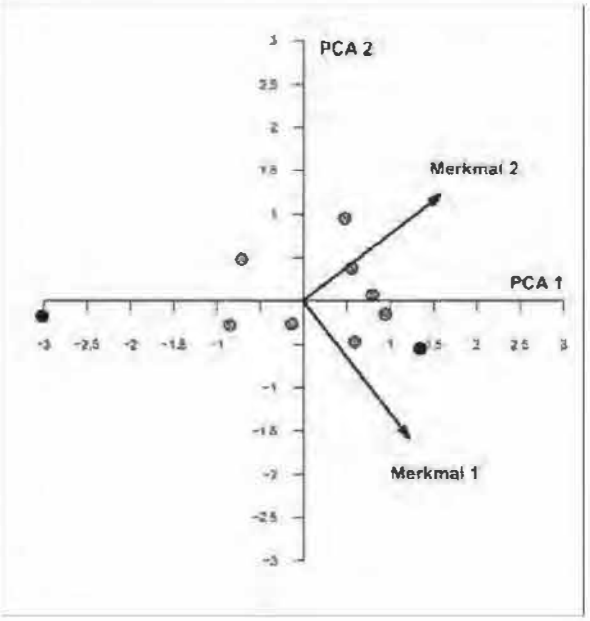
Der zweite Hauptschritt der RDA besteht darin, die Schätzwerte der VM und die Residuen der VM beide jeweils einer Hauptkomponentenanalyse (PCA) zu



a) Streudiagramm zweier ratioskalierter Merkmale mit den Fällen als Punkte. Zwei Extremwerte sind zum besseren Verständnis hervorgehoben.



b) Zentrierung der Merkmale entspricht einer Verschiebung der Punktwolke



c) Drehung der Längsachse des Ensembles aus Fallpunkten und Merkmalsachsen zur optimalen Ansicht

Abb. 3. Prinzip der Hauptkomponentenanalyse nach Legendre und Legendre (2012, 431 Fig. 9.2). Die beiden Extremwerte entlang der Punktwolkenlängsachse sind jeweils schwarz gefüllt. Das Zentrum (schwarzes Kreuz) der Punkte in a) wird zum Ursprung der Achsen in b) und ist Drehpunkt in c).

unterwerfen. Die PCA wurde zuerst von Karl Pearson (1901) als *‚best fitting line'* bzw. *‚plane'* vorgeschlagen. Die mathematische Formalisierung in ihrer heutigen Form stammt von Harold Hotelling (1933a; 1933b). Die PCA selbst ist eigentlich nur eine Verschiebung und Drehung der als Punktwolke (vgl. o.) verstandenen Fälle (Abb. 3; siehe v. a. Wickens 1995, Kap. 9; vgl. Legendre und Legendre 2012, Kap. 9.1 spez. 431 fig. 9.2). Die PCA wurde früher oft mit der für mathematische Laien sperrigen Fachbegrifflichkeit beschrieben, was zu unnötig unverständlichen Darstellungen führte. Hier wird die Funktionalität der PCA geometrisch beschrieben – und das ist keine Metapher, sondern tatsächlich exakt das Prinzip PCA! Die PCA ist nämlich in Wirklichkeit ein einfacher, geometrisch vorstellbarer Vorgang, der allen Archäologinnen und Archäologen, die schon einmal Funde in einem 3D-Grabungssystem ausgewertet haben, leicht eingängig sein sollte.

Oben wurden zum Verständnis der euklidischen Distanz Fälle als Punkte in einem mehrdimensionalen Raum beschrieben. Nimmt man alle Fälle einer Datentabelle zusammen, so ergäbe sich bei dieser Vorstellung eine Punktwolke – oder man denke an einen riesigen Vogelschwarm. Die Merkmale sind dabei durch die Koordinatenachsen dieses Raumes repräsentiert. Uns wiederum reicht es, sich die Punktwolke in 3D vorzustellen, da die sogleich beschriebenen Prinzipien auch auf höherdimensionale Räume verallgemeinerbar sind. Als erstes wird die Punktwolke verschoben, so dass ihr Zentrum auf der Position des Nullpunktes eines neuen Achsensystems zu liegen kommt. Das ist ganz einfach durch Abziehen jeweils des Mittelwertes einer Spalte von allen Werten der zugehörigen Spalte erreichbar. Das Zentrum der Punktwolke liegt jetzt im neuen Koordinatensystem auf dem Achsenursprung (Null-Koordinaten). Jetzt wird die Punktwolke um dieses Zentrum so gedreht, dass ihre Längsachse zur ersten Achse (1. Hauptkomponente) des neuen Achsensystems der Ansicht nach der Drehung wird. Ihre Querachse wird zur zweiten Achse (2. Hauptkomponente) des neuen Achsensystems, ihre Tiefenachse zur dritten

Achse usw. (vgl. Borcard *et al.* 2018, 152). Die Achsen des neuen Koordinatensystems nach der Drehung werden auch als Hauptkomponenten bezeichnet und manchmal selbst im Deutschen nummeriert als PCA 1, PCA 2 usw.

Warum aber wird die Punktwolke so gedreht? Weil entlang der Punktwolkenlängsachse die größten Abstände zwischen den Punkten bzw. die größten Merkmalsunterschiede zwischen den Fällen existieren! Wird also entsprechend obiger Anweisung gedreht, zeigt ein Streudiagramm, das als Koordinatenachsen die ersten beiden neuen Achsen (nach der Drehung) verwendet, die 2D-Ansicht mit dem höchsten Informationsgehalt, nämlich mit den größten (Punktwolken-Längsachse) und den zweitgrößten (Punktwolken-Breitenachse) Unterschieden zwischen den Fällen. Die üblichen PCA-Ergebnisgrafiken zeigen die Fälle als Punkte nach dem Verschieben und Drehen auf den ersten beiden Achsen des neuen Systems. Aber nur weil die Punktwolke gedreht wurde, gibt es immer noch die gleiche Anzahl von Dimensionen. Sie werden nur meist nicht mehr beachtet, da sie deutlich weniger Information beinhalten als die ersten beiden bzw. ersten paar. Man denke an den Fisch Flunder: Die Ansicht von der Seite, in der die Gestalt der Flunder nur eine schmale Silhouette ist, informiert wenig über das eigentliche Aussehen der Flunder, aber die Aufsicht. Eine PCA einer flundergestaltigen Punktwolke würde diese also so drehen, dass die Aufsicht angezeigt wird. Die Seitenansicht wäre noch vorhanden, interessierte aber kaum.

Jetzt stelle man sich nicht nur die Punktwolke, sondern auch noch Merkmale (VM), die die Punkte beschreiben, als quasi die Punktwolke rahmende Achsen vor. Man könnte auch an ein Mobile aus Punkten (Fälle) und Stäben (Merkmale) denken. Dreht man dieses gesamte Ensemble, dreht man auch diese (alten) Achsen in das neue System. In die PCA-Ergebnisgrafik lassen sich also auch die Merkmale als Pfeile einzeichnen. Sie erlauben es zu bestimmen, welche Position im neuen System welchen Werten auf den alten Achsen entspricht. Die Ergebnisgrafik einer PCA heißt Biplot, wobei die Vorsilbe Bi- auf die Darstellung zweier Entitäten, der Fälle und der alten Merkmalsachsen, verweist (zur Deutung s. u.). Weil bei einer solchen Ergebnisgrafik die ähnlichen Fälle nach wie vor nahe beieinander abgebildet werden – eine Drehung verändert ja die Abstände nicht (!) – zählt die PCA (und entsprechend die RDA) zu den Verfahren der multivariaten Ordination (Ähnlichkeitsanordnung). Die Deutung dieser RDA-Ergebnisabbildung wird später im Detail erläutert (s. u. Triplot).

Der Autor legt allen akademischen Lehrerinnen und Lehrern multivariater Methoden in der Archäologie wärmstens an Herz, die Bezeichnung Ordination zur Methodenunterteilung zu nutzen, da sie Lernenden nach eigener Erfahrung ein besser strukturiertes Verständnis des gesamten so bezeichneten Methodenfeldes erlaubt.

Jetzt seien noch technische Aspekte ergänzt, um den Anschluss an die Sprache der Lehrbücher zu halten. Um die Punktwolke zu drehen, bedarf es einer sog. Rotationsmatrix, die angibt in welche Richtung wie stark gedreht wird. Diese Rotationsmatrix wird aus den Maßen für die Streuung abgeleitet, die jeweils zwei Merkmalen gemeinsam ist (Kovarianz bzw. Korrelation). Gemeinsame Streuung ist nur eine bestimmte Ausdruckweise dessen, dass zwei Merkmale die Fälle in mehr oder weniger gleicher Weise beschreiben. Oder nochmals anders formuliert: je größer die gemeinsame Streuung, desto ähnlicher die Abfolge der Fallwerte bei beiden Merkmalen. Es steckt also Information über die Fallähnlichkeiten in einer Streuungsmatrix, was an der möglichen Berechnung der Streuungsmatrix aus den Mittelwertsdifferenzen der Fälle erkennbar ist (Legendre und Legendre 2012, 147). Bei einer RDA sind die für die gemeinsame Streuung benutzten Merkmale nur die VM. Wenn diese Merkmale alle die gleiche Maßeinheit (oder keine Maßeinheiten) aufweisen, wird als Streuungsmaß die sog. Kovarianz verwendet (a. a. O., 445). Sie wird für alle Merkmalspaare ermittelt und in eine quadratische Tabelle (Streuungs-Matrix) geschrieben. Auf der Diagonalen steht die Streuung eines Merkmals um seinen Mittelwert, ausgedrückt als sog. Varianz, jeweils links unterhalb und rechts oberhalb stehen die paarweisen gemeinsamen Streuungen (Kovarianzen). Diese Streuungsmatrix heißt (Varianz-/)Kovarianzmatrix. Wenn die Merkmale dagegen in verschiedenen Maßeinheiten ausgedrückt sind, sind sie vorher einer Z-Transformation zu unterwerfen (vgl. o.). Die Kovarianz z-transformierter Ratiomerkmale heißt Korrelation (a. a. O., 152). Die gerade beschriebene Streuungsmatrix heißt dann Korrelationsmatrix. Sie weist auf der Diagonalen nur Einsen auf, da die Korrelation (gemeinsame Streuung) eines Merkmals mit sich selbst den maximalen Korrelationswert 1 ergibt. Wenn es sich um Chord- oder hellinger-transformierte Anzahlen handelt, darf keine Z-Transformation eingesetzt werden, da sonst Informationen über die Rolle häufiger und seltener VM verloren gehen (a. a. O. 634). Außerdem verkompliziert dies die Deutung der Abstände in der PCA und der RDA als multivariate Distanzmaße für Anzahldaten.

