UNIVERSITÄT POTSDAM

Wirtschafts- und Sozialwissenschaftliche Fakultät

Hans Gerhard Strohe (Hrsg.)

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 43

Hannes-Friedrich Ulbrich

Höherdimensionale Kompositionsdaten
Gedanken zur grafischen Darstellung und Analyse

Potsdam 2010

ISSN 0949-068X

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 43

Hannes-Friedrich Ulbrich

Höherdimensionale Kompositionsdaten
Gedanken zur grafischen Darstellung und Analyse

Herausgeber: Prof. Dr. Hans Gerhard Strohe, Lehrstuhl für Statistik und Ökonometrie
Wirtschafts- und Sozialwissenschaftliche Fakultät
der Universität Potsdam
August-Bebel-Str. 89, D-14482 Potsdam
Tel. +49 (0) 331 977-3225
Fax. +49 (0) 331 977-3210
Email: strohe@uni-potsdam.de
2010, ISSN 0949-068X

Zusammenfassung

Kompositionsdaten sind mehrdimensionale Daten, deren Komponenten im Wesentlichen nur relative Informationen enthalten (und die sich deshalb meist zu einem festen Wert wie 1 oder 100 % addieren). Wegen ihres geschlossenen Charakters sind sie mit herkömmlichen Methoden (basierend auf einem n-dimensionalen Raum $\mathbb{R}^n$) nicht konsistent analysierbar. Methoden der Kompositionsdatenanalyse existieren seit etwa 30 Jahren, sie werden kurz vorgestellt.

Ein besonderes Problem ist die adäquate Darstellung von Kompositionsdaten. Für (bis zu) drei Komponenten gibt es verschiedene Methoden, für vier und mehr hingegen sind allen Komponenten gleichartig gerecht werdende Darstellungen kaum vorhanden. Ausgehend von den etablierten Methoden der Kompositionsdatenanalyse wird eine neue Darstellungsform vorgeschlagen, Vor- und Nachteile werden theoretisch and anhand von Beispielen diskutiert.

Inhaltsverzeichnis

1 Was ist Kompositionsdatenanalyse?

In der ersten Hälfte der 80er Jahre des 20. Jahrhunderts wurde in Hongkong untersucht, welcher Anteil des monatlichen Einkommens von Einpersonenhaushalten für (a) Wohnen einschließlich Heizung und Elektrizität, (b) Lebensmittel einschließlich Alkohol und Tabak, (c) Bekleidung und andere längerfristige Anschaffungen, (d) Inanspruchnahme von Dienstleistungen einschließlich öffentlichem Transport und eigenem Fahrzeugaufgewendet werden. (Die Kategorien sind eindeutig, einander ausschließend und erschöpfend.) In einer nach Geschlecht geschichteten Stichprobenerhebung unter alleinlebenden Einpersonenhaushalten in Mietwohnung wurden je 20 weibliche und männliche Bürger hinsichtlich der Ausgaben eines Monats in obigen Kategorien befragt (Aitchison, 2003, Kap. 1.7 & Appendix D). Alle Angaben erfolgten in HK-$. Kennt man nun für einen jeden Haushalt seine Gesamtausgaben in jenem Monat als Summe der Angaben zu (a) bis (d), realisiert man, dass jede der Komponenten in ihrem Wertebereich der Einschränkung unterliegt, weder negativ noch größer als die Gesamtsumme sein zu können. Mehr noch, mit der sukzessiven Kenntnis der Werte für die ersten Komponenten verringert sich der mögliche Wertebereich einer jeden verbleibenden weiterhin. Standardisiert man nun die Werte der Einzelkomponenten bezüglich der Gesamtausgaben, ergeben sich Anteilszahlen bzw. Quoten (Rönz and Strohe, 1994, S. 142) bezogen auf eine konstante Summe von 1 (oder 100 %).

Daten konstanter Summe werden im Englischen seit geraumer Zeit als *compositional data* bezeichnet (vgl. z. B. Aitchison, 1982). Noch viel länger (Pearson, 1897) bekannt ist ihre bemerkenswerteste Eigenschaft, zwischen ihren Komponenten einen negativ-verzerrenden Anteil an Korrelation zu erzwingen, der damals noch als *spurious* (hier in etwa: unberechtigt, störend) erschien. Heute weiß man, dass sich dieses direkt aus der Beschränktheit der Wertebereiche der Komponenten und deren Zusammenhang ergibt – für zwei Komponenten p und $1 - p$ gilt dies mit $\rho(p, 1 - p) = -1$ offensichtlich.

Erst etwa 60 Jahre nach Pearson (1897) wurden zu diesem Problem neue Erkenntnisse publiziert. Mit Chayes (1960) Artikel „On Correlation between Variables of Constant Sum“ begann die intensive Beschäftigung diesen Problemen. Treibende Kraft war (und ist bis heute) die Petrografie mit ihrem Interesse an der Beschreibung und Analyse der mineralischen Zusammensetzung von Gesteinen. Die bis heute wegweisende Publikation zur statistischen Analyse von *compositional data* lieferte ein Vierteljahrhundert später Aitchison (1986, 2003); Aitchison selbst wird nicht müde, „pragmatisch“ vereinfachenden Umgang mit *compositional data* kritisch zu beleuchten (s. z. B. Aitchison et al., 2001, und andere Beiträge in Mathematical Geology). Unter den bisher nur wenigen Zentren der Weiterentwicklung (und Propagierung) der Analyse von *compositional data* ragen die Arbeitsgruppen in Barcelona und Girona heraus (z. B. Egozcue and Pawlowsky-Glahn, 2006).

Bisher gibt zu *compositional data* es kaum Publikationen auf Deutsch, eine einheitliche Begriffsverwendung etabliert sich vergleichsweise langsam. Nach *Zusammensetzungsdaten* (van den Boogaart, 2008) und *Statistik von Zusammensetzungen* sowie *Zusammensetzungen / Kompositionen* (van den Boogaart, 2009) soll hier in Anlehnung an Pawlowsky-Glahn and Egozcue (2007) von der *Analyse von Kompositionsdaten* bzw. von *Kompositionsdatenanalyse* gesprochen werden.

2 Kompositionsdatenanalyse in der Wirtschaftsstatistik

Im aktuellsten Überblick zu Stand und Weiterentwicklung der Methoden der Kompositionsdatenanalyse erwähnen Aitchison and Egozcue (2005) die Anwendungsgebiete wie Geologie und Biologie nur insoweit, dass diese mit ihren Fragestellungen der Entwicklung der Kompositionsdatenanalyse Auftrieb verschafften. Aktuelle Anwendungen in der Ökonometrie werden ebensowenig erwähnt wie Anwendungen in Teilgebieten der Medizin, der Materialkunde, der Archäologie oder der Epidemiologie.

Ökonometrisch lässt sich mit den Daten der bereits erwähnten Beobachtungsstudie (Aitchison, 2003, Kap. 1.7) zu den Ausgaben von Einpersonenhaushalten untersuchen, ob für Männer und Frauen generell eine unterschiedliche Verteilung ihrer Ausgaben hinsichtlich der Quoten vier Kategorien (a) bis (d) aufweisen. Mehr noch, betrachtet man die jeweilige absolute Summe der Ausgaben (in HK-$) als Kovariate, kann die Frage adressiert werden, ob sich mit steigenden Gesamtausgaben die Quoten der vier Kategorien gegeneinander verändern, und ob auch dieses einem geschlechtsspezifischen Einfluss unterliegt. Beide Fragen dieser Querschnittsstudie sind von wirtschaftlichem Interesse.

Methoden der Zeitreihenanalyse sind klassischerweise bedeutsam für die Ökonometrie und die statistische Analyse von Wirtschaftsdaten. Obwohl weder Ravishanker et al. (2001) noch Aguilar Zuil et al. (2007) in ihrem Überblicksartikel direkt ökonometrische Fragestellungen bearbeiten, weisen beide bereits in der Einleitung auf die Bedeutung der Kompositionsdaten-Zeitreihenanalyse für die Ökonometrie hin. Mills (2009, 2010) untersucht neben der Zusammensetzung der Ausgaben in der britischen Volkswirtschaft auch die Veränderungen in der britischen Bevölkerung hinsichtlich Adipositas (Fettleibigkeit). Anhand der Kompositionsdaten-Zeitreihen werden Vorhersagen für die kommenden Jahre abgeleitet, die (zunehmende) Fettleibigkeit wird dabei als Problem der Gesundheit der Bevölkerung, als gesellschaftliche Erscheinung und in ihrer Bedeutung für Löhne, Einkommen und Wohlstand betrachtet.

Weitere Untersuchungen mit Kompositionsdaten finden sich bei Graf (2006) vom Bundesamt für Statistik der Schweiz. Sie betrachtet für die Jahre 2002 und 2004 die Einkommensstruktur der Schweizer Erwerbsbevölkerung mit fünf Komponenten, während Fry et al. (2000, 2001) vom

Department of Econometrics and Business Statistics der Monash University anhand der Daten des 1988/1989 Australian Household Expenditure Survey eine Kompositionsdatenanalyse mit 18 Warengruppen als Komponenten anstrengen. Sie versuchen dabei eine Lösung zu finden für die Analyse von Kompositionen, deren Einzelkomponenten einen Wert von 0 aufweisen können.

3 Methoden der Kompositionsdatenanalyse

Kompositionsdaten werden meist als Daten konstanter Summe definiert. Wie das Eingangsbeispiel der Anteile an den Haushaltsausgaben jedoch zeigt, ist eine Erweiterung auf Daten, die nur relative Informationen als Teil eines Ganzen tragen, sinnvoll (Filzmoser et al., 2009). Die gewonnenen relativen Informationen lassen sich bezüglich des jeweiligen Ganzen standardisieren und man erhält im Falle von D Komponenten Daten in einem Unterraum des D-dimensionalen Euklidischen Raumes $\mathbb{R}^D$:

$$S^D = \{x \in \mathbb{R}^D \mid x_i > 0,\ x_1 + \ldots + x_D = const\}$$

.

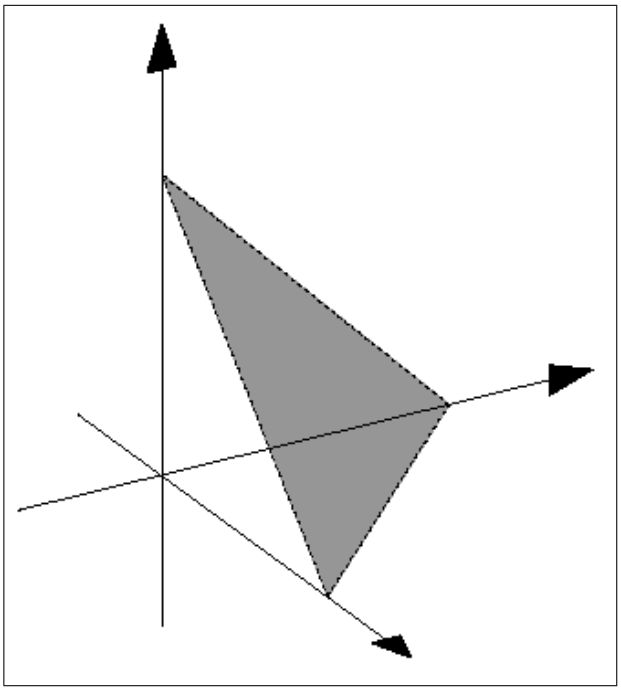


Abbildung 1: Simplex $S^3 \subsetneq \mathbb{R}^3$ – ein gleichseitiges Dreieck

Diesen wohldefinierten Unterraum S^D (der Dimension D-1) nennt man entsprechend der Dimension des Umgebungsraumes (und der Anzahl der Komponenten) ein D-dimensionales Simplex (seltener auch (D-1)-Simplex); niedrigdimensionale Beispiele sind eine Strecke als S^2, das gleichseitige Dreieck als S^3 (Abbildung 1) und als S^4 das gleichseitige Tetraeder. Im strengen Sinne obiger Definition sind diese Simplizes topologisch offene Räume, Ränder sind nicht Teil des Simplex. Ein geschlossener Simplex schließt seine Ränder mit ein, die Definitionsbedingung $x_i > 0$ wird zu $x_i \geq 0$ erweitert.

Zu Beginn der 80er Jahre des vergangenen Jahrhunderts stellte Aitchison (1982, 1986) seine völlig neue Herangehensweise an die Analyse von Kompositionsdaten vor. Kompositionsdaten lassen sich als reellwertige Vektoren mit positiven Komponenten darstellen, die ein Simplex S^D im D-dimensionalen Raum aufspannen. Ein Simplex S^D ist somit ein begrenzter (D-1)-dimensionaler Teilraum des reellen D-dimensionalen Raumes $\mathbb{R}^D$.

Da sich die relativen Informationen sich bezüglich des jeweiligen Ganzen standardisieren lassen, kann jeder reellwertige Vektor einer Äquivalenzklasse zugeordnet werden, der Konstantsummenwert ist unerheblich, alle Äquivalenzklassen finden sich in einem Simplex S^D zur Konstantsumme 1 repräsentiert (Abbildung 2).

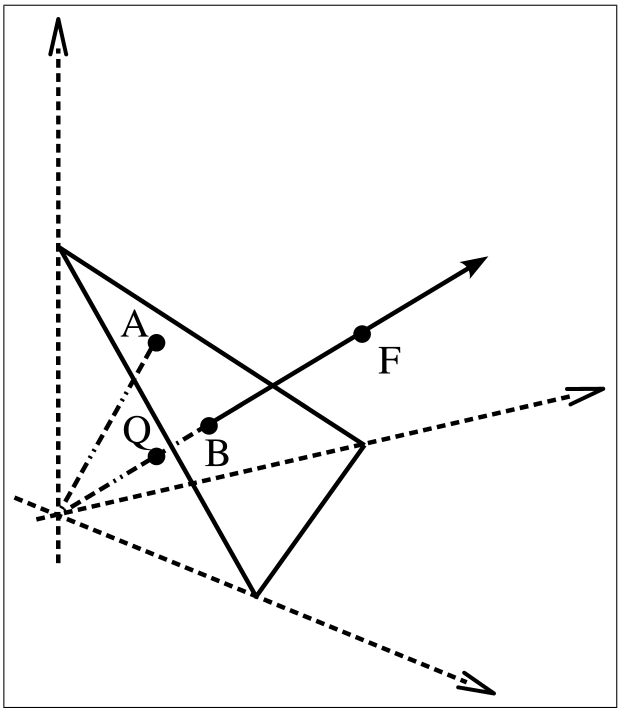


Abbildung 2: Simplex S^3 als Äquivalenzklasse:
der Vektor mit F, Q (und B selbst) wird durch B repräsentiert

Die Punkte $\vec{x} = (x_1, \ldots, x_D)^T$ eines Simplex S^D wiederum kann man durch geeignete ein-

eindeutige Transformationen in Punkte eines (D-1)-dimensionalen unbegrenzten Raumes Euklidischer Struktur überführen, nennen wir diesen einen Koordinatenraum. Drei Transformationstypen wurden dafür bisher vorgeschlagen:

$$\mathbf{alr}(\vec{x}) = \left(ln\left(\frac{x_1}{x_D}\right), \ldots, ln\left(\frac{x_{D-1}}{x_D}\right)\right)^T = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \cdot ln\,(\vec{x})$$

$$\mathbf{clr}(\vec{x}) = \left(ln\left(\frac{x_1}{\mathrm{g}(\vec{x})}\right), \ldots, ln\left(\frac{x_D}{\mathrm{g}(\vec{x})}\right)\right)^T = \frac{1}{D \cdot \mathrm{g}(\vec{x})} \cdot \begin{pmatrix} D-1 & 1 & \cdots & 1 \\ 1 & D-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & D-1 \end{pmatrix} \cdot ln\,(\vec{x})$$

$$\mathbf{ilr}(\vec{x}) = \mathbf{V} \cdot \mathbf{clr}(\vec{x}) = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{D(D-1)}} & \frac{1}{\sqrt{D(D-1)}} & \frac{1}{\sqrt{D(D-1)}} & \cdots & -\frac{D-1}{\sqrt{D(D-1)}} \end{pmatrix} \cdot \mathbf{clr}(\vec{x})$$

Die Transformationen **alr** (additive logratio) und **clr** (centered logratio, mit dem geometrischen Mittel aller Komponentenwerte $\mathrm{g}(\vec{x}) = \mathrm{g}(x_1, \ldots, x_D) = \sqrt[D]{x_1 \cdots x_D}$ im Nenner) gehen auf Aitchison (1982) zurück, **ilr** (isometric logratio) wurde erstmals von Egozcue et al. (2003) vorgestellt. Alle drei Transformationen basieren auf dem Logarithmieren; zum dabei möglicherweise auftretenden Problem mit Nullwerten in einzelnen Komponenten der Kompositionen siehe z. B. Martín-Fernández and Thío-Henestrosa (2006).

Während **alr** und **ilr** ein Simplex S^D direkt in den $\mathbb{R}^{D-1}$ als Koordinatenraum abbilden, wird mittels **clr** in eine durch den Ursprung des $\mathbb{R}^D$ verlaufende Hyperebene der Dimension D-1 transformiert, so dass jeder Punkt im Simplex in Koordinaten des $\mathbb{R}^D$ überführt werden, deren Summe 0 ergibt.

Weder **ilr** noch **alr** sind eindeutig definiert, es handelt sich eher um Klassen von Transformationen. Für die **ilr**-Transformation wurde als Beispiel die Transformation über eine (um die erste Zeile verminderte) Helmert-Matrix $\mathbf{V}$ der Dimension $D\times(D\text{-}1)$ (Harville, 1997, S. 86) angegeben. Der Vorteil der **ilr**- gegenüber der **alr**-Transformation besteht darin, dass mittels **ilr** in einen $\mathbb{R}^{D-1}$ mit orthonormaler Basis überführt wird. Jede orthonormale Rotation einer **ilr**-Transformation mittels einer orthonormalen Matrix $\mathbf{A}$

$$\mathbf{V}' \cdot \mathbf{clr}(\vec{x}) = \mathbf{A} \cdot \mathbf{V} \cdot \mathbf{clr}(\vec{x})$$

ist somit selbst wieder **ilr**-Transformation. In der Klasse der **alr**-Transformationen kann jede Komponente die Rolle der Referenzkomponente (hier: x_D) einnehmen.