Die Information der Streuungsmatrix kann man auch so verstehen, dass sie angibt in welche Richtungen – immer definiert als Paar zweier Merkmale – sich die Punktwolke vor allem erstreckt bzw. eine langgestreckte Gestalt aufweist und in welche Richtungen sie eher diffus bzw. kugelig ausgebildet ist, also keine ‚echte' Längserstreckung besitzt. Die Rechnung zur Bestimmung der Rotationsmatrix anhand der Streuungsmatrix heißt Eigenwertzerlegung und ist ein Polynom p-ten Grades. Dabei ist p meist gleich der Anzahl der Merkmale (VM) und selten kleiner. Ihr Ergebnis ist die Rotationsmatrix. Diese Rotationsmatrix heißt technisch Eigenvektormatrix oder seltener Ladungsmatrix. Die Matrix der zentrierten Werte (= auf Nullkoordinaten verschobene Tabelle) wird mit ihr postmultiplizert (siehe dazu eine Einführung zur Matrix-Algebra

etwa bei Legendre und Legendre 2012, Kap. 2), was de facto einer Drehung entspricht. Das Ergebnis ist die zur optimalen Ansicht gedrehte Punktwolke. Die Eigenvektoren selbst werden in der gebräuchlichen Variante der PCA-Ergebnisgrafik (s. u., Distanz-Biplot bzw. Skalierung I; Legendre und Legendre 2012, 443 f.) als Positionen für die Pfeilspitzen benutzt, die die Lage der alten Merkmale im neuen System bezeichnen. Durch die Eigenwertzerlegung wurden die Informationen über die Fallunterschiede, die ja in den gemeinsamen Streuungen stecken und damit die Streuungsmatrix bilden, aufgeteilt und mit neuen Werten, den VM-Koordinaten im neuen System, dargestellt. Damit nachvollziehbar ist, wie viel Information über die ursprünglichen Fallunterschiede (Streuungen) im neuen System von den jeweiligen Hauptkomponenten (neuen Achsen) repräsentiert wird, gibt es als Maßzahl für jede neue Achse einen sog. Eigenwert. Zu jeder Hauptkomponente (neuen Achse nach Drehung) bzw. jeder Spalte der Eigenvektor-Matrix gehört solch ein Eigenwert. Bei einem korrelationsbasierten PCA-Schritt ist die Summe der Eigenwerte i. d. R. gleich der Spaltenanzahl der Datentabelle, deren Streuungsmatrix analysiert wurde (a. a. O. 94 und 432). Bei einem kovarianzbasierten PCA-Schritt entspricht die Eigenwertsumme der Summe der Diagonalwerte der Streuungsmatrix. Nach dem oben beschriebenen Vorgehen des Drehens der Punktwolke in ihre Längsrichtung ist leicht verständlich, dass der erste Eigenwert, also der Eigenwert der ersten Hauptkomponente, der größte sein muss. Denn er ist ja auch ein Maß für Unterschiede entlang ‚seiner' Hauptkomponente. Der Eigenwert der zweiten Hauptkomponente muss der zweitgrößte sein usw. Weil aber der Zahlenwert eines Eigenwerts von den konkreten Merkmalswerten abhängt, wird i. d. R. stets der Anteil eines Eigenwertes an der Summe aller Eigenwerte angegeben. Verständlicherweise weist ein erfolgreicheres PCA-Ergebnis also für wenige erste Achsen sehr hohe Informationsanteile auf und niedrige für alle weiteren. Die erste Abbildung einer PCA wie einer RDA ist denn auch stets ein Säulendiagramm – der sog. *screeplot* –, das die Zahlenbeträge oder Anteilswerte der Eigenwerte pro neuer Achse darstellt, und somit die Güte der Informationszusammenfassung einer multi-D-Tabelle durch eine 2D-Ansicht (Biplot bzw. Triplot, s. u.) beurteilbar macht (Borcard *et al.* 2018, 157 f.). Wenn die Zahlenwertbeträge der Eigenwerte als Säulenhöhen benutzt werden, steckt die Anteilsinformation in den Größenverhältnissen der Säulen zueinander. Bei einer RDA wird zugleich visuell beurteilbar, wie viel Ähnlichkeit (auf welchen Achsen) als kausal darstellbar ist und wie viel nicht.

In der einfachen alten Definition des RDA-Algorithmus wird nur die Tabelle der Regressions-Schätzwerte einer PCA unterworfen. Die nach dem Verschieben und Drehen definierten neuen Achsen heißen RDA 1 usw. Dabei passiert allerdings etwas auf den ersten Blick Verstörendes. Wenn in der RDA nur ein ratioskaliertes KM verwendet wird, hat der RDA-Ergebnisraum nur eine neue Achse! Bei genauerem Nachdenken wird aber klar, dass solch ein Effekt durch den Regressionsschritt erzwungen wurde. Denn dadurch enthalten die Schätzwerte in allen Spalten der VM-Tabelle doch immer nur die Information, die in dem einen Ratio-KM enthalten ist. Wenn als Ausgang also nur ein Ratio-KM (eine Merkmalsachse) diente, kann das Ergebnis ebenfalls nicht mehr als eine Achse/Dimension aufweisen! Bei mehreren ratioskalierten KM besitzt das RDA-Ergebnis dagegen maximal so viele neue Achsen, wie es KM gibt (zur Anzahl von RDA-Achsen siehe Borcard *et al.* 2018, 206). Wenn diese aber wiederum in Abwandlungen sehr ähnliche Informationen enthalten, können sich auch weniger RDA-Achsen ergeben. Bei einem nominalskalierten KM hängt die Zahl der RDA-Achsen von der Anzahl der Nominalausprägungen – der Anzahl der Kategorien des Merkmals oder sage, der Gruppen – ab. Um zwei Gruppen auseinander zu halten, bedarf es einer Achse, entlang derer man sie scheidet. Bei drei Gruppen braucht es schon zwei Achsen, eine, um die ersten beiden zu unterscheiden und eine, um die dritte von den ersten beiden zu unterscheiden. Bei k Gruppen werden also k-1 Achsen nötig. Bei mehreren gemischten KM ergibt sich die Zahl der RDA-Achsen entsprechend aus der gemeinsamen Anwendung beider Prinzipien. Unabhängig davon, wie viele RDA-Achsen entstehen, immer bilden stets die ersten beiden die informationsreichste Ansicht für die Kausalbeziehung. In der älteren RDA-Definition ist damit der Anordnungsvorgang abgeschlossen. Wenn nur ein (Ratio)KM vorliegt, ‚degeneriert' der Ergebnisraum auf 1D; deshalb erlaubt die Software PAST (Hammer *et al.* 2001) etwa keine RDA mit nur einem KM – obwohl das rechentechnisch lösbar wäre.

In der mittlerweile üblichen neueren Vorgehensweise, die vor allem durch die RDA-Umsetzung im Erweiterungspaket ‚vegan' (Funktion ‚rda'; Oksanen *et al.* 2020) der statistischen Programmieroberfläche R (R Core Team 2020) inspiriert ist, wird auch die Tabelle der Residuen der VM einer PCA unterworfen (Borcard *et al.* 2018, 205). In gleicher Weise wird die Rotationsmatrix ermittelt und die Punktwolke, deren Koordinaten jetzt nur die nicht auf das KM zurückgehende Informationen enthalten – das sind die Residuen – wird wiederum gedreht. Und wieder ergeben sich Koordinaten für die Fälle und die alten Merkmale (Eigenvektoren) im neuen System. Man könnte nun sagen, die PCA der Schätzwerte beschreibt den Teil des Gesamtraumes der Ähnlichkeit zwischen den Fällen, der auf die KM zurückgeht, während die PCA der Residuen den Teil des Gesamtraumes der Ähnlichkeit zwischen den Fällen beschreibt, der nicht auf die KM zurückgeht. Es gibt also bei modernen RDA-Implementierungen zwei Ergebniskoordinatentabellen (und zwei Eigenvektortabellen), die aber als gemeinsame Tabelle aus den Blöcken RDA zuerst und dann PCA ausgegeben werden. Die ersten Spalten von links bilden RDA 1, RDA 2 usw. danach kommen die Spalten PCA 1, PCA 2 usw. Denn beide Teilräume zusammengenommen ergeben den Gesamtraum der Ähnlichkeit zwischen den Fällen. Weil bei

einer RDA vor allem die von der vermuteten Ursache bedingte Ähnlichkeit interessiert, werden die aus den Schätzwerten gewonnenen RDA-Koordinaten, die Achsen RDA 1 usw., als erster Block in die zusammengefügte Koordinatentabelle aufgenommen und die aus den Residuen gewonnenen PCA-Koordinaten als zweiter Block. Analog wird mit der aus den beiden Eigenvektoren-Matrizen zusammengesetzten Tabelle verfahren. Es ergeben sich so bei modernen RDA-Implementierungen zwei aus RDA- und PCA-Teil zusammengesetzte Koordinaten-Tabellen, einmal für die Fälle und einmal für die VM.

Die Wertveränderungen bei den KM im neuen System müssen nicht unbedingt parallel zu den RDA Achsen verlaufen! Gibt es nur ein Ratio-KM, dann zeigt der Pfeil für das KM in die gleiche Richtung wie die Achse RDA 1. Bei mehreren KM können sie, als Pfeile repräsentiert, in der Ergebnisgrafik in verschiedene Richtungen zeigen (s. u.). Praktisch werden sie anhand der Beziehungen (Korrelation) zwischen den RDA-Koordinatenwerten der Fälle und den KM-Werten der Fälle berechnet (Legendre und Legendre 2012, 639 Formel 11.20). Wenn Nominal-KM vorliegen, wird der Mittelpunkt aller Fälle einer Gruppe (einer Kategorie), der sog. Zentroid, als Durchschnittskoordinatenposition berechnet und als großer Punkt abgebildet, nicht als Pfeil. So kann bei der Deutung die Art des KM besser berücksichtigt werden. Da es bei der Information eines Nominal-KM keine kontinuierlichen Wertänderungen gibt, sollte er nicht als ein wie eine Achse deutbarer Pfeil, sondern eben als feste Position, als ein Punkt erscheinen. Zentroidenpositionen ergeben sich wie gesagt aus den Durchschnittswerten aller Fälle einer Gruppe auf den RDA-Achsen.

c) Ausschluss von Zufallseinflüssen (Permutationstest)

„Testing is the central step of inferential statistics. It allows one to generalize the conclusions of statistical estimation to the reference population from which the observations have been drawn and that they are supposed to represent" (Legendre und Legendre 2012, 22). Denn bei allen wissenschaftlichen Beobachtungen, die nur auf Teilen (Stichprobe) eines interessierenden Phänomens basieren, gibt es immer eine zunächst nicht auflösbare Ungewissheit darüber, ob die an diesem Ausschnitt erkennbaren Aspekte auf das gesamte Phänomen (Population) verallgemeinert werden dürfen (Hedderich und Sachs 2020, 9 und 21 f.).

Im Fall einer RDA besteht die Stichprobe aus den Fällen, die multivariat beschrieben und analysiert werden. Dass die Fälle für ein übergeordnetes, homogenes Phänomen stehen, ergibt sich schon durch eine sinnvolle Forschungsfrage. Homogen bedeutet hier, die Fälle können alle gemeinsam als gleichartige archäologische, zum Forschungsgegenstand passende, Quellen angesehen werden. Die Daten sind also vor Beginn der Analyse daraufhin zu prüfen, ob sie sinnvoll gemeinsam ausgewertet werden können. Inhomogenität sei mit einem Gegenbeispiel verdeutlicht: Aspekte der Entwicklung spätlatènezeitlicher Geldwirtschaft in Mitteleuropa sind zielführend anhand von spätlatènezeitlichen Münzhorten untersuchbar; ist (unglücklicherweise) ein Münzhort der späten römischen Kaiserzeit in den zu analysierenden Daten enthalten, sind sie inhomogen und ihre Zusammenstellung sollte überarbeitet werden. Homogenität hängt dabei von der Fragestellung ab. Wenn beispielsweise jungsteinzeitliche Haustiernutzung untersucht wird, sollten nur die Anzahldaten der Haustierarten vorhanden sein. Ist dagegen die Tiernutzung von Interesse, können auch Wildtierarten-Anzahlen enthalten sein. In Bezug auf die Datenqualität ist noch anzumerken, dass gerade für multivariate Anzahldaten einer tb-RDA gelten muss: Die Typenbestimmung bzw. die Auszählung nach Kategorien erfolgte durch Personen mit möglichst guter Materialkenntnis. Es macht die anspruchsvollste Analyse hinfällig, wenn die Bestimmungssicherheit des Materials berechtigterweise angezweifelt werden kann. Für alle Arten quantitativ untersuchbarer Phänomene gilt: ohne gute Daten keine sinnvolle Analyse. Natürlich gilt auch: ohne quantitative Analyse keine sinnvolle moderne Deutung der Daten.

Der Test einer RDA prüft, ob die Effekte der KM auf die VM überzeugend vom Zufall unterscheidbar sind, also ob eine RDA deutungswürdig ist oder ob die Beobachtungen auch auf zufällige Einflüsse zurückgehen können. Da die Struktur des Zufalls bei einem multivariat erfassten Kausalbeziehungsphänomen nicht bekannt ist, können klassische Zufallsvariablenverteilungen nicht angewendet werden (Borcard *et al.* 2018, 219). Stattdessen kommt ein sog. Permutationstest zum Einsatz (Hedderich und Sachs 2020, 599–601; Legendre und Legendre 2012, 25–28). Dabei wird als Teststatistik ein sog. Pseudo-F-Wert als Bruch aus erklärter und nicht-erklärter Streuung berechnet (a. a. O., 634 Formel 11.7) – hierbei geht das ‚kanonische R^{2}' in die Formel ein. Die als Permutations-P-Wert bezeichnete Chance, dass die vorliegenden Auswirkungen, ausgedrückt mit dem Pseudo-F, auf Zufallseinflüssen beruhen könnten, wird folgendermaßen berechnet (Legendre *et al.* 2011, 272). Zuerst wird die RDA der echten Daten gerechnet und ein erster Pseudo-F-Wert bestimmt. Dann beginnt das Permutationsverfahren – Permutation bedeutet hier einfach Vertauschen bzw. Neu-Kombinieren (vgl. Hedderich und Sachs 2020, 61 f.). Und zwar werden die KM zufällig in den Zeilen vertauscht. Nach einer Vertauschung (Permutation) sind sie also nicht mehr mit den Fällen verbunden, für die sie erfasst wurden, sondern stehen in anderen Zeilen, also an anderer, ‚falscher' Stelle. Der datenphilosophische Aspekt dabei ist, durch eine Vertauschung wird der Zusammenhang zwischen VM und KM zufällig. Jetzt wird anhand dieses simulierten Zufalls wiederum ein Pseudo-F-Wert berechnet. Das Vertauschen und erneute Pseudo-F-Berechnen wird jetzt viele Male wiederholt, etwa 999- oder

9999-mal. Nimmt man die eigentliche Analyse hinzu ergeben sich so 1000 oder 10000 Pseudo-F-Werte. Jetzt lässt sich bestimmen, mit welchem Anteil Pseudo-F-Werte auftreten, die gleich groß oder noch größer als das tatsächliche Pseudo-F sind. Dieser Anteil ist der oben erwähnte Permutations-P-Wert. Er misst quasi, wie hoch die Chance ist, dass bei Zufall eine gleiche oder bessere Erklärung der Daten auftreten kann. Ist der Permutations-P-Wert kleiner als fünf Prozent ($p \leq 0{,}05$), bedeutet dies, die Chance, dass der Zufall das beobachtete RDA-Ergebnis hätte erzeugen können, ist kleiner als 1-zu-19 bzw. 1 in 20. Wenn etwa bei 999 RDA-en der permutierten Daten und der einen realen Analyse, also 1000 Pseudo-F-Werten insgesamt, nur 49 Werte davon größer oder gleich dem Pseudo-F-Wert der realen Analyse sind, kann das Ergebnis schon als nicht-zufällig und daher deutungswürdig beurteilt werden.

Wie bei anderen Testen auch gilt, schwache Effekte kann ein Test erst bei einer großen Stichprobe vom Zufall unterscheiden, starke Effekte schon bei kleineren Stichproben. Die Größenordnungen, was schwache und starke Effekte bzw. was kleine und große Stichproben sind, lässt sich nicht genau beziffern. Aus der Praxis empfiehlt der Autor bei einem Ratio-KM mindestens 20 Fälle und bei einem Nominal-KM mit zwei Kategorien wiederum mindestens 20 Fälle mit jeweils mindestens 10 Fällen pro Gruppe/Kategorie. Nominal-KM sollten grundsätzlich immer mindestens 10 Fälle pro Gruppe/Kategorie aufweisen. Mit Stichproben solchen Umfangs können Permutationsteste bereits stärkere Effekte (> 20 Prozent erklärte Streuung) erkennen. Bei einem oder wenigen Ratio-KM, einem Nominal-KM mit wenigen Kategorien, oder einer kleinen gemischten KM-Tabelle sind Anteile von weniger als 10 Prozent kausal erklärter Streuung nicht selten (kanonisches $R^2 < 0.1$). Es empfiehlt sich daher eine größere zweistellige Fallzahl, am besten aber mehr als einhundert Fälle.

Zunächst ist immer vor jeder weiteren Darstellung der RDA diese Grundprüfung des Gesamtergebnisses durchzuführen. Gibt es aber mehr als ein KM, werden sogar bereichsweise Teste bzw. Teste von Teilen des RDA-Ergebnisses möglich (siehe die Argumente bei Funktion ‚anova.cca' in R-Paket ‚vegan'; Oksanen *et al.* 2020). Dabei ist es möglich, ein KM getrennt einzeln (engl. *by term*) oder einzeln, aber unter Betrachtung aller anderen KM (engl. *by margin*), auf die Nichtzufälligkeit seines Einflusses zu testen. Die englischen Ausdrücke sind die Begrifflichkeiten der Funktion ‚anova.cca' (*ibid.*). Bei mehr als einem KM sollten auf diese Weise alle ‚Einzelteile' der RDA überprüft werden. Da bei einer RDA mit mehreren Ratio-KM oder einem Mehr-Kategorien-Nominal-KM deren Effekte getrennt auf mehrere RDA-Achsen verteilt werden können, ist es außerdem sinnvoll, auch die einzelnen RDA-Achsen getrennt zu testen (engl. *by axis*). Jede RDA-Achse repräsentiert schließlich einen von allen anderen Achsen jeweils getrennten und unabhängigen Teileffekt, der also auch einzeln interpretiert werden kann. Um sicher zu sein, keine Zufallseinflüsse zu deuten, sollten vor einer Einzeldeutung von RDA-Achsen also stets auch diese Achsen einzeln getestet worden sein. Dabei kommt es nicht zu selten vor, dass zwar der Test das Gesamtergebnis als nicht-zufällig beurteilt, aber die letzten paar RDA-Achsen durch Einzelteste als nicht vom Zufall scheidbar ausgewiesen werden. Dies ist kein Widerspruch, sondern nur eine Auswirkung des ‚genaueren Hinsehens' aufgrund der Einzelteste. Angemerkt sei noch, dass die Technik der bereichsweisen Teste wesentlich komplizierter aufgebaut ist, als der oben skizzierte Permutationstest des Gesamtergebnisses und deshalb hier nicht einzeln erläutert wird. Die Details dazu finden sich in einem Artikel der führenden Methodenentwickler Legendre, Oksanen und ter Braak (Legendre *et al.* 2011).

Zum Umgang mit einer erfolgreich getesteten RDA sei nochmals angemerkt: Weist der Test das RDA-Ergebnis als nicht-zufällig aus, dann kann die Deutung der RDA auf das Ganze, durch die Fälle ja nur ausschnittweise repräsentierte, Phänomen erweitert werden. Wenn etwa die Entfernung von der Meeresküste die Tierarten-Zusammensetzungen bei einigen Dutzend Jäger:innen-Sammler:innen-Fundstellen mit verursacht, dann gilt dieser Effekt nicht nur für diese Stichprobenfälle, sondern grundsätzlich für diese Jäger:innen-Sammler:innen-Gesellschaft in dieser Zeit und in dieser Region. Die RDA erlaubt es durch ihre Koppelung mit einem Test also, multivariate (Anzahl-)Datensätze in einer bisher nicht-möglichen Weise kausal deutend auszuwerten und stellt – zusammen mit den anderen kanonischen Ordinationsverfahren – die einzige Möglichkeit dar, Deutungen zu Anzahl-Zusammensetzungsdaten replizierbar zu erzeugen und objektiv historisch zu verallgemeinern. Damit wird für diese Daten auch die jüngst geforderte Replikation von Analysen (Marwick 2017), also das Wiedererzeugen einer Einsicht mittels gleicher Methoden und Merkmalsdefinitionen aber anhand anderer konkreter Datenwerte, möglich. Dies ist traditionellen, nicht-quantitativ arbeitenden Ansätzen methodologisch schlicht unmöglich – was bedauerlicherweise bei den besonders das Deutende heraushebenden theoretischen Ansätzen der letzten Jahrzehnte ignoriert wird (Schulte 2020). Unsere gesellschaftliche Gegenwart der Jahre 2020 und 2021 sollte hier aber allen ein Lehrstück dafür gewesen sein, wie unverzichtbar richtig angewendete quantitative Verfahren beim Verallgemeinern von Aussagen über eine komplizierte Wirklichkeit sind. Wie viel mehr muss das erst für Verallgemeinerungen von Aussagen über die nur in ausschnitthaften Informationen erfassbare Vergangenheit gelten.

6. Ergebnisgrafik (Triplot)

Das Ergebnis einer RDA lässt sich eleganterweise in nur einer sehr informationsreichen und gleichzeitig anschaulichen Ergebnisgrafik zusammenfassen. Dieses Diagramm wird als Triplot bezeichnet, weil drei Entitäten mit ihren gegenseitigen Beziehungen gemeinsam dargestellt werden. Ein Triplot enthält nämlich zugleich

die ursächlich bedingte Ähnlichkeitsanordnung der Fälle, die Kausaleffekte bei den verursachten Merkmalen (VM) und die Beziehungen dieser Effekte (bei Fällen und VM) zu dem oder den KM. Da eine solche Ergebnisgrafik nach etwas Einüben der Triplot-Interpretationsregeln auf einen Blick einen enormen Informationsgehalt vermitteln kann, gehören solche Diagramme zu den informativsten Illustrationen in den Wissenschaften überhaupt, drücken sie doch komplexe Kausalbeziehungen schlicht als Abstände zwischen Punkten, Winkeln zwischen Achsen und Pfeilen, sowie Punktprojektionen auf die letzteren (s. u.) aus. Bevor hier diese Triplot-Deutungsregeln erläutert werden, ist allerdings noch einmal auf zwei technische Aspekte der Fallkoordinaten im Ergebnisraum der RDA einzugehen. Es sei hier aber nochmals daran erinnert, so spannend ein RDA-Ergebnis auch scheinen mag, wenn der Test des Gesamtergebnisses den Zufall nicht ausschließen kann, soll und darf außer der Mitteilung über das negative Testergebnis kein weiteres RDA-Ergebnis vorgelegt werden.

a) Grundlage der Fall-Darstellung im RDA-Ergebnis (Koordinaten-Varianten)

Der erste Aspekt betrifft die Art, wie die Fallkoordinaten für den Triplot berechnet werden. Tatsächlich gibt es nämlich zwei Varianten, diese Koordinaten zu berechnen (Borcard *et al.* 2018, 205; Legendre und Legendre 2012, 638), die sich quasi in ‚datenphilosophischer' Hinsicht unterscheiden. Dieser Unterschied ließe sich auch mit dem Sprichwort ‚Man sieht nur, was man sehen will' umschreiben. Denn es geht um eine grundsätzliche Frage, welcher Informationsaspekt der Fälle im Ergebnisdiagramm erscheinen soll.