Jede konkret ausgewählte Transformation aus der Klasse der **ilr**- bzw. **alr**-Transformation ist wie die **clr**-Transformation ein-eindeutig: Zu jeder der Transformationen existiert eine inverse Transformation vom $\mathbb{R}^{D-1}$ zurück in das Simplex S^D (bzw. die jeweiligen Äquivalenzklasse).

Die Zeilen der verminderten Helmert-Matrix $\mathbf{V}$ selbst bilden ein orthonormales Koordinatensystem im Simplex S^D.

Aitchison (1982, 1986, und Folgepublikationen) zeigen, dass statistische Analysen von Kompositionsdaten im Koordinatenraum (der transformierten Daten) mit Hilfe der (üblichen) multivariaten Verfahren durchgeführt werden können. Mehr noch, der Koordinatenraum $\mathbb{R}^{D-1}$ ist ein Euklidischer Vektorraum (zu Statistik in Vektorräumen siehe Eaton, 2007, der neueren Ausgabe von Eaton, 1983). Mit den (Inversen der) ein-eindeutigen Transformationen überträgt sich die Vektorraum-Struktur auf den Simplex S^D. Die Geometrie dieses Vektorraumes im Simplex S^D wird Aitchison-Geometrie genannt, die Skalarmultiplikation und die Vektoraddition werden mit „powering“ und „perturbation“ bezeichnet. Das Skalarprodukt zweier Vektoren (Kompositionen) und der Abstand zwischen zwei Vektoren sind wohldefiniert Aitchison (2003, Postscript 2003, S. 3-6); entsprechende Verschiebungen von Punktewolken z. B. in das Zentrum eines Koordinatensystems verändern die Beziehungseigenschaften der Punkte untereinander nicht.

Das Zentrum einer Verteilung im Simplex ist durch das geometrische – und nicht das arithmetische – Mittel der Einzelkompositionen $\vec{x}_j = (x_{j,1}, \ldots, x_{j,D})$ abzuleiten. Für das theoretische Pendant, den Erwartungswert, gilt (mittels der **clr**-Transformation formuliert)

$$\mathrm{cen}(\vec{x}) = \mathrm{clr}^{-1}(\mathrm{E}[\mathrm{clr}(\vec{x})]).$$

Modell-Prädiktionen im Koordinatenraum können mittels der Invers-Transformationen in das Simplex des Stichprobenraumes zurückgeführt werden. Gegenüber vielen, auch heute noch propagierten Alternativ-Ansätzen (wie z. B. Srivastava et al., 2007), sichert der Aitchison-Ansatz, dass Vorhersagen und Konfidenzbereiche außerhalb des Stichprobenraumes – des Simplex S^D – unmöglich sind.

Eine Subkomposition einer Komposition $\vec{x} \in S^D$ besteht aus d ($1 \leq d \leq D$) Komponenten, während Amalgamation die additive Zusammenfassung zweier (oder mehrerer) Komponenten einer Komposition bedeutet (Aitchison, 2003, Abschnitte 2.5 & 2.6). In beiden Fällen sind die Ergebnisse Kompositionen in einem Simplex S^d ($1 \leq d \leq D$). Eine wichtige Eigenschaft dabei ist „subcompositional coherence“ (Pawlowsky-Glahn et al., 2007, Abschnitt 2.2.3): die Subkomposition entspricht einer Projektion in einen niedriger-dimensionalen Teilraum des ursprünglichen Simplex, der Abstand zweier projizierter Punkte ist (nicht-negativ und) nicht größer als der Abstand dieser Punkte im ursprünglichen Raum.

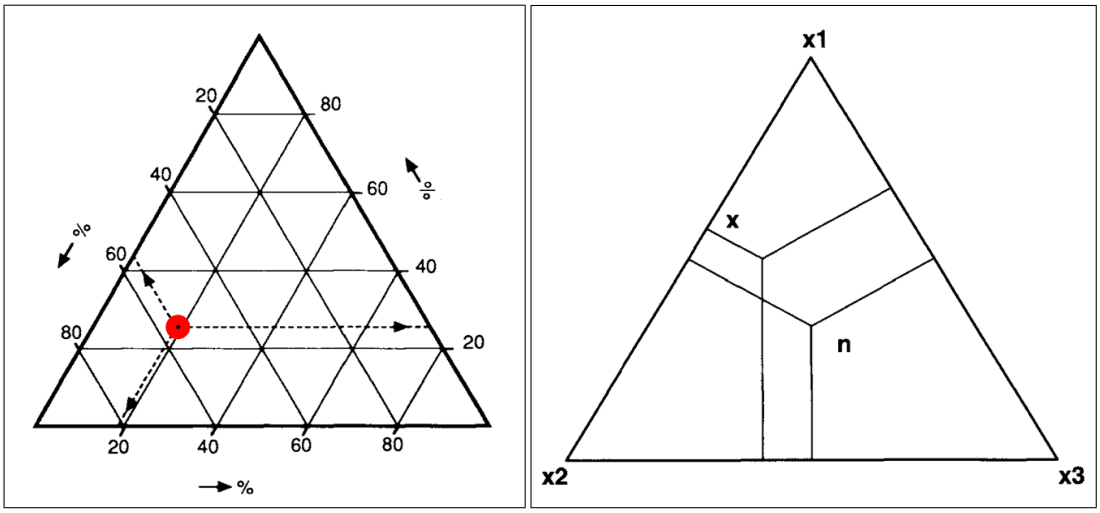
4 Grafische Darstellung von Kompositionsdaten

Adäquate grafische Darstellungen von Kompositionsdaten sollten widerspiegeln, dass es sich um Datenpunkte in einem Simplex S^D handelt. Darstellungen im unbeschränkten $\mathbb{R}^D$ erscheinen hier generell nicht als erstrebenswerte Lösung.

Für den Zwei-Komponenten-Fall S^2 verstanden als Raten existieren etablierte Darstellungsformen, z. B. gestapelte Säulen- oder Balkendiagramme und Kurven der logistischen Regression. Beide Komponenten ergänzen sich zu Balken konstanter Länge und werden darstellungstechnisch gleichwertig behandelt; tauscht man die Komponenten, ergibt sich ein Diagramm welches sich symmetrisch von ersterem unterscheidet.

Für Kompositionsdaten von 3 oder mehr Komponenten dargestellt im Balkendiagramm gilt diese Symmetrie nicht.

Kompositionen dreier Komponenten lassen sich in Dreiecksdiagrammen (engl.: *ternary plot*) visualisieren (z. B. Egozcue and Pawlowsky-Glahn, 2006), die gleichseitigen Dreiecke garantieren eine visuelle Gleichwertigkeit der Komponenten bei Drehungen um 120° (siehe Abbildung 3).



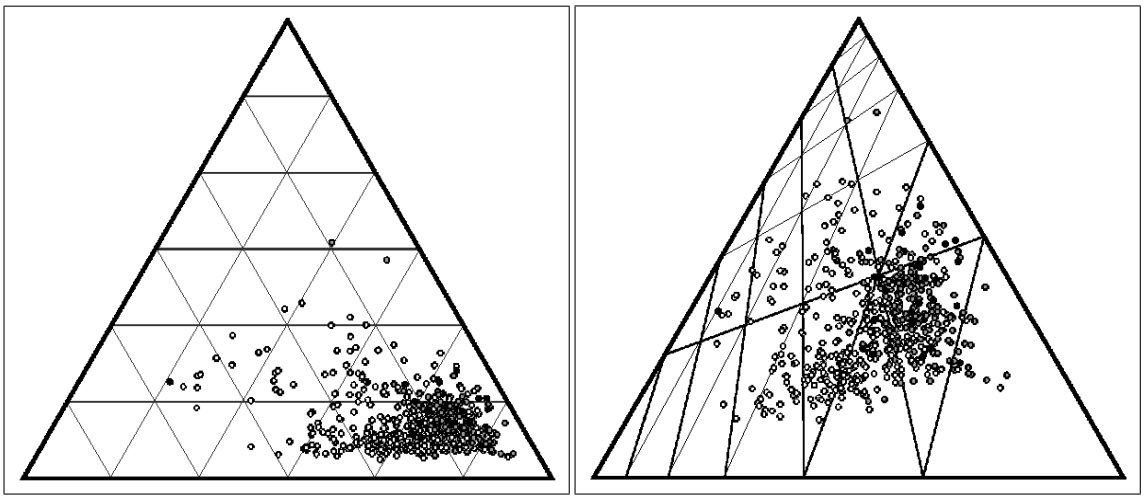
(a) (b)