Die erste Variante wurde oben beim PCA-Schritt schon beschrieben. Noch einmal in technischen Worten ausgedrückt: Die Fallkoordinaten für den Triplot werden dort mittels einer Postmultiplikation der zentrierten VM-Tabelle (zentrierte Daten-Matrix) mit der Matrix der Eigenvektoren der Streuungsmatrix der Regressionsschätzwerte erzeugt. Das wichtige für den hier behandelten Aspekt ist aber, dass die beteiligte Datentabelle (zentrierte VM-Daten-Matrix) noch die gesamte Information über die Unterschiede bzw. Ähnlichkeit der Fälle enthält. Die Drehung wurde aber einzig anhand der kausal beeinflussten Unterschiede (Eigenvektoren der Schätzwert-Matrix) definiert. Etwas abstrakter gesagt, der Ähnlichkeitsraum wurde anhand der Schätzwerte bestimmt, die Punkte, die in diesen Raum ‚hineingedreht' werden, enthalten aber auch Information, die mit der Raumdefinition nichts zu tun hatte. Wieder in technischen Begriffen: Die Eigenvektoren der Regressionsschätzwerte-Streuungsmatrix definieren einen Ergebnisraum, der nur die kausal bedingte Ähnlichkeit enthält. In diesen Raum wird aber die gesamte Ähnlichkeitsinformation der Daten (der Originaltabelle der VM) eingebettet. Die sich so ergebenden RDA-Fallkoordinaten enthalten also auch noch Informationsanteile, die nicht strikt im Sinne der RDA-Logik rein als verursachte Ähnlichkeit zu bezeichnen sind. Und diese Informationsteile bestehen in dieser Variante der Fallkoordinaten fort und prägen so teilweise die Fallanordnung. Die Positionen der Fallpunkte sind also bei diesem Vorgehen nicht im strengen Sinne ‚nur' Repräsentation der verursachten Ähnlichkeit.

Die zweite Variante der Koordinatenerzeugung bedient sich dagegen nur der Regressionsschätzwerte der VM. Wie bei der ersten Variante werden wiederum aus der Streuungsmatrix der Regressionsschätzwerte die Eigenvektoren – also die ‚Drehanweisung' für die Punktwolke – berechnet. Aber gedreht wird jetzt eine Punktwolke, deren Koordinaten durch die Regressionsschätzwerte der VM gegeben sind. Diese Zahlen enthalten also nur den verursachten Teil der Ähnlichkeitsinformation. Ihre Deutung kann sich auf die strikte RDA-Philosophie berufen. Die Positionen der Fallpunkte sind im strengen Sinne ‚nur' Repräsentation der verursachten Ähnlichkeit.

In der Zusammenfassungsfunktion des R-Paketes ‚vegan' für das Ergebnis einer RDA wird die erste Koordinaten-Variante als ‚*Site scores* (*weighted sums of site scores*)' bezeichnet und als ‚wa' für *weighted averages* abgekürzt (Borcard *et al.* 2018, 205). Jari Oksanen selbst (ders. 2020, 4–5) bezeichnet sie als ‚WA' oder *Weighted Averages Scores*. Folgerichtig heißt die zweite Variante in der RDA-Zusammenfassungsfunktion ‚*Site constraints* (*linear combinations of constraining variables*)' und ihre Abkürzung lautet ‚lc'. Oksanen selbst wiederum benutzt die Bezeichnungen ‚LC' oder ‚*Linear Combination Scores*'. Während in der Begrifflichkeit für die zweite Variante die Worte ‚linear' und ‚Kombination' noch auf den Regressionsschritt der RDA verweisen, geht die Bezeichnung ‚gewichtete Mittel' forschungsgeschichtlich darauf zurück, dass Oksanen die RDA-Funktion von ‚vegan' entsprechend der CCA-Funktion aufbaute und von dort – leider – auch diese Bezeichnungen übernahm und nicht an die RDA anpasste. In der kommerziellen Software CANOCO wird die zweite Variante mit Verweis auf den Regressionsschritt als ‚*sample scores that are linear combinations of environmental variables*' bezeichnet (ter Braak und Smilauer 2012). In der freien Software PAST (Hammer *et al.* 2001) werden sie als ‚*fitted site scores*' bezeichnet.

Hier seien im Deutschen einmal die zugegebenermaßen etwas sperrigen, aber dafür an die RDA-Rechnung inhaltlich angepassten Bezeichnungen ‚Gesamtähnlichkeits-Koordinaten' (GK) für die ‚WA' Variante und ‚Schätzähnlichkeits-Koordinaten' (SK) für die ‚LC' Variante vorgeschlagen. Nochmal zur Präzisierung: Die Entscheidung über die Varianten GK und SK betrifft nur die Fallkoordinaten im RDA-Ergebnis, nicht die VM.

Darüber, welche Variante zu benutzen sei, wurde vor allem bei Anwendungen der verwandten Methode der

kanonischen Korrespondenzanalyse (CCA) diskutiert (Oksanen 2020). Bei der CCA gibt es wegen ähnlicher Berechnung ebenfalls beide Koordinaten-Varianten für die Fälle. Während Palmer (ders. 1993) aus Gründen der genauen technischen Deutung die SK als Standard empfahl, plädierte McCune für die GK (ders. 1997). Zunächst seien kurz zwei Effekte der SK-Variante für die Triplotdarstellung genannt. Vorgreifend auf die Regeln zur Triplot-Deutung (Borcard *et al.* 2018, 156 f. und 216 f.) sei angemerkt, dass dort VM und Ratio-KM als Pfeile dargestellt werden und Kategorien von Nominal-KM als große Punkte, sog. Zentroiden (s. u.), abgebildet werden. Teilweise schrumpfen bei der Verwendung der SK die Pfeile unter maßstabsgetreuer Abbildung der RDA-Zahlenwerte u. U. bis zur Unkenntlichkeit (Legendre und Legendre 2012, 646). Gravierender ist aus Sicht des Autors aber, dass bei rein nominalen KM die mit SK abgebildeten Fallpunkte auf die Position der Zentroiden ‚zusammenfallen'. Denn in den SK steckt ja durch den Regressionsschätzungsschritt (s. o.) jeweils nur die Information, die die Position des durchschnittlichen Falles einer Kategorie/Gruppe bezeichnet. Die Position des durchschnittlichen Falles einer Kategorie/Gruppe ist aber zugleich die Definition für die Zentroidkoordinate. Es geht also in einem solchen Triplot (und auch in den Zahlenwerten) sämtliche Information über die interne Unterschiedlichkeit in den Gruppen einer Nominal-KM verloren.

Der Autor schließt sich hier der Präferenz von Oksanen (ders. 2020) und McCune (ders. 1997) für die GK-Variante aber vor allem deswegen an, weil McCunes Argument für diese Variante auch in den Archäologien zum Tragen kommt. Er konnte zeigen, dass ungenau erfasste oder nur unscharf messbare KM zu verzerrten SK-Werten und damit zu schlecht deutbaren Ergebnisstrukturen führen. Nun sind aber gerade in den Archäologien die Werte der KM für die einzelnen Fälle manchmal auch nur vage oder grob bestimmbar. Die Archäologien besitzen sogar ein geradezu klassisches Beispiel zu McCunes Argumentation: wenn es als KM genutzt werden soll, muss ein Radiokarbondatum mit nur einer einzelnen Zahl ausgedrückt werden. In Wahrheit aber ist die Radiokarbondatierungsinformation eine Wahrscheinlichkeitsverteilung, also ein unscharfes Merkmal, das durch einen einzelnen Wert nur vage zusammengefasst werden kann. Ähnlich liegt die Situation, wenn als nominales KM eine relativchronologische Gruppierung – etwa nach Zeitstufen – verwendet wird. Dabei könnten zwei Fälle einmal von den entgegengesetzten ‚Enden' zweier benachbarter Zeitstufen stammen, ein andermal von den aneinandergrenzenden. Wieder ist die Bestimmung des KM also eher vage. Deshalb plädiert der Autor dafür, dass auch die Archäologien bei kanonischen Ordinationen (RDA und CCA) die GK-Variante als Standard verwenden, die im R-Paket ‚vegan' sowieso die Grundeinstellung ist. Nochmals im originalen englischen Fachbegriff: Empfohlen wird ausdrücklich die Verwendung sog. ‚WA-*scores*' (der Fälle).

b) Art der Informations-Darstellung im RDA-Ergebnis (Koordinaten-Skalierung)

Ging es im letzten Abschnitt darum, welche Information von den Fällen im Ergebnisdiagramm erscheinen soll (GK oder SK), so ist hier zu besprechen auf welche Weise die so gewählten Informationen dargestellt werden sollen. Das wird bei Ordinationen mit dem Begriff Skalierung erfasst. Damit ist i. d. R. gemeint, dass bereits errechnete Koordinatenwerte noch einmal mit Zahlen bzw. Zahlenmatrizen multipliziert und dadurch vergrößert oder verkleinert werden (Multiplikation mit Zahlen kleiner 1) bzw. durch den Skalierungsschritt auf bestimmte Größen fixiert werden. Dies erfolgt üblicherweise, weil die Koordinaten noch eine Bedingung erfüllen sollen, die bessere Deutungs- oder Darstellungseigenschaften besitzt.

Heutzutage werden bei der Standardvariante der Eigenwertzerlegung die Eigenvektoren von vornherein auf die Länge 1 normiert (Legendre und Legendre 2012, 94). Im oben beschriebenen Ablauf des PCA-Schrittes ist also implizit eine Skalierung der Eigenvektoren enthalten. Die Eigenvektoren werden, wie auch schon angemerkt, in der Ergebnisgrafik zur Darstellung der VM benutzt. Sie bezeichnen die Position der Spitzen der als Pfeile (vgl. u.) abgebildeten VM. Daher beinhaltet also diese Standardvariante der Eigenwertzerlegung implizit eine Skalierung der Darstellung der Merkmale in der Ergebnisgrafik. Diese Konstellation aus RDA-Fallkoordinaten und VM-Koordinaten wird in der Ökologie seit einigen Jahren für die PCA als Distanz-Biplot bzw. Biplot in Skalierung I und für die RDA als Distanz-Triplot bzw. Triplot in Skalierung I bezeichnet (Legendre und Legendre 2012, 443 f. und 639 f.). Diese Bezeichnungen seien wiederum für die Archäologien empfohlen.

Daneben gibt es die Möglichkeit, die Koordinaten der VM – also die Eigenvektoren – anders zu skalieren, nämlich so, dass der einzelne Merkmalspfeil (Eigenvektor) mit der Wurzel des Achseneigenwertes multipliziert wird (a. a. O., 435). Weil so diesmal die VM-Koordinaten Informationen (Achseneigenwert-Wurzel) zu den Fallunterschieden beinhalten, sind die Winkel zwischen ihnen unverzerrte Repräsentationen ihrer Beziehungen (Kovarianzen oder Korrelationen). Daher wird in der Ökologie diese Biplotvariante als Korrelations-Biplot bzw. Biplot in Skalierung II und für die RDA als Korrelations-Triplot bezeichnet (a. a. O., 640 f.). Dies sei wiederum als Begriff empfohlen. Die Fallkoordinaten werden dafür ebenfalls umgerechnet (skaliert). Ihre Abstände sind dann in einem eher schwer verständlichen Distanzmaß, der Mahalanobis-Distanz (a. a. O., 303 und 441), ausgedrückt. Bei einer tb-RDA wird die Deutung der Fall-Abstände damit *de facto* hinfällig.