Abbildung 3: S^3 – Kompositionen im Dreiecksdiagramm
(a) zu 100 % mit Hilfslinien
(b) $x = (1/2, 1/3, 1/6)$ und $n = \frac{1}{3}(1, 1, 1)$ – das Baryzentrum des Simplex

Abbildung 4 zeigt für eine Menge von Datenpunkten (Kompositionen dreier Komponenten) deren Lage im Dreiecksdiagramm[1]. Teilbild (a) zeigt, dass sich wegen überwiegend hoher Werte

[1] Quelle: The American Statistician 55: 214–217, 2001

einer Komponente die Datenpunkte hauptsächlich in einer Ecke des Dreiecksdiagramms wiederfinden, eine Diskriminierung zwischen einzelnen Punkten ist nicht gut möglich. Teilbild (b) hingegen zeigt dieselbe Punktewolke in Richtung Mitte des Dreiecks verschoben. Die Verschiebung erfolgte so, dass sich das empirische Zentrum der Verteilung nun genau im Baryzentrum des Dreiecksdiagramms befindet, die Punkte der Datenwolke sind deutlich besser voneinander zu unterschieden. Diese Verschiebung verändert die Verhältnisse zwischen den Datenpunkten nicht, sie entspricht einfach einer üblichen, orthogonalen Verschiebung im zugehörigen Koordinatenraum (z. B. Transformation per **ilr**, Verschiebung des Koordinatensystems – hier in das Zentrum der Verteilung –, Rücktransformation per $\mathbf{ilr}^{-1}$). Dass sich dieses Vorgehen auf die Referenzwerte-Einteilung an den Rändern des Dreiecksdiagramms auswirkt, muss beim Betrachten des Diagramms zusätzlich beachtet werden.

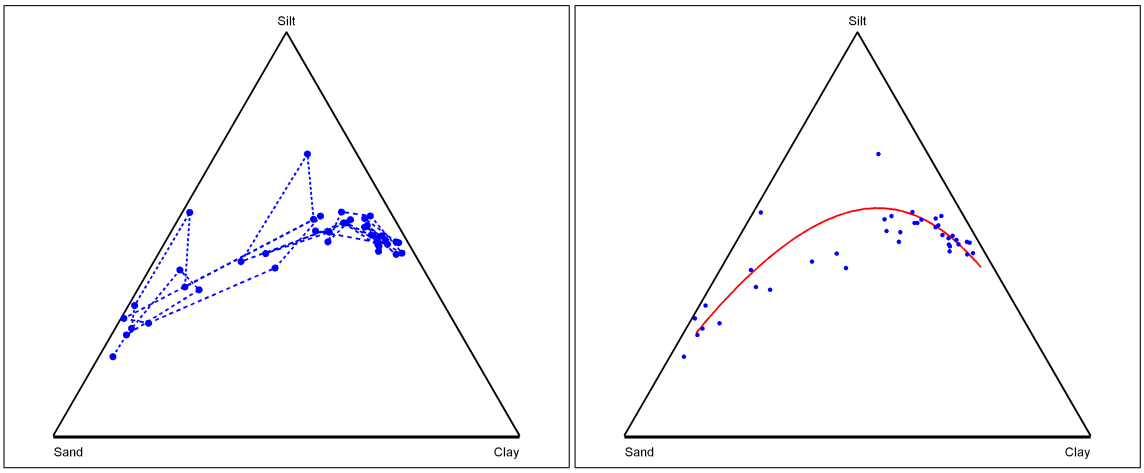


(a) (b)

Abbildung 4: S^3 – Dreiecksdiagramm (Hilfslinien alle 16,6 %)
(a) Originalwerte
(b) nach Verschiebung in das Baryzentrum des Simplex (siehe Text)

Mit Dreiecksdiagrammen lassen sich somit Kompositionen dreier Komponenten darstellen, allerdings ist bereits deren Darstellung erweitert um eine zusätzliche (Einfluss-) Variable auf kategorieller, ordinaler oder metrischer Skala nur sehr eingeschränkt möglich. Aitchison (1986, Fig. 1.3) zeigt eine perspektivische Abfolge von Dreiecksdiagrammen, die man sich auch als eine Sequenz nebeneinanderliegender Diagramme vorstellen kann – in beiden Fällen sind kleine Veränderungen kaum und in ihrer Richtung nicht sicher wahrnehmbar. Alternativ können (bei nicht zu großer Anzahl an Punkten) die Dreiecksdiagramme übereinander gelegt werden (siehe Abbildung 5 zu Daten aus Aitchison, 1986). Man sieht, dass mit der Veränderung der Tiefe des Sees eine Veränderung in der Zusammensetzung (Komposition) der Ablagerungen am Grunde des Sees einhergeht, weder die Tiefenwerte selbst – hier logarithmiert als metrisches Merkmal

und Regressor – noch die Richtung (rechts sind die größeren Tiefen) sind ersichtlich.



(a) (b)

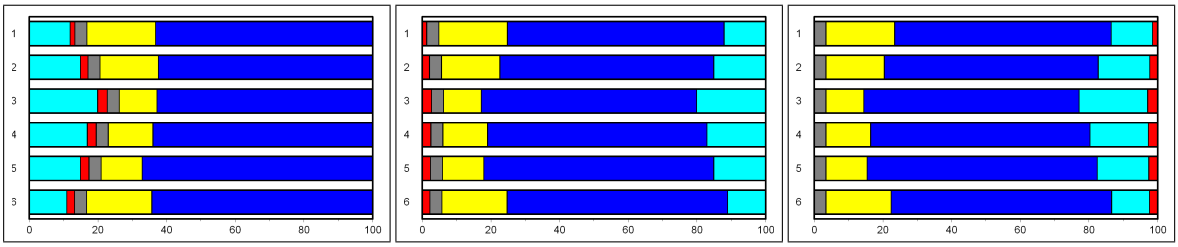
Abbildung 5: Sedimente als Kompositionen aus Sand, Schluff (silt) und Lehm (clay) in verschiedenen Tiefen eines arktischen Binnensees
(a) Messdaten (verbunden in Reihenfolge der Tiefe unter Wasseroberfläche)
(b) Regression mit log(Tiefe) als Regressor (siehe Text)

Für Komponentenanzahlen größer als 3 nehmen die Schwierigkeiten beim Versuch der Visualisierung zu. Während für vier-komponentige Kompositionen eine dynamisch-räumliche Darstellung des Tetraeders als Abbild des Simplex S^4 zumindest vorstellbar erscheint, ist dieser Weg für Komponentenanzahlen von 5 und darüber nicht gangbar.

Einzel-Kompositionen können jederzeit mit einem Kuchendiagramm veranschaulicht werden, für eine Sequenz von Kuchendiagrammen gilt jedoch das für eine Sequenz nebeneinanderliegender Dreiecksdiagramme bereits gesagte: kleine Veränderungen sind kaum und in ihrer Richtung nicht sicher wahrnehmbar. Außerdem ist in einer Folge von Kompositions-Kuchendiagrammen immer nur eine Richtung (die 0°-Achse) fixierbar, nur die beiden anliegenden Komponenten haben im Diagramm eine von den anderen Komponenten unbeeinflusste Position. Somit sind, obwohl kreisrund, Kuchendiagramme in einer Sequenz bezüglich der einzelnen Komponenten nicht gleichwertig.

Häufiger als Kuchendiagramme werden (Sequenzen von) gestapelten Säulen- oder Balkendiagrammen zur Veranschaulichung von Kompositionsdaten herangezogen. Hier ist ebenfalls bereits ab 3 Komponenten eine völlige Gleichbehandlung der Komponenten nicht mehr möglich. Abbildung 6 zeigt das Problem anhand der Daten aus Tabelle 1: sechs Kompositionen zu 100 %, dabei seien $z1$ bis $z6$ beliebige unterschiedliche Ereignisse, möglicherweise über ein metrisches Merkmal miteinander verknüpft. Die Teilbilder unterscheiden sich nur in der Ordnung der Kom-

ponenten zueinander. Teilbild (a) ist zu entnehmen, dass der Anteil der größten Komponente (5) ein wenig um 63 % mit einem Maximum bei $z5$ schwankt; Komponenten 2 und 3 sind in ihrer jeweiligen Veränderung kaum beurteilbar. Das ändert sich für Komponente 2 in Teilbild (b) durch den Positionswechsel an den (linken) Rand des Diagramms – hier zuungunsten der Sichtbarkeit der Komponente-5-Veränderungen. Jedoch zeigt erst Teilbild (c), dass die Werte für Komponente 3 keinerlei Veränderungen unterliegen. Mit dem Anstieg der Anzahl der Komponenten vergrößert sich das Problem, dass nur die randständigen Komponenten wegen ihrer fixierten Lage gut hinsichtlich ihrer Veränderungen beurteilbar sind. Handelt es sich bei den Komponenten um ein System mit innerer Rangordnung von Komponente 1 „klein“ zu Komponente 5 „groß“ verletzte ein Positionswechsel ziemlich die Forderung nach Klarheit eines Diagramms.



(a) (b) (c)

Abbildung 6: Kompositionen zu 100 % (Tab. 1) – Reihenfolge der Komponenten
(a) 1–2–3–4–5 (b) 2–3–4–5–1 (c) 3–4–5–1–2

Alternativ zum gestapelten Säulen- oder Balkendiagramm findet man vielfach die überlagerte Darstellung einer jeden Komponente wie in Abbildung 7 für den Beispieldatensatz. Handelt es sich bei $z1$ bis $z6$ um beliebige Kategorien zur Identifikation, so können die Verbindungen lediglich als Hilfe für die optische Zuordnung verstanden werden. Häufiger handelt es sich bei $z1$

#	Komponenten [%]				
z1	12	1.4	3.5	20	63.1
z2	15	2.2	3.5	17	62.3
z3	20	2.8	3.5	11	62.7
z4	17	2.6	3.5	13	63.9
z5	15	2.5	3.5	12	67.0
z6	11	2.3	3.5	19	64.2

Tabelle 1: Kompositionen zu 100 % (Beispieldaten)

bis $z6$ jedoch um Werte eines metrischen Merkmals (z. B. in äquidistantem Abstand), geeignete Modellierung wie z. B. in Buccianti and Pawlowsky-Glahn (2005, Fig. 12) führt dann zu glatten Funktionen als Anpassung an die Daten.

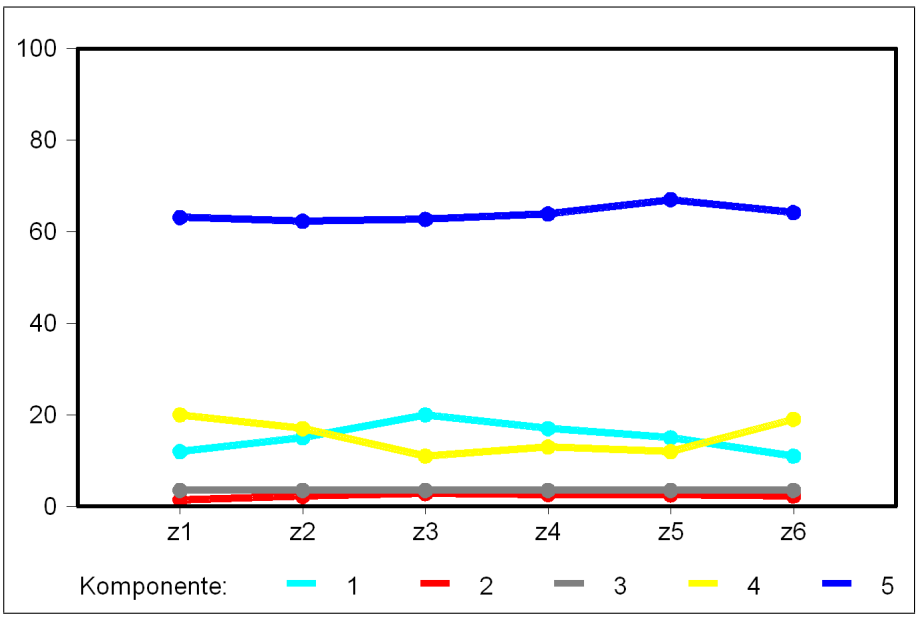


Abbildung 7: Kompositionen zu 100 % (zu Tab. 1)

Abbildung 7 zeigt deutlich, dass Komponente 5 wenig Veränderung erfährt, Komponenten 1 und 4 zeigen gegenläufiges Verhalten. Komponenten 2 und 3 werden kaum wahrgenommen, obwohl Komponente 2 ihren Anteil zwischenzeitlich „verdoppelt“ (von 1.4 % auf 2.8 %) – natürlich auf „Kosten“ aller anderen, deren gemeinsamer Anteil von 98.6 % auf 97.2 % sinkt.

5 Das Kompositionsabweichungsdiagramm – eine neue Darstellungsmöglichkeit für Kompositionsdatenanalyse

Grafische Darstellung von Kompositionsdaten so dass jede der Komponenten gleich behandelt wird, ist somit für bis zu 3 Komponenten möglich. Da nicht jede auf Kompositionsdaten basierende Fragestellung durch entweder Amalgamation auf 3 Komponenten oder durch (sukzessive) Analyse von 3er Subkompositionen ausreichend gut beantwortet werden kann, bleibt die Frage nach Darstellungen für Kompositionsdaten höherer Komponentenanzahl.

Vorgeschlagen wird dafür ein Diagramm auf Basis der **clr**-Transformation

$$\mathbf{clr}(\vec{x}) = \frac{1}{D \cdot \mathrm{g}(\vec{x})} \cdot \begin{pmatrix} D-1 & 1 & \cdots & 1 \\ 1 & D-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & D-1 \end{pmatrix} \cdot ln\,(\vec{x})\,;$$

es wird im folgenden Kompositionsabweichungsdiagramm (engl.: *compositional deviation plot*) genannt.

Das Kompositionsabweichungsdiagramm nutzt die Eigenschaften der **clr**-Transformation: je Komponente aus S^D gibt es eine Koordinate im $\mathbb{R}^D$ und diese Koordinaten-Werte addieren sich zu 0 (sie liegen auf einer entsprechenden Hyperebene der Dimension (D-1)). Das Diagramm zeigt für jede Kompositionen die **clr**-transformierten Werte einer jeden Komponenten auf der Ordinate, während die Abszisse Raum lässt für die Anordnung mehrere Kompositionen über einem interessanten Einflussmerkmal, sei dieses eine (Einfluss-) Variable auf kategorieller, ordinaler oder metrischer Skala. Der „Verlauf“ der einzelnen Komponenten über dem Einflussmerkmal lässt sich leicht hinsichtlich zweier Eigenschaften begutachten: der mittleren Lage bezogen auf das (jeweilige) geometrische Mittel und der Veränderungen (Anstieg, Abfall oder „Konstanz“) gegenüber den anderen Komponenten.

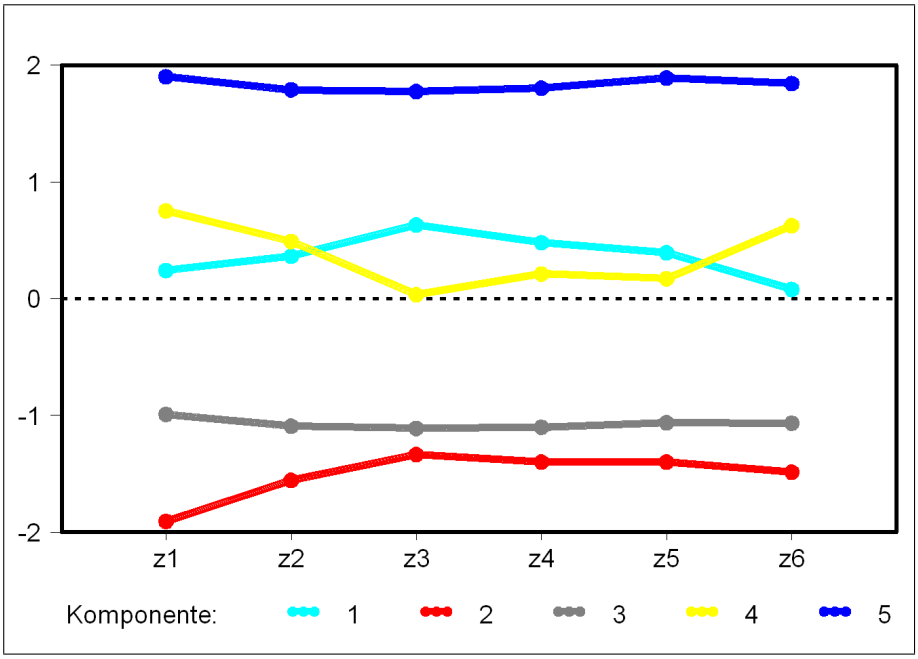


Abbildung 8: Kompositionsabweichungsdiagramm (zu Tab. 1)
(Verbindungen zwischen den Punkten für bessere Lesbarkeit)

Abbildung 8 zeigt, dass Komponente 5 die anteilsmäßig bedeutendste Komponente ist und dass

die Veränderungen in Komponente 5 relativ unbedeutend scheinen gegenüber dem deutlichen Anstieg Komponente 2 zwischen $z1$ und $z2$. Die Veränderungen in Komponente 2 zeigen sich so bedeutsam, wie die in den Komponenten 1 und 4. Bemerkenswert ist auch, dass die „Konstanz“ der Werte der Komponente 3 im Bezug auf die anderen Komponenten keine Konstanz ist.

Was auch immer man Abbildung 7 entnommen hat, das Kompositionsabweichungsdiagramm in Abbildung 8 zeigt, was die Kompositionsdatenanalyse analysiert: in einen linearen Raum der Dimension D-1 transformierte Punkte des Simplex S^D.