Der Grund für die Präfixe ‚Distanz-' und ‚Korrelation' liegt in einem Problem, das PCA, RDA, CA und CCA gemeinsam ist, und dass man etwa mit dem Sprichwort „Wasch mir den Pelz, aber mach mich nicht nass"

umschreiben könnte. Aufgrund der Vorgehensweise für die Erzeugung von Koordinatensystemen (die Projektion) kann ein Biplot oder auch ein Triplot nämlich entweder nur die Beziehungen zwischen den Fällen unverzerrt darstellen oder nur die Beziehungen zwischen den VM (Borcard *et al.* 2018, 156 bzw. 211 f.), beides zusammen geht nicht. Da die Fälle als Punkte und ihre Ähnlichkeiten untereinander durch ihre Abstände dargestellt werden, sollten diese Abstände nicht verzerrt werden, wenn die Fallbeziehungen interessieren. Daher das auch sinnvolle Präfix ‚Distanz-' für die ansonsten als ‚Skalierung I' bezeichnete Darstellungskonvention. Umgekehrt erscheinen die VM als Pfeile (vgl. u.), die alle vom Achsenursprung des Triplots ausgehen. Ihre Beziehungen können so nur in den Winkeln zwischen den Pfeilen enthalten sein. Wenn also die VM-Beziehungen untereinander interessieren, sollten die Winkel zwischen den Pfeilen nicht verzerrt werden. Weil diese Beziehungen zwischen z-transformierten (Ratio-)Merkmalen auch als Korrelationen bezeichnet werden, ist das Präfix ‚Korrelation-' für die Darstellungskonvention ‚Skalierung II' ebenso sinnvoll. Die Bezeichnung Korrelations-Triplot wird auch dann verwendet, wenn die Rotation anhand der Kovarianzen – und nicht der Korrelationen – berechnet wurde, wie es bei Chord- oder Hellinger-transformierten Anzahlen zwingend notwendig ist.

Das Problem der Wahl der Skalierungsart ist also diesmal kein philosophisches, sondern eine rein praktische Interessens-Entscheidung. Interessiert die Fallanordnung, dann wird ‚Skalierung I' bzw. die ‚Distanz-Variante' gewählt. Interessieren die Beziehungen zwischen den VM, dann wird ‚Skalierung II' bzw. die ‚Korrelations-Variante' gewählt. Bei einer tb-RDA von multivariaten Zusammensetzungs-Anzahldaten liegt i. d. R. der Fokus auf der Ähnlichkeit zwischen den Fällen, weshalb dann der Distanz-Triplot verwendet werden sollte. Ein Korrelations-Triplot kommt bei der tb-RDA dann zum Einsatz, wenn interessiert, welche Typen gemeinsam in ähnlichen Anzahlen vorkommen – dies ist die praktisch-archäologische Deutung des Konzepts ‚Korrelation', wenn die VM nach Typen ausgezählte Anzahldaten sind.

Der Unterschied zwischen dem Inhalt des vorhergehenden und dieses Abschnittes sei nochmals betont: Oben ging es darum, welche Information für die Fälle dargestellt werden sollte, Koordinaten mit der gesamten Ähnlichkeitsinformation (GK) oder solche nur mit dem kausalen Teil dieser Information (SK). In dem hiermit beendeten Abschnitt dagegen wurde vorgestellt, wie das RDA-Ergebnis (Fälle und VM) je nach Erkenntnisinteresse darzustellen ist, als Distanz-Triplot zur Deutung der Fallähnlichkeiten oder als Korrelations-Triplot zur Deutung der VM-Beziehungen untereinander. Während die Frage des vorigen Abschnittes nur bei kanonischen Ordinationen auftaucht, muss die Frage, welche Skalierung verwendet wurde, stets auch bei einfachen, erkundenden Ordinationen (PCA und CA) klar beantwortet werden.

c) Aufbau eines Triplots

Im Triplot werden, wie schon mehrfach erwähnt, die mit der (tb-)RDA herausstellbaren (kausalen) Beziehungen in und zwischen allen drei Datenteilen, den Fällen, den VM und den KM, in einer Grafik zusammengefasst. Sie sind somit zwar im Prinzip ‚auf einen Blick' erkennbar, aber dafür müssen Auge und Hirn zunächst im Verständnis des Triplot-Aufbaues trainiert werden (Abb. 4). Erst dann kann der Vorteil visuell hochkompakt zusammengefasster Information voll ausgeschöpft werden.

Der Triplot ist nur eine Weiterentwicklung des 1971 vom deutsch-israelisch-britischen Statistiker Kuno Gabriel formalisierten sog. Biplots (ders. 1971), welcher seitdem weltweit das Grundprinzip aller Visualisierungen von Ordinationen und in der Folge von kanonischen Ordinationen darstellt (zum Aufbau von Bi- und Triplot auch: ter Braak 1994). Abstrakt gesagt stellt solch ein Biplot die Information einer Zahlentabelle (Matrix) als zerlegt in zwei Teile (zwei Matrizen) dar, einen Teil, der die Zeilen und einen anderen Teil, der die Spalten repräsentiert. Beide, Zeilen wie Spalten, werden dabei von Zahlenreihen repräsentiert, die als (Gruppen von) Vektoren bezeichnet werden und entweder als Punkt oder als Pfeil (Vektorkoordinaten = Pfeilspitze) darstellbar sind. Ihre Koordinaten stehen in den beiden Teilmatrizen. Damit es sich um einen ‚echten' Biplot handelt, müssen die beiden Teile eine Bedingung erfüllen, nämlich dass ihre rechnerische Verbindung (mit Matrixalgebra) wieder die ursprüngliche Tabelle rekonstruiert. Das ist mathematisch weitgehend gleichbedeutend damit, dass die beiden in einem Biplot gemeinsam dargestellten Vektorengruppen (Zeilen und Spalten) über eine Rechnung (inneres Produkt bzw. Skalarprodukt), die ihre Verbindung in einer Zahl (Skalar) ausdrückt, verbunden werden können (Gabriel 1971, 454). Ins Praktische übersetzt erlaubt die Existenz eines Skalarproduktes, dass die Abstände und Winkel zwischen den beiden Vektorengruppen so gedeutet werden können, dass die einen – gedacht als Punkte – ‚auf' den anderen – gedacht als Pfeile – ‚abgelesen' werden können bzw. dass die einen auf die anderen projiziert werden können. Oder, geradezu gefährlich vereinfacht, dass Abstände und Winkel zwischen zwei Exemplaren aus den verschiedenen Vektorgruppen eine sinnvolle Deutung besitzen. Dies, nämlich dass gemeinsam Abgebildetes auch gemeinsam deutbar ist, ist nicht so banal wie es klingen mag, gerade wenn Vereinfachungen multivariater Beziehungen dargestellt werden sollen. Bezogen auf eine Datentabelle mit Fällen in den Zeilen und Merkmalen in den Spalten ist ein Biplot also die Möglichkeit der gemeinsamen Darstellung von Fällen und Merkmalen innerhalb des gleichen Systems. Beide werden technisch als Vektoren gedacht, aber die einen als Punkte, die anderen als Pfeile dargestellt.

Die Standardansicht eines Biplots (und damit eines Triplots) besteht i. d. R. aus den beiden informationsreichsten Achsen nach einer Drehung der Fallpunkt-

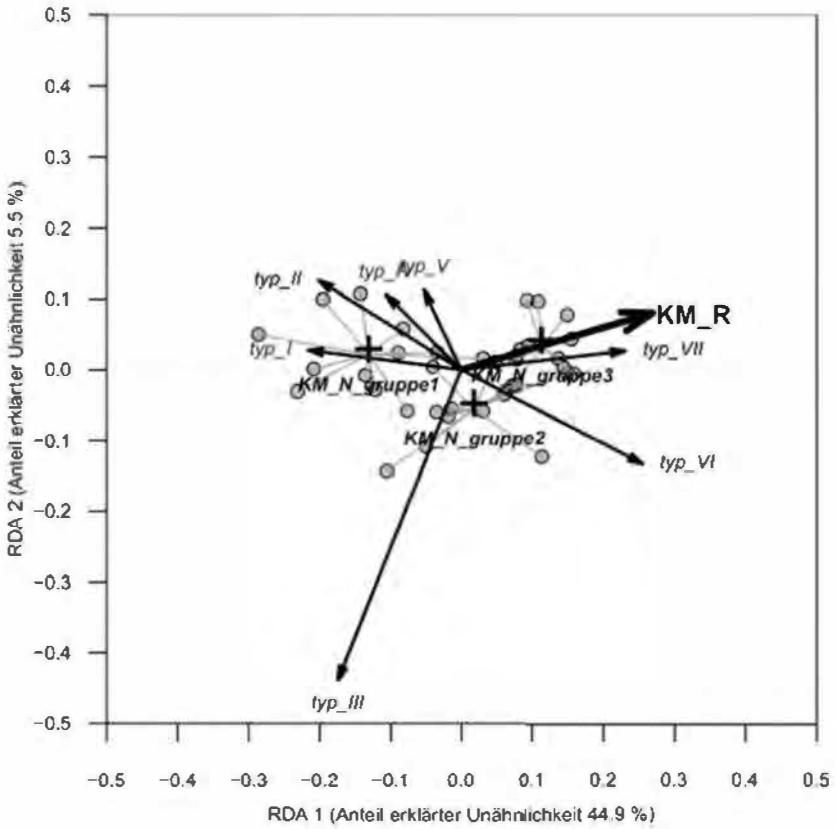


Abb. 4. Beispiel für einen tb-RDA-Distanz-Triplot künstlicher Daten. Fälle als graue Punkte, verursachte Merkmale (VM) als dünne Pfeile, kausales Ratiomerkmal (KM_R) als dicker Pfeil und Kategorien des kausalen Nominalmerkmals (KM_N) als Zentroide (Gruppe1 etc.); Fälle jeweils einer Kategorie des KM_N sind mit Linien (‚ordispider') verbunden. Nach den Deutungsregeln (*siehe Text*) kann z. B. erkannt werden:
- Erhöhung der Werte des KM_R verursacht höhere Anteile von Typ VI und VII, und niedrigere bei allen anderen,
- die Ähnlichkeit der Fälle der ‚Gruppe1' wird durch niedrige, die der ‚Gruppe2' durch mittlere und die der ‚Gruppe3' durch höhere Werte der KM_R verursacht,
- Fälle rechts weisen hohe Anteile der Typen VI und VII auf, Fälle links unten hohe Anteile bei Typ III und Fälle links oben hohe Anteile bei den Typen I, II, IV und V.

wolke. Als 2D-Ansicht ähnelt sie stark einem zweidimensionalen Streudiagramm (*scatterplot*). Obwohl mit digitalen Medien heute auch (filmartig animierte) dreidimensionale Darstellungen möglich sind, werden diese m. W. fast nie benutzt. Das im wahrsten Wortsinne Grundgerüst des Plots bilden dabei seine beiden Koordinatenachsen. Diese werden üblicherweise von den ersten beiden Koordinatenachsen des RDA-Ergebnisses gebildet. Dass die Achsen stets senkrecht aufeinander stehen, ist eine grafische Konvention, um zu zeigen, dass die einzeln RDA- bzw. PCA-Achsen voneinander unabhängige Information enthalten. Bewegt man sich z. B. waagrecht durch ein 2D-Diagramm, gibt es wohl Veränderungen der X-Koordinate, aber die Y-Koordinate bleibt unverändert. Die Information über X kann sich ändern, während die über Y gleichbleibt. Das ist mit unabhängig gemeint. Tatsächlich besteht ein PCA- bzw. ein RDA-Ergebnis meist aus viel mehr Achsen als dargestellt. Aber die ersten paar Achsen zeigen eben die großen, also wichtigeren Unterschiede. Die Bedeutung der Achsen war bereits am *‚screeplot'* (s. o.) ablesbar. Welchen Anteil an der Gesamtinformation eine Achse durch die an ihr ablesbaren Fall-Unterschiede abdeckt, kann in einem Bi- oder Triplot auch als Prozentangabe neben den Namen der Achse (RDA oder PCA) gesetzt werden (ter Braak 1994, 134). Diese Angabe ergibt sich wie gesagt durch den Anteil des Achsen-Eigenwertes an der Summe aller Eigenwerte.