Da sich

$$\mathbf{ilr}_{(\mathbf{A}\cdot\mathbf{V})}(\vec{x}) = \mathbf{A}\cdot\mathbf{V}\cdot\mathbf{clr}(\vec{x})$$

aus der Klasse der **ilr**-Transformationen als orthonormale Drehung der **clr**-Transformations-Hyperebene $\mathbf{1}^T\cdot\mathbf{clr}(\vec{x}) = 0$ erweist, spiegeln die **clr**-Koordinaten die Kompositionen im Simplex genauso unverzerrt wider wie die **ilr**-Koordinaten. Auf diesen Gedanken aufbauend lassen sich nach Modellierungen zum jeweiligen Kompositionsabweichungsdiagramm Residuendiagramme ableiten – der Informationsfülle wegen ggf. für jede Komponente separat gezeichnet.

Nachteile des Kompositionsabweichungsdiagramms könnten in den Augen des (substanzwissenschaftlichen) Betrachters vor allem darin liegen, dass transformierte Daten auf arbiträr erscheinender Skala dargestellt werden. Das menschliche Auge sieht additiv, die Situation im Simplex kann wegen der Konstantsummeneigenschaft additiv nicht sein. Datendeskription und -Modellierung im Koordinatenraum lassen jedoch sinnvollerweise additiv gestalten, das Kompositionsabweichungsdiagramm kann hilfreich sein bei der adäquaten Darstellung der den Daten innewohnenden Beziehungen, die datengesteuert willkürlich erscheinende Skalierung der Ordinate nimmt man in Kauf.

6 Beispiele mit höherdimensionalen Kompositionsdaten

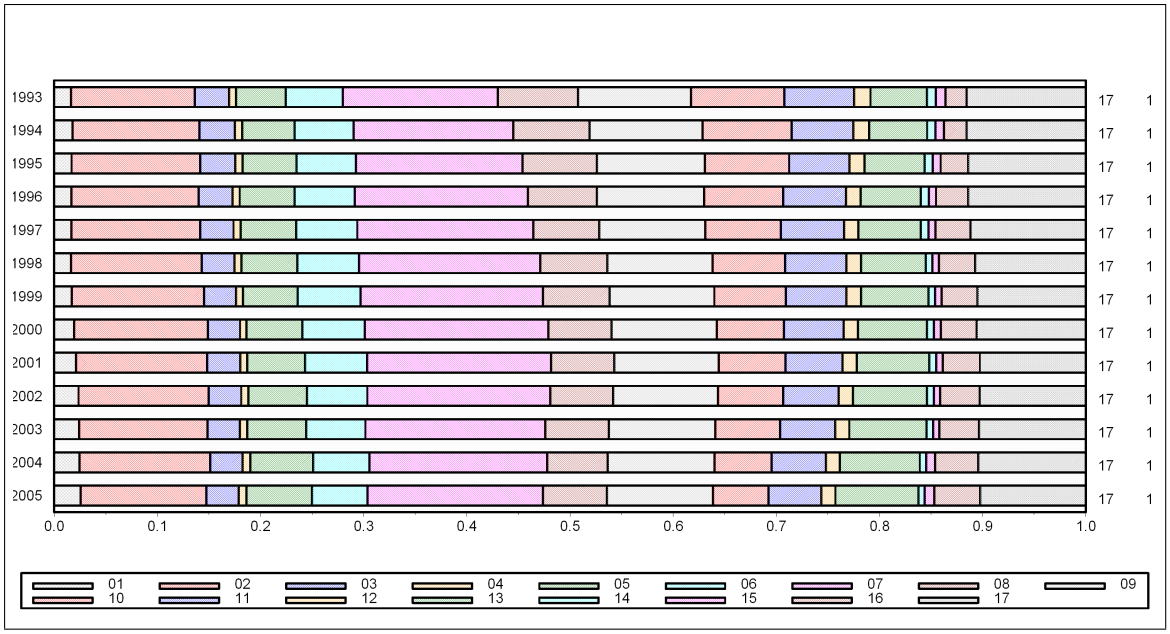
Für eine Untersuchung zu Veränderungen der wirtschaftlichen Situation der ost- und westdeutschen Krankenhäuser während der Transition nach der Wiedervereinigung wurden die Daten der Krankenhausstatistik herangezogen. Die Krankenhausstatistik ist seit 1991 für alle Krankenhäuser Deutschlands obligatorisch und umfasst drei Teile: Grunddaten (Sitz des Krankenhauses, personelle und materielle Ausstattung, Anzahl der Fälle) und Kostendaten werden von Anfang an erhoben, Diagnosedaten kommen mit dem Jahr 1993 hinzu. Die Diagnosedaten enthalten für jeden Fall (d. h. Aufenthalt eines Patienten) die Hauptdiagnose, die Behandlungstage und als demografische Daten Alter, Geschlecht und Wohnort); Neueinweisung, wiederholter Aufenthalt selbst in dasselbe Krankenhaus machen den Patienten zu einem neuen Fall.

Die Erfassung erfolgt jährlich und bundeslandweise, die Daten werden in anonymisierter Form durch das FDZ der Statistischen Landesämter für die Forschung zur Verfügung gestellt. Die erwähnte Untersuchung basiert auf den Daten der ostdeutschen Bundesländer einschließlich Berlins und denen des Landes Rheinland-Pfalz als Altbundesland – zum Vergleich – für die Jahre 1991 bis 2005. Diagnosedaten umfassen somit den 13-Jahreszeitraum von 1993 bis 2005.

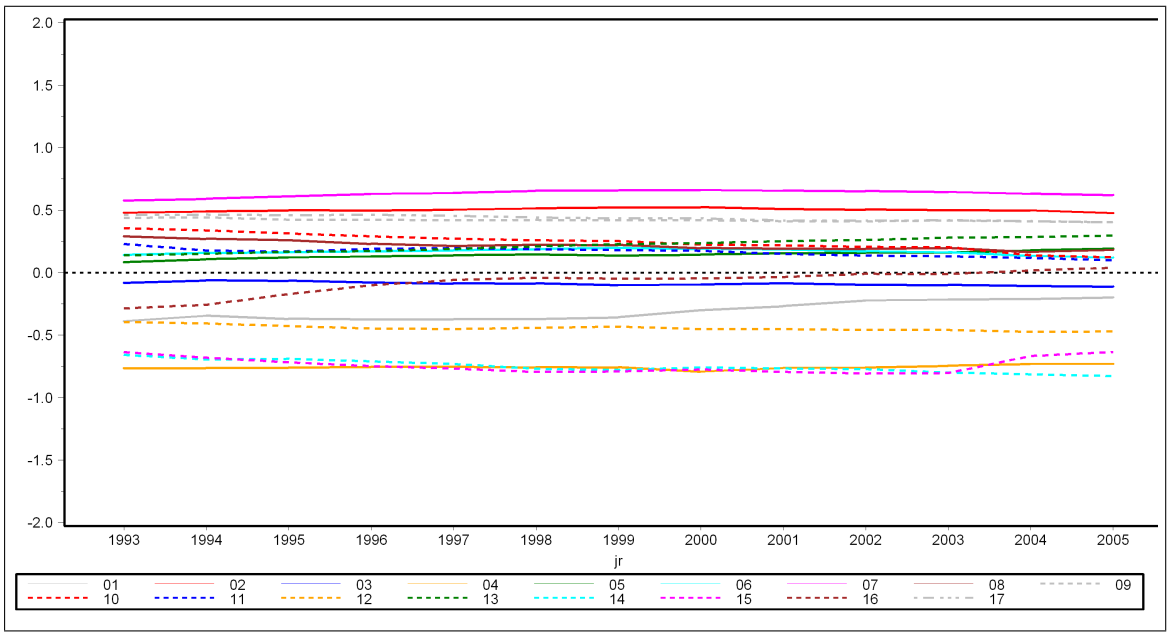
Neben den Transitionserscheinungen in Ostdeutschland ist dieser Zeitraum auch geprägt von den häufigen gesetzlichen Änderungen der Rahmenbedingungen für das Gesundheitswesen mit den Zielen Effizienzsteigerung und Kostenbegrenzung. Das führte auch in der Krankenhausstatistik zu Veränderungen: Bis 1995 wurden die Kosten nach dem Bruttoverfahren erhoben, für die Jahre 1996 bis 2001 wurde das Nettoverfahren angewendet, 2002 kehrte man zum Bruttoverfahren zurück. Mit dem Übergang ins Jahr 2002 wurden die Erfassungsbögen aktualisiert, dabei wurden mehrere Fragen hinzugenommen, einige verändert und andere weggelassen. Die Umstellung des Schlüssels für die Kodierung der Hauptdiagnose wurde für das Jahr 2000 angewiesen, ICD-10 ersetzte ICD-9, ein alphanumerischer Schlüssel in 21 Kapiteln ersetzte ein (i. W.) numerischen Schlüssel in 17 und 2 Ergänzungs-Kapiteln; dafür war ebenfalls eine Umstellung der Erfassungsbögen nötig. Nach langer, mit Verweigerung seitens wichtiger zu Beteiligender gestrafter und kontrovers geführter Diskussion nahmen 2003 überraschenderweise 50 % der Krankenhäuser an der neuen DRG-Statistik teil, bevor diese 2004 für alle Krankenhäuser (mit Ausnahme psychiatrischer Einrichtungen) verbindlich wurde.

Die DRG-Statistik (G-DRG: Diagnose Related Groups – German Version) ist Teil der amtlichen Statistik und ebenfalls bei den Bundesländern angesiedelt, ein Teil der Krankenhausstatistik ist sie nicht. Sie ist ein wichtiges Mittel bei der Änderung der Krankenhausfinanzierung – weg von einem Budget-basierten hin zu einem Fallpauschalensystem. Eine Vergütung der Krankenhäuser (oder jeglicher medizinischer Leistung) ist nach einem streng medizinischen Diagnosesystem nicht möglich, schon gar nicht, wenn für den Krankenhausaufenthalt „nur“ die Hauptdiagnose festgehalten wird. Im Fallpauschalensystem wird jeder Fall nach der für seine DRG-Einstufung vorgesehenen Fallpauschale vergütet. Für eine Übergangszeit beginnend 2003 war vorgesehen, die DRG-Einstufungen zu erheben und unter Kostenneutralität für das Gesamtgesundheitswesen jeder DRG-Einstufung eine Vergütung – genannt Basisfallwert – zuzuordnen. Aktuell (2010) versucht man, die gewonnenen bundeslandspezifischen Basisfallwerte zu bundeseinheitlichen konvergieren zu lassen. Ein enger Bezug der DRG-Einstufung zu den jeweils zugrunde liegenden Krankheitsbildern ist sinnvollerweise anzunehmen.

Soweit solcherart Veränderungen die Krankenhausstatistik (oder die DRG-Statistik oder einen andere amtliche Statistik) betreffen, wird dieses von Seiten der amtlichen Statistik vorab kund getan. Trotzdem gibt es Krankenhäuser, denen es gelang (und Statistische Landesämter, die dieses zuließen), auch nach 1999 Angaben zu Diagnosen nach ICD-9 zu verschlüsseln.



(a)



(b)

Abbildung 9: 17 Kapitel der ICD-9 Klassifikation für Hauptdiagnosen im Krankenhaus
(a) (gestapelte) Balkendiagramme
(b) Kompositionsabweichungsdiagramm (siehe Text)

Vom Deutschen Institut für Medizinische Dokumentation und Information (DIMDI) in Köln gibt es jährlich aktualisierte Ausgaben der gültigen Versionen der Schlüssel ICD-9 und ICD-10. Basierend auf den Angaben des DIMDI ist es möglich, eine Rekodierung von ICD-10-verschlüsselten Angaben auf ICD-9 vorzunehmen, dieses wurde für die erwähnte Untersuchung getan.

Krankheiten werden medizinischerseits auf verschiedenen Hierarchiestufen gruppiert, dieses spiegelt sich in den Kapiteln und Unterkapiteln der ICD-Kodierungssysteme wider. Ein Automatismus sinnvoller Zusammenfassung von Krankheiten entsteht damit bekanntermaßen jedoch nicht. (So wurden aus dem Kapitel VI der ICD-9 Krankheiten des Nervensystems und der Sinnesorgane drei der ICD-10: VI Krankheiten des Nervensystems, VII Krankheiten des Auges und der Augenanhangsgebilde und VIII Krankheiten des Ohres und des Warzenfortsatzes.)

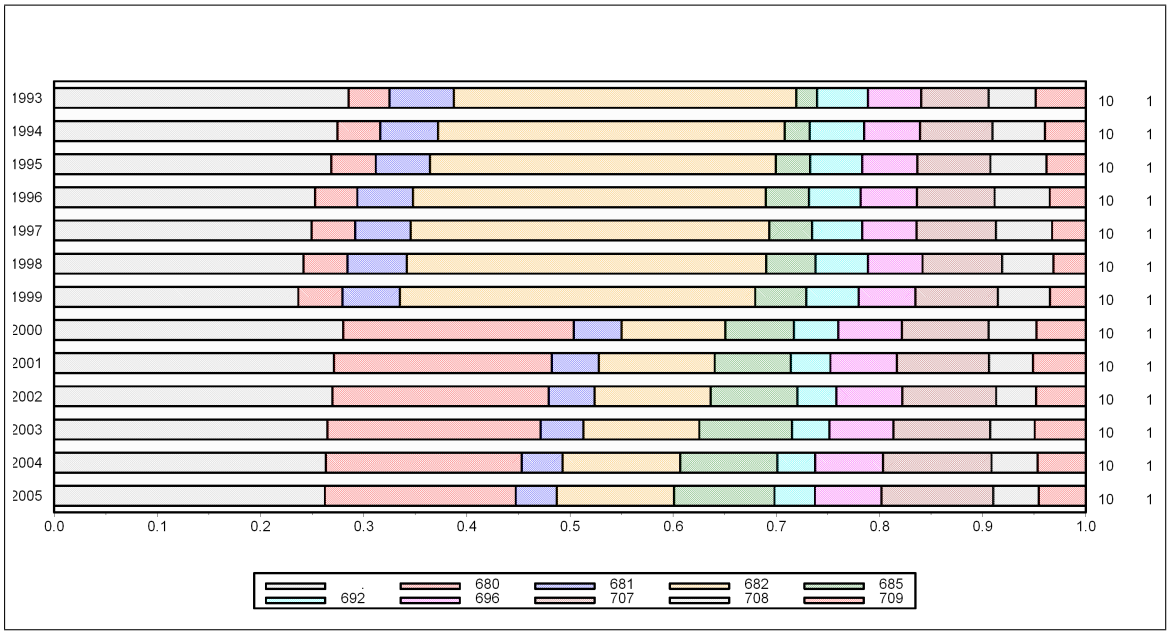
Nimmt man für die Jahre 1993 bis 2005 alle Diagnosen der Krankenhausfälle, so ergibt sich vom gestapelten Balkendiagramm (Abbildung 9 (a)) der Eindruck, dass die Kompositionen (die Kapitel I bis XVII) unterschiedlich stark sind, doch alle über die Zeit wenig Veränderung erfahren. Das wird im Teilbild (b) bestätigt, doch lässt sich im Kompositionsabweichungsdiagramm die Ordnung Komponenten untereinander besser ermitteln. Zusätzlich sieht man, dass es im Verlaufe der 13 Jahre einzelne Kapitel gab, die ihren Anteil an der Komposition in insgesamt ruhigem Umfeld in bestimmten Phasen erhöhen konnten.

Die 17 Kapitel der ICD-9 sind hier Amalgamationen aller Diagnosen im jeweiligen Kapitel. Die 17 Kapitel enthalten etwa 900 dreistellige numerische Krankheitscodes (zwischen 001 und 999 gibt es unbesetzte). Wegen der Eigenschaft der „subcompositional coherence“ können sinnvoll gewählte Kapitel oder Teilkapitel selbst einer Kompositionsdatenanalyse unterworfen werden.

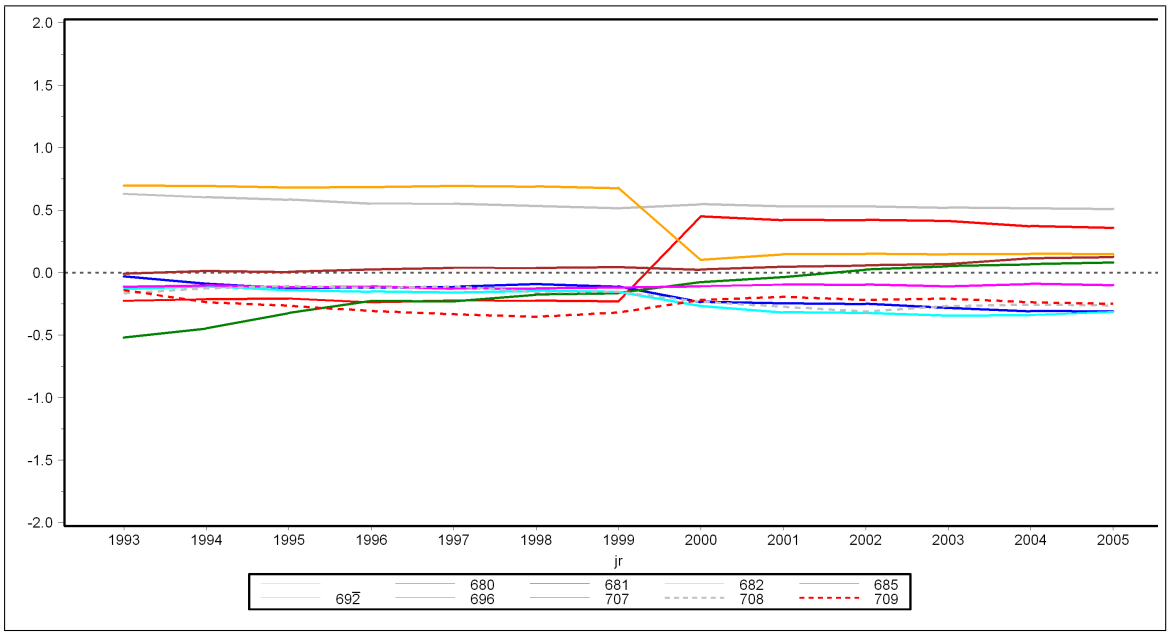
Ein deutlich anderes Bild als Abbildung 9 ergibt die Analyse der Diagnosen des ICD-9-Kapitels XII Hautkrankheiten (Abbildung 10 (a) & (b)). Die gestapelten Balkendiagramme zeigen einen sprunghaften Verlauf zwischen 1999 und 2000 für drei Komponenten: die Diagnosen 680 (rötlich) und 682 (beige) sowie die Aggregat-Komponente aller verbleibenden Diagnosen dieses Kapitels. Im Kompositionsabweichungsdiagramm wird ersichtlich, dass die sprunghaften Veränderungen der Diagnosen 680 und 682 viel bedeutsamer sind als die der Aggregat-Komponente und dass sie sich gegenseitig kompensieren.