Anders aber als bei Streudiagrammen liegt der Achsenursprung des Plots, die Nullkoordinate, im Zentrum des Diagramms und die Achsen besitzen folglich positive und negative Abschnitte. Die Nullkoordinate kann sowohl als die Position des durchschnittlichen Falles verstanden werden als auch als das Zentrum der Fallpunktwolke, um das diese auf die optimale Ansicht gedreht wurde. Fallpunkte die ganz nahe am Zentrum liegen, sind also einem imaginären Durchschnittsfall ähnlich.

Der RDA-Triplot ist wie gesagt als Gerüst i. d. R. aus den Achsen RDA 1 und RDA 2 aufgebaut. Er ist schlicht die Erweiterung des Biplots um die rechnerisch und geometrisch angemessene Darstellung der Beziehungen zu den KM. Bei nur einem Ratio-KM oder nur einem Nominal-KM mit nur zwei Kategorien gibt es nur eine kano-

nische Achse (RDA 1). Dann ist es mittlerweile üblich, als Y-Achse des Triplots die erste PCA-Achse der PCA der Residuen zu verwenden. Da ist inhaltlich gesprochen also ein Raum, der aus dem eindimensionalen kausalen Ähnlichkeitsraum und der ersten Dimension des nicht kausal herleitbaren restlichen Ähnlichkeitsraumes besteht. Gibt es mehrere KM können bei fragestellungsrelevanten Ergebnissen auch andere RDA-Achsenpaare als das aus RDA 1 und RDA 2 in einem dann zusätzlich abgebildeten und gedeuteten Triplot dargestellt werden – wenn sie getestet wurden (s. o.). Der Triplot der – falls vorhanden – ersten beiden RDA-Achsen bzw. der RDA-Achse und der ersten PCA-Achse sollte aber immer zuerst dargestellt und gedeutet werden.

In einem PCA-Bi- wie in einem RDA-Triplot sind die Fälle stets als Punkte dargestellt. Die Variablen der PCA bzw. die VM der RDA werden als Pfeile dargestellt. Die Pfeildarstellung betont, dass die so repräsentierten Merkmale als Achsen gedeutet werden dürfen, entlang derer Werte abgelesen werden können. Die Pfeile beginnen dabei stets im Achsenkreuzursprung und ihre Spitze sind die Ergebniskoordinaten der VM (Eigenvektoren). Das Ablesen eines Fallwertes bei einer VM erfolgt, indem man von einem Fallpunkt ein Lot auf einen Merkmalspfeil fällt und damit ungefähr ablesen kann, wie hoch der Wert des Falles bei diesem Merkmal ist (ter Braak 1994, 132). Dass dies Sinnvolles ergibt, wird durch die oben beschriebene Existenz des Skalarproduktes ermöglicht. Wie beim PCA-Schritt erwähnt, wird ja das gesamte Ensemble aus Fallpunkten und originalen Merkmalsachsen zur optimalen Ansicht gedreht, weshalb in einem solchen Diagramm grundsätzlich immer noch die ursprünglichen Fallwerte auf den originalen Merkmalsachsen ablesbar sind – allerdings rotiert und skaliert und nicht als Zahlenwerte, sondern nur als relative Abfolge. Was aber, wenn der Fallpunkt links unten liegt, der Merkmalspfeil aber vom Achsenursprung (der Diagramm-Mitte) nach rechts oben zeigt, wie fällt man dann das Lot? Nun, man verlängere den Pfeil quasi ‚nach hinten' durch den Achsenursprung nach links unten und fälle dann das Lot und lese die relative Position ab. Wenn nämlich der Achsenursprung die Position des durchschnittlichen Falles repräsentiert, dann können Ablesungen durch Lotfällen auf den Pfeil selbst nur überdurchschnittliche Werte repräsentieren. Die Merkmalspfeile repräsentieren also streng genommen nur den überdurchschnittlichen Wertebereich eines originalen Merkmals. Sie weisen quasi vom Fuß bei den durchschnittlichen zu den höchsten Werten bei ihrer Spitze. Die Darstellung als Pfeil bildet den Wertebereich für unterdurchschnittliche Werte nicht ab, sonst würden Biplot- und Triplot-Diagramme schnell ‚zu voll'. Diese Verbindungen zwischen Punkten und Pfeilen machen klar, dass in einem solchen Diagramm neben Punktabständen auch Richtungen eine entscheidende Rolle für das Deuten spielen (*ibid.*). Wenn gerade gesagt wurde, es ließen sich in einem Diagramm die Werte aller Fälle auf allen VM ablesen, ist dies ist etwas zu relativieren. Die vereinfachende 2D-Abbildung einer Multi-D-Beziehung bedingt, dass die Fall-Werte nur angenähert ablesbar sind. Es ist schlicht die bestmögliche 2D-Vereinfachung eines Multi-D-Komplexes.

Ein VM (der Eigenvektor eines Originalmerkmals) hat als Pfeil durch alle Dimensionen des multivariaten Ähnlichkeitsraumes bei einem Distanz-Biplot oder -Triplot per Definition (s. o. Skalierung) die Länge Eins (eine Einheit an der RDA- oder PCA-Achsen-Wertbeschriftung; vgl. Legendre und Legendre 2012, 436 f.). Der Plot ist als 2D-Phänomen quasi die informationsreichste Bildebene ‚durch' den Multi-D-Raum. Sehr kurz erscheinende Pfeile kann man sich plastisch als ‚fast senkrecht aus der Bildebene herausweisende' Merkmalsachsen vorstellen. Die Unterschiede bei diesen Merkmalen sind dann nicht gut in der Bildebene des Plots zu sehen. Sehr lange Pfeile weisen auf weitgehend ‚in der Bildebene liegende' Merkmale hin. Die mit ihnen verbundenen Unterschiede ‚sieht' man gut im Plot. Je länger die VM-Pfeile, desto ‚besser' sind die Merkmale im Plot erfasst (a. a. O., 444). Dies lässt sich auch in Zahlen fassen. Die neuen Achsen selbst, also RDA- oder PCA-Achsen, sind ja quasi neue, die Information mehrerer alter Merkmale gemeinsam zeigende Merkmale/Variablen. Was eine Achse vereint, kann bei Bedarf von einem Säulendiagramm der Eigenvektorwerte (= Koordinaten) der VM für diese Hauptkomponente abgelesen werden. Die Eigenvektoren werden auch als Ladungen (*loadings*), und solch ein Zusatzdiagramm als Ladungsdiagramm bezeichnet. Es zeigt die Eigenvektorwerte aller VM für eine Hauptkomponente (neue Achse nach Drehung). Eine neue RDA- oder PCA-Achse misst demnach vor allem die Unterschiede der Merkmale gemeinsam (!), die eine hohe Ladung (einen großen Eigenvektorwert) auf dieser Achse besitzen. Bei einer ausführlicheren RDA-Ergebnisdiskussion zeigt man für die zwei im Plot dargestellten Achsen auch jeweils das Ladungsdiagramm. Unter Verwendung dieser Sichtweise ist es auch möglich, bei Bedarf für jedes VM den Anteil der durch die KM insgesamt (über alle RDA-Achsen hinweg) erklärten Streuung als Säulendiagramm darzustellen (Funktion ‚inertcomp' im R-Paket ‚vegan'). So lässt sich für jedes VM einzeln seine Abhängigkeit von den KM aufzeigen, und damit, ‚wie stark' es von diesem/n ‚verursacht' wird.

Die Darstellung der KM im Triplot richtet sich danach, ob es Ratio- oder Nominalmerkmale sind. Ein Ratio-KM wird als Pfeil, ein Nominal-KM als großer oder auffälliger Punkt (*Zentroid*) abgebildet. Das ist nach dem über die Verbindung von Fällen und VM bzw. über das Ablesen der Fallwerte auf den VM Gesagten analog auf die KM übertragbar. Wenn es ein Ratio-KM ist, dann besitzt es eine Werteskala, die sinnvollerweise als Pfeil, also eine Art Werte-Achse darstellbar ist. Ist es ein Nominal-KM dann besteht der Informationswert dieses Kausalmerkmals einfach in einer kategorialen Zuweisung ohne Details über relative oder gar absolute Beträge (!) des Wertunterschiedes. Die Darstellung als Punkt (Zentroid) soll also Verwechslungen bei

der Deutung unterbinden. Die Deutung der Beziehung zwischen Fall- und Zentroid-Punkten benutzt nur ihren Abstand, hier werden keine Lote gefällt (s. u.).

Noch ein Hinweis: Je nach verwendetem Computer und/oder Software kann die Situation auftreten, dass bei der Wiederholung einer Analyse – sowohl bei einer PCA als auch bei einer RDA – alle Positionen (Fallkoordinaten, VM-Pfeilspitzen und KM-Pfeil- und Zentroidinformationen) entlang einer Achse gespiegelt sind. Das ist kein Fehler oder gar ein neues Ergebnis. Aufgrund allerkleinster Rundungsungenauigkeiten verschiedener CPU-Fabrikate bei der Matrixalgebra der Eigenwertzerlegung oder auch aufgrund bestimmter Chipsatz-spezifischer Regeln zur Abarbeitung von Rechenanweisungen können sich bei der Reproduktion eines Ergebnisses die Vorzeichen für die Informationen entlang einer Ergebnisachse umkehren. Das hat keinerlei Bedeutung (Legendre und Legendre 2012, 460)! Stellen Sie sich den Triplot einfach als einen Ausdruck auf einer durchsichtigen Folie vor. Wenn Sie die Folie umdrehen, sind alle informationstragenden Aspekte des Triplots, nämlich Abstände und Winkel, unverändert. Die Achsenvorzeichen tragen, für sich allein genommen, keine Information und sind ohne Veränderung des dargestellten Inhaltes vertauschbar.

d) Regeln zur Deutung eines Distanz-Triplots

Nach diesen Erläuterungen zum Aufbau und der ganz allgemeinen Betrachtung werden die unterschiedlichen Regeln zur Deutung eines Distanz-Triplots (Borcard *et al.* 2018, 156 f. und 216 f.; Legendre und Legendre 2012, 640) für jedes der drei Ergebnisteile (Fälle, VM und KM) getrennt im Einzelnen aufgelistet. Bei Bedarf finden sich die Regeln zur Deutung eines Korrelationstriplots an gleicher Stelle (*ibid.*), dort aber noch als Skalierung II bezeichnet.

Fälle (einfache/kleine Punkte; Fallpunkte = FP):

- Der Abstand der FP entspricht der in 2D bestmöglichen Annäherung an ihre euklidische Distanz; je näher zwei FP desto ähnlicher sind sie sich; wurde eine tb-RDA berechnet, entspricht der FP-Abstand der bestmöglichen Annäherung in 2D an dasjenige Distanzmaß, welches zur Transformation benutzt wurde.

- Ein FP kann durch Lot-Fällen auf den Pfeil eines VM projiziert werden und dadurch sein ungefährer Wert beim VM bestimmt werden; dazu gilt Obiges zum Projizieren und ‚Pfeilverlängern'.

- Ein FP kann durch Lot-Fällen auf den Pfeil eines Ratio-KM projiziert werden und dadurch sein ungefährer Wert beim KM bestimmt werden; dazu gilt Obiges zum Projizieren und ‚Pfeilverlängern'.

- Je näher ein FP einem Nominal-KM-Zentroiden, desto wahrscheinlicher gehört er zu dieser Kategorie des Nominal-KM; bei anhand der kausalen-VM-Information scharf trennbaren Gruppen bilden die FP einer Nominal-KM-Kategorie sichtbar aufgeteilte Punktcluster mit ihrem Zentroiden ungefähr in der Mitte; bei anhand der kausalen-VM-Information nur unscharf trennbaren Gruppen bilden sie sich überlagernde Punktcluster (Wenn das Nominal-KM einzeln (*by margin*) erfolgreich getestet wurde, sind auch sich überlagernde Gruppen trotzdem noch deutungswürdig, weil rechnerisch ‚in der Tendenz' noch auftrennbar).

- Die Ergebnisachsen (RDA oder PCA) können als neue, künstlich Information zusammenfassende, Merkmale verstanden werden; die FP-Koordinate auf einer Achse ist ihr ‚Fall-Wert' bei dieser synthetischen Variable (synthetisch = zusammengesetzt); wie diese neue Variable gemeinsam Unterschiede bei den alten Merkmalen misst, zeigt das Ladungsdiagramm (s. o.) dieser Achse.

Verursachte (= originale) Merkmale ((dünne) Pfeile = VM):

- Der Winkel zwischen zwei VM-Pfeilen ist nur (i. d. R. leicht) verzerrt proportional zur Korrelation zwischen zwei KM; ganz grob lässt sich sagen, je kleiner der Winkel, desto stärker korreliert sind die beiden VM.

(Die Verzerrungen der Winkel im Distanztriplot untersagen eine Deutung der Korrelationen, aber sie lassen zumindest ein grobes Bild der Korrelationsbeziehungen zu. Interessieren die VM-Korrelationen, ist zusätzlich ein Korrelations-Triplot zu erstellen).

- Der Winkel zwischen einem VM-Pfeil und einem KM-Pfeil (Ratio-KM) ist proportional zur Korrelation zwischen den beiden Merkmalen; je kleiner der Winkel, desto stärker die Korrelation zwischen beiden.

Kausale Ratio-Merkmale ((dicke) Pfeile = Ratio-KM)

- Der Winkel zwischen zwei KM-Pfeilen ist nur (i. d. R. leicht) verzerrt proportional zur Korrelation zwischen zwei KM; ganz grob lässt sich sagen, je kleiner der Winkel, desto stärker korreliert sind die beiden KM.

Kausale Nominal-Merkmale ((dicke) Punkte/Zentroiden = Nominal-KM)

- Ein Zentroid kann durch Lot-Fällen auf den Pfeil eines VM projiziert werden und dadurch der durchschnittliche Wert dieser Gruppe beim VM ungefähr bestimmt werden; dazu gilt Obiges zum Projizieren und ‚Pfeilverlängern'.

- Der Abstand zweier Zentroiden entspricht der in 2D bestmöglichen Annäherung an ihre euklidische Distanz; je näher zwei Zentroiden desto ähnlicher sind sich die von ihnen repräsentierten Gruppen im Hinblick auf die VM; wurde eine tb-RDA berechnet, entspricht der Abstand zweier Zentroiden der bestmöglichen Annä-

herung in 2D an dasjenige Distanzmaß, welches zur Transformation benutzt wurde.

Diese vielen Details belegen, wie informationsreich ein Triplot ist. Dieser Reichtum kann zu umfangreichen, deutenden Beschreibungen der zahlreichen ablesbaren Beziehungen benutzt werden. Aspekte, die durch Umsetzung obiger Regeln dabei diskutiert werden können, sind etwa, welche Fälle sich ähnlich und welche sich unähnlich sind, welche Werte diese Fälle ungefähr bei Ratio-KM haben, und zu welcher Nominal-KM-Kategorie sie anhand ihrer VM-Werte tendieren – das muss nicht ihre ‚wahre' Gruppe sein. Tatsächlich können sie etwa schlecht in der 2D-Ebene des Triplots erfasst worden sein. Weiterhin lässt sich erörtern, wie die VM auf die KM reagieren, also bei welchen VM die Werte gemeinsam mit Ratio-KM zunehmen oder wo die VM-Werte zunehmen, wenn die Ratio-KM-Werte abnehmen. Schließlich lassen sich noch die Kategorien der Nominal-KM (die Zentroide) im Hinblick auf die bei ihnen auftretenden VM und Ratio-KM-Werte charakterisieren. Die Deutung eines Triplots erlaubt eben umfangreiche Einblicke in komplexe Strukturen. Es lohnt sich also, einen Triplot auch in Worte zu fassen, um gerade denjenigen, die noch nicht ausreichend im visuellen Interpretieren trainiert sind, die wichtigsten Ergebnisaspekte zugänglich zu machen. Die Triplot-Koordinatentabellen sind als Inhalte eigentlich nur für Anhänge und Supplemente sinnvoll.

Wer eigene grafische Kompositionen bevorzugt, beachte Folgendes. Die Zahlengrundlagen eines Triplots bestehen aus drei Koordinatentabellen (je eine für Fälle, VM und KM). Für alle drei ist ausschließlich die zur jeweils darzustellenden Skalierung (s. o.) passende Koordinatenvariante zu verwenden. Die Darstellungskonventionen, was als Punkt, als dicker Punkt, als dünner und als dicker Pfeil abgebildet wird, sind unbedingt einzuhalten. Ansonsten werden gerade die Betrachter:innen verunsichert, die schon mit Triplots umgehen können und man säht bei ihnen Zweifel über die Korrektheit der ihnen vorliegenden Analyse.

7. Software

Für die Praxis wird hier kein umfassender Software-Überblick geboten. Es sei nur auf wenige Implementierungen hingewiesen, aber dafür die besten, wichtigsten und – vermutlich – am häufigsten verwendeten.

Den internationalen Standard bildet m. W. mittlerweile in zahlreichen Wissenschaften die RDA-Implementierung im Erweiterungspaket ‚vegan' (Funktion ‚rda'; Oksanen *et al.* 2020) der freien statistischen Programmieroberfläche R (R Core Team 2020). ‚Vegan' steht für *vegetation analysis*. Als Freeware im Range wissenschaftlicher Publikationen genügen sowohl die eigentlich eine Programmiersprache darstellende Software R wie das Erweiterungspaket ‚vegan' höchsten Ansprüchen an Qualität, Stabilität, FOSS-Kriterien und vor allem Nachhaltigkeit. Da alle Analysen mit R-Paketen als Programmskripte vorgelegt werden können, lassen sich Daten und Analysen von Dritten direkt überprüfen – und vor allem weitergeben und zur Weiterbildung nutzen. Speziell in Bezug auf die RDA ist anzumerken, dass der Programmcode für die RDA und für den RDA-Test vom Autor Oksanen persönlich verfasst wurden. Und nur ‚vegan' bietet die Teste für die Teilbereiche (KM einzeln, RDA-Achsen einzeln) eines RDA-Ergebnisses. Die beiden hier umfangreich benutzten Bände von Borcard *et al.* (dies. 2018) sowie Legendre und Legendre (dies. 2012) lassen sich dabei als allumfassende Lehrbücher für die mathematische Theorie (Legendre und Legendre 2012) und die praktische Umsetzung mit ‚vegan' in R (Borcard *et al.* 2018) verwenden, die konsistent ineinandergreifen, da Daniel Borcard ehem. Mitarbeiter und Nachfolger von Pierre Legendre ist.

Eine weitere RDA-Umsetzung findet sich im R-Erweiterungspaket ‚ade4' – lies: *‚A-D-four-E'* (Dray und Dufour 2007). Die französischen Biometriker:innen um den Hauptautor Stephane Dray folgen in diesem Paket der französischen Statistiktradition, weshalb die RDA hier mit der Funktion ‚pcaiv' bezeichnet wird, also der ursprünglichen Methodenbezeichnung Raos (ders. 1964). Allerdings bietet ‚ade4' nur den Test für das RDA-Gesamtergebnis. Die Teste für die Teilbereiche werden nicht angeboten und Bereiche eines Ergebnisses aus ‚ade4' können nur mit viel manuellem Programmieraufwand mit der Test-Funktion aus ‚vegan' überprüft werden. Daraus ergibt sich, dass in ‚ade4' nur eine RDA mit einem einzigen KM gerechnet werden kann, denn bei zwei und mehr KM ist das Ergebnis in diesem Paket nicht vom Zufall abgrenzbar. Auch zu ‚ade4' gibt es ein begleitendes Lehrbuch (Thioulouse *et al.* 2018).

Von einer RDA mit dem R-Erweiterungspaket ‚calibrate' (Graffelman und van Eeuwijk 2005) sei abgeraten, da es keinerlei Test anbietet. Und Ergebnisse aus ‚calibrate' können nur mit viel manuellem Programmieraufwand mit der Test-Funktion aus ‚vegan' überprüft werden.

Die für die tb-RDA grundlegenden Transformationen (Chord oder Hellinger) sind im R-paket ‚vegan' implementiert (Funktion ‚decostand', Argument „method= ‚norm'" bzw. Argument „method= ‚hell'"; Oksanen *et al.* 2020). So transformierte Datentabellen können in R auch mit ‚ade4' oder ‚calibrate' ausgewertet werden.

Die freie Software PAST des norwegischen Paläontologen Oyvind Hammer (Hammer *et al.* 2001) erlaubt die Durchführung einer RDA ohne Programmierkenntnisse. Allerdings benötigt sie mindestens zwei KM und bietet trotzdem nur den Test des Gesamtergebnisses an. Dafür handelt es sich um ein menübasiertes Programm, das ein Tabellenkalkulationsprogramm nachahmt und so die Durchführung stark vereinfacht. PAST bietet

zudem unter dem *Menu Transform* mit der Option ‚*row normalize length*' eine Chordtransformation für Anzahldaten an.

Die kommerzielle Software CANOCO 5 (ter Braak und Smilauer 2012; Lizenzkosten ca. 300 $) konnte nur nach den frei zugänglichen Informationen (www.canoco5.com/index.php/resources/8-overview-article) beurteilt werden. Danach wird die RDA und mindestens die Hellinger Transformation angeboten. CANOCO 5 ist m. W. weltweit ein Standardprogramm in vielen Archäobotanik-Laboren und v. a. in vielen Botanik-Instituten.

8. Zusammenfassung und epistemologischer Ausblick

Bevor noch ein paar weitergehende Gedanken zur RDA in der Archäologie formuliert werden, sei das Grundprinzip und Ergebnispotenzial noch einmal zusammenfassend wiederholt. Ausgangslage ist die Beschreibung von Entitäten (Fällen/Objekten) mit vielen Merkmalen (Ratiomerkmalen oder Kategorieauszählungen = transformierte Anzahl-Zusammensetzungsdaten) und mit weiteren Ratio- oder Nominalmerkmalen, die zu jedem Fall die Werte möglicher Ursachen für die zuvor beobachteten Merkmale vermerken. Wenn die auf ihre Kausalreaktion betrachteten Merkmale Anzahldaten sind, müssen diese für eine tb-RDA angemessen transformiert werden (Chord- oder Hellinger). Das Ergebnis der RDA ist dann die Anordnung der Entitäten zuvorderst nach dem Teil der Ähnlichkeit, der durch die mögliche/n Ursache/n erzeugt wird. Dabei wird zunächst geprüft, ob die Auswirkungen wirklich als nicht-zufällig anzusehen sind (Permutationstest). Dann wird die Gesamtstärke der Auswirkungen in einer Maßzahl (bimultivariate Redundanzstatistik), dem adjustierten kanonischen R^2_{adj}, ausgedrückt. Und schließlich wird ein extrem informationsreiches Diagramm (Triplot) erzeugt, dass die vielfältigen Beziehungen zwischen den drei beteiligten Phänomenen, den Fällen, den zunächst beobachteten Merkmalen und den kausal wirkenden Merkmalen in einer Grafik visualisieren kann. Damit wird umgesetzt, ob, wie stark und wie genau im Einzelnen ein bereits in sich facettenreiches Phänomen (die multivariaten Daten) von einem – möglicherweise vielfältigen – Bündel von Ursachen beeinflusst wird. Der Fortschritt durch die tb-RDA ist es, beliebige Ursachen für Phänomene zu erforschen, die mit Anzahldaten multivariat beschrieben wurden. Eine erfolgreiche RDA führt zu verallgemeinerbaren historischen Deutungen der Ursachen für kompliziert zusammengesetzte Phänomene.