Sowohl aus medizinisch-epidemiologischer als auch aus wirtschafts- und sozialwissenschaftlicher Sicht interessiert die Frage nach den Ursachen dieser Veränderung, vielleicht besonders, da diese Veränderung bei ansonsten veränderungsarmen Diagnosen auftreten und für die 6 Jahre bis zum Ende des Untersuchungszeitraumes auf ihrem jeweils neuen Niveau verbleiben. Hautpatienten deutscher Krankenhäuser scheinen plötzlich mit dem Jahr 2000 anders krank.

Hält man sich jedoch vor Augen, dass mit dem Jahreswechsel 1999 zu 2000 die Kodierung nach ICD-10 die bis dahin geltende nach ICD-9 ablösen sollte, kann man vermuten, dass dieser



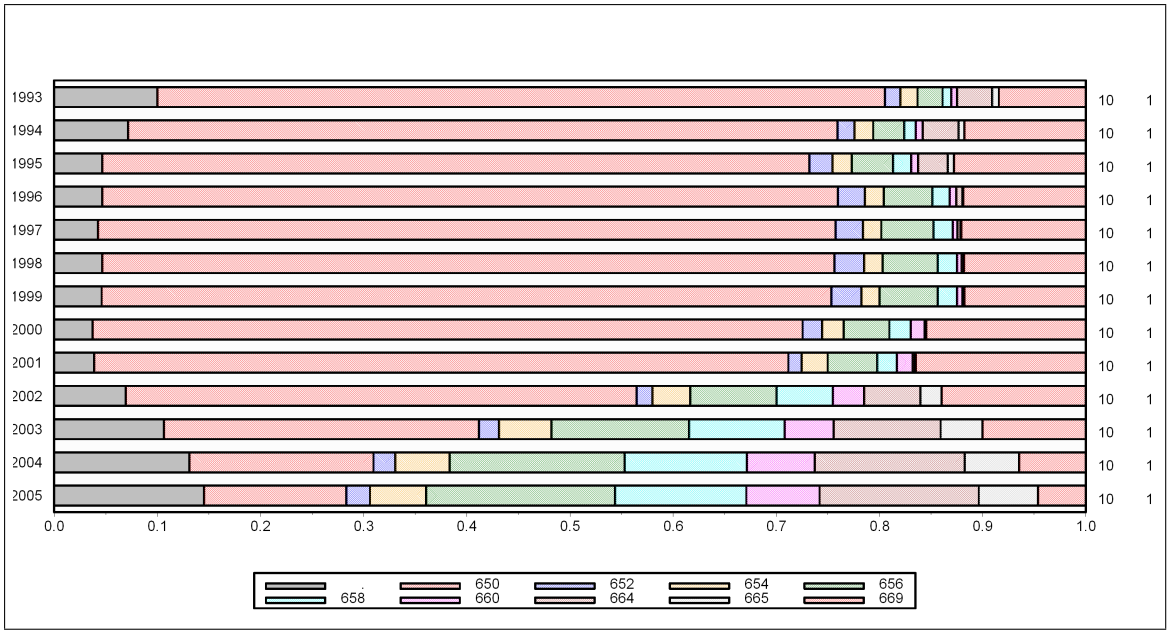
(a)



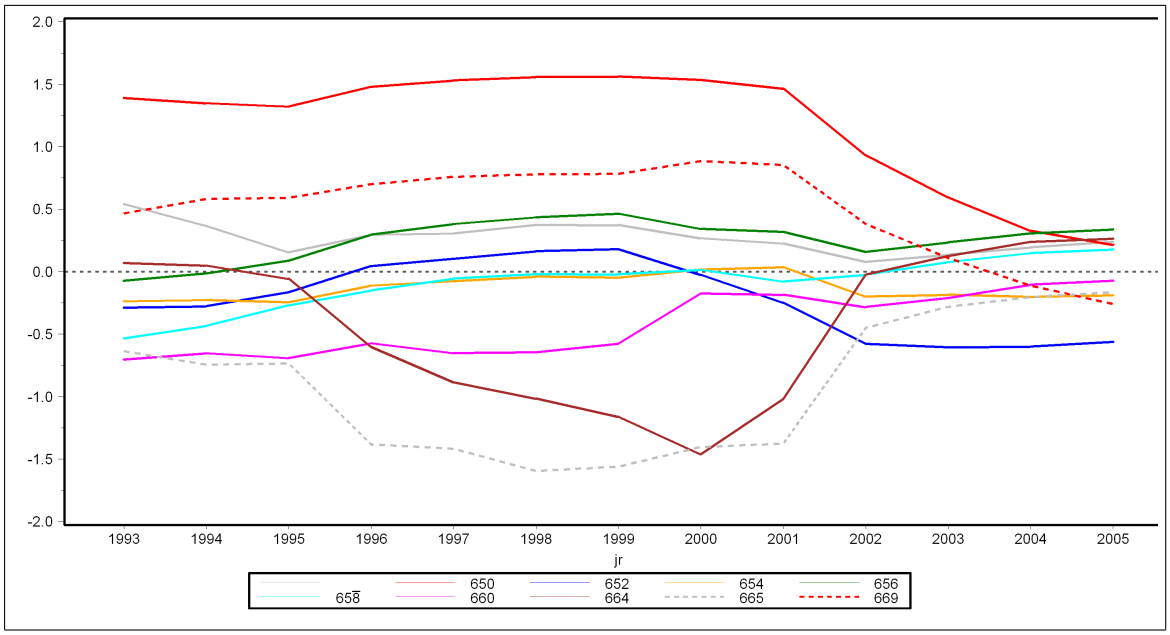
(b)

Abbildung 10: ICD-9 Kapitel XII Krankheiten der Haut (30 Diagnosen):
9 wichtigste & Aggregat (per Amalgamation) aller anderen

Wechsel entweder das Diagnose- und Kodierverhalten hinsichtlich der betreffenden Krankheiten änderte oder dass das ICD-Rekodierungsprogramm von ICD-10 auf ICD-9 an dieser Stelle einen (systematischen) Fehler aufweist.



(a)



(b)

Abbildung 11: ICD-9 Kapitel XI, Diagnosen die Entbindung betreffend (20 Diagnosen): 9 wichtigste & Aggregat (per Amalgamation) aller anderen

Wiederum anders ist die Situation bei Diagnosen, die zu den auf die Entbindung bezogenen Un-

terkapiteln des Kapitels XI Komplikationen in der Schwangerschaft, während der Entbindung und im Wochenbett gehören. Abbildung 11 (a) & (b) zeigt, dass etwa mit dem Jahr 2001 gravierende Veränderungen zwischen mehreren Komponenten stattfinden und dass sich dieser Prozess über mehrere Jahre hinzieht. Da ein Fehler im Zusammenhang mit Kodierveränderungen wegen ICD-Wechsels für die Jahre eher ausgeschlossen werden kann, sollten andere Erklärungen gefunden werden.

Schaut man sich die vier veränderungsintensivsten Komponenten an, so ergibt sich: 650 (normale Entbindung – rot durchgezogene Linie) und 669 (Sonstige Wehen- und Entbindungskomplikationen, anderweitig nicht klassifiziert – rot gebrochene Linie) nehmen stark ab, kompensiert wird dies zum größten Teil durch die Zunahme von 664 (Verletzung des Dammes und der Vulva während der Entbindung – braun durchgezogene Linie) und 665 (Sonstige Geburtsverletzungen – grau gebrochene Linie), diese beiden waren zuvor mit stetig sinkenden Anteilen registriert worden. Es klingt beunruhigend, wenn die bis 2000 für mehr als 70 % der entsprechenden Fälle gestellte Diagnose einer normalen Geburt 2005 gerade so noch für 13 % der Fälle gemeldet wird. Die zugrunde liegenden Fallzahlen bewegen sich zwischen 131.000 (1993, dabei fehlen die Daten für etwa die Hälfte der Krankenhäuser Mecklenburg-Vorpommerns) und 163.000 (2000), danach sinken diese wieder bis auf 136.000 (2005), diese Kovariate kann somit das Phänomen nicht erklären.

Um das Jahr 2001 herum beginnen die Vorbereitungen zur Einführung des Fallpauschalensystems zur Vergütung der Krankenhäuser, die Schwere eines