An dieser Stelle ist noch einmal auf die durch eine (tb-) RDA möglichen Fragen und die in ihr umgesetzten Konzepte, insbesondere der Kausalität einzugehen. Eine RDA kann anhand des Testes nahelegen, dass der zahlenmäßige Ausdruck einer Beziehung nicht Zufall im Sinne eines willkürlichen gemeinsamen Auftretens von bestimmten VM-Werten und bestimmten KM-Werten bei den Fällen ist. Der hinter den Zahlen vermutete Wirkmechanismus ist nun zunächst *a priori* theoretisch zu begründen. In dem Augenblick allerdings, wo er anhand von Daten mittels der RDA überprüft werden kann, ist eine erfolgreiche Analyse zugleich Beleg der Wirkungsbeziehung. Wie genau der kausale WirkMechanismus ist, kann wie hier durch die detaillierte Beschreibung des Algorithmus klar wurde, nicht durch eine RDA ermittelt werden. Dass es die Wirkung gibt, ist jedoch nach erfolgreicher Analyse nur dann noch vernünftigerweise abstreitbar, wenn in den Daten steckende Messungen und Typenzuweisungen als fehlerhaft identifiziert werden können. Werden die Daten akzeptiert, müssen konsequenterweise auch reproduzierbare Ergebnisse richtig angewendeter Datenanalysemethoden anerkannt werden. Dass die dabei benutzten Methoden kompliziert sein mögen, ist kein Argument gegen sie. Die hier oben als Handlungsanweisungen formulierten Erklärungen einer komplizierten Methode wie der RDA zeigen, dass die Prinzipien solcher Werkzeuge immer noch allgemein verständlich sind.

Im Zeitalter der computergestützten, datenauswertenden Archäologie werden schließlich gerade die deutenden Ansätze in den Archäologien durch diese Werkzeuge, bei der RDA etwa die Erweiterung der RDA zur partiellen RDA (Legendre und Legendre 2012, 649–652), vor neue Aufgaben gestellt. Historische Ursachen und Zusammenhänge sind hier neu zu bedenken, wenn Kausalität so untersuchbar wird – insbesondere wenn bei der partiellen RDA zuvor andere Einflüsse entfernt werden konnten. Wenn sich dann noch bei Datensätzen die ‚Lieblingsursache' der Archäologien, die durch die Zeit bedingte Veränderung, herauspartialisieren (= entfernen) ließe, und dann immer noch beispielweise ein weiteres nominales Kausalmerkmal wirksam bleibt. Solche Ergebnisse sind dem Autor noch nicht bekannt, aber sie könnten in Zukunft auftreten.

Dies sind Herausforderungen an die Archäologie-Theorie, der sie aber nur dann gerecht werden kann, wenn sie die Kenntnis moderner quantitativer Methoden (wieder) als integralen Bestandteil des Faches akzeptiert und anerkennt, dass quantitative Werkzeuge eine positive – im Sinne besserer und weitergehender Erkenntnis – qualitative Änderung des Faches darstellen und damit einen Fortschritt hin zur Archäologie des 21. Jahrhunderts.

Literatur

Borcard *et al.* 2018: D. Borcard, Fr. Gillet und P. Legendre, *Numerical Ecology with R*. 2. Aufl. (Cham 2018).

Dray, Dufour 2007: St. Dray und A. Dufour, The ade4 Package. Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software* 22 (4), 2007, 1–20. DOI: 10.18637/jss.v022.i04

Fox, Weisberg 2019: J. Fox und S. Weisberg, *An R companion to Applied Regression*. 3. Aufl. (Los Angeles 2019).

Gabriel 1971: K. Gabriel, The Biplot Graphic Display of Matrices with Application to Principal Component. *Biometrika* 58 (3), 1971, 453–467.

Gehlen *et al.* 2020: B. Gehlen, N. Schneid, G. Roth und A. Zander, Typo-Chronology for the Mesolithic between 9000 and 7800 cal BC in Central Europe. A New Approach to Use Constrained Correspondence Analysis (CCA) Of Microliths for Dating. In: A. Zander und B. Gehlen (Hrsg.), *From the Early Preboreal to the Subboreal Period. Current Mesolithic Research in Europe. Studies in Honour of Bernhard Gramsch.* Edition Mesolithikum 5 (Kerpen-Loogh 2020) 315–367.

Gittins 1985: R. Gittins, *Canonical Analysis. A Review with Applications in Ecology*. Biomathematics 12 (Berlin 1985).

Graffelman, van Eeuwijk 2005: J. Graffelman und F. van Eeuwijk, Calibration of Multivariate Scatter Plots for Exploratory Analysis of Relations within and between Sets of Variables in Genomic Research. *Biometrical Journal* 47 (6), 2005, 863–879. DOI: 10.1002/bimj.200510177

Hammer *et al.* 2001: Ø. Hammer, D. Harper und P. Ryan, PAST-Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica* 4 (1), 2001 (ohne Seitenzählung).

Hedderich, Sachs 2020: J. Hedderich und L. Sachs, *Angewandte Statistik. Methodensammlung mit R.* 17. Aufl. (Berlin 2020).

Hotelling 1933a: H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology* 24 (6), 1933, 417–441.

Hotelling 1933b: H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology* 24 (7), 1933, 498–520.

Lambert *et al.* 1988: Z. Lambert, R. Durand und A. Wildt, Redundancy Analysis. An Alternative to Canonical Correlation and Multivariate Multiple Regression in Exploring Interset Associations. *Psychological Bulletin* 104 (2),1988, 282–289.

Lebreton *et al.* 1991: J. Lebreton, R. Sabatier, G. Banco und A. Bacou, Principal Component and Correspondence Analyses with Respect to Instrumental Variables. An Overview of Their Role in Studies of Structure-Activity and Species-Environment Relationships. In: J. Devillers und W. Karcher (Hrsg.), *Applied Multivariate Analysis in SAR and Environmental Studies* (Dordrecht 1991) 85–114.

Legendre, Anderson 1999: P. Legendre und M. Anderson, Distance-Based Redundancy Analysis. Testing Multi-Species Responses in Multi-Factorial Ecological Experiments. *Ecological Monographs* 69 (1), 1999, 1–24.

Legendre, Gallagher 2001: P. Legendre und Eu. Gallagher, Ecologically Meaningful Transformations for Ordination of Species Data. *Oecologia* 129, 2001, 271–280. DOI 10.1007/s004420100716

Legendre *et al.* 2011: P. Legendre, J. Oksanen und C. ter Braak, Testing the Significance of Canonical Axes in Redundancy Analysis. *Methods in Ecology and Evolution* 2, 2011, 269–277. DOI 10.1111/j.2041-210X.2010.00078.x

Legendre, Legendre 2012: P. Legendre und L. Legendre, *Numerical Ecology.* Developments in Environmental Modelling 24. 3. Aufl. (Amsterdam 2012).

Maier 2015: A. Maier, *The Central European Magdalenian. Regional Diversity and Internal Variability* (Dordrecht 2015).

Marwick 2017: B. Marwick, Computational Reproducibility in ArchaeologicalResearch. Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory* 24, 2017, 424–450. DOI 10.1007/s10816-015-9272-9

McCune 1997: B. McCune, Influence of Noisy Environmental Data on Canonical Correspondence Analysis. *Ecology* 78, 1997, 2617–2623.

Noy-Meir *et al.* 1975: I. Noy-Meir, D. Walker und W. Williams, Data Transformation in Ecological Ordination. II. On the Meaning of Data Standardization. *Journal of Ecology* 63, 1975, 779–800.

Oksanen 2020: J. Oksanen, *Design Decisions and Implementation Details in vegan.* Erläuternder Begleittext (Vignette), beiliegend in: J. Oksanen, F. Guillaume Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. Minchin, R. O'Hara, Gavin Simpson, P. Solymos, M. Stevens, E. Szoecs und H. Wagner, *vegan. Community Ecology Package. R package version 2.5-7* (2020). URL: https://cran.r-project.org/package=vegan

Oksanen *et al.* 2020: J. Oksanen, F. Guillaume Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. Minchin, R. O'Hara, Gavin Simpson, P. Solymos, M. Stevens, E. Szoecs und H. Wagner, *vegan. Community Ecology Package. R package version 2.5-7* (2020). URL: https://cran.r-project.org/package=vegan

Orlóci 1967: L. Orlóci, An Agglomerative Method for Classification of Plant Communities. *Journal of Ecology* 55, 1967, 193–206.

Orlóci 1978: L. Orlóci, *Multivariate Analysis in Vegetation Research.* 2. Aufl. (The Hague 1978).

Palmer 1993: M. Palmer, Putting Things in Even Better Order. The Advantages of Canonical Correspondence Analysis. *Ecology* 74, 1993, 2215–2213.

Pearson 1901: K. Pearson, LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11), 1901, 559–572.

Rao 1964: C. Rao, The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhyā. The Indian Journal of Statistics, Series A* 26, 1964, 329–358.

Rao 1995: C. Rao, A Review of Canonical Coordinates and an Alternative to Correspondence Analysis Using Hellinger Distance. *Qüestiió. Quaderns d'Estadística i Investigació Operativa* 19 (1–3), 1995, 23–63.

R Core Team 2020: R Core Team, *R. A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2020. URL: https://www.R-project.org/

Schulte 2020: L. Schulte, „Massendinghaltung" oder „Big Data". *Praehistorische Zeitschrift* 95 (1), 2020, 206–247. DOI 10.1515/pz-2020-0030

Stevens 1946: S. Stevens, On the Theory of Scales of Measurement. *Science* 103 (2684), 1946, 677–680.

ter Braak 1986: C. ter Braak, Canonical Correspondence Analysis. A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology* 67, 1986, 1167–1179.

ter Braak 1994: C. ter Braak, Canonical Community Ordination. Part I. Basic Theory and Linear Methods. *Ecoscience* 1, 1994, 127–140.

ter Braak, Smilauer 2012: C. ter Braak und P. Smilauer, *CANOCO 5.0 Reference Manual and CanoDraw for Windows User's Guide. Software for Canonical Community Ordination* (Ithaca 2012).

Thioulouse *et al.* 2018: J. Thioulouse, St. Dray, A.-B. Dufour, A. Siberchicot und Th. Jombart, *Multivariate Analysis of Ecological Data with ade4* (New York 2018).

Wickens 1995: Th. Wickens, *The Geometry of Multivariate Statistics* (Hillsdale 1995).

Zerl 2019: T. Zerl, *Archäobotanische Untersuchungen zur Landwirtschaft und Ernährung während der Bronze- und Eisenzeit in der Niederrheinischen Bucht*. Rheinische Ausgrabungen 77 (Darmstadt 2019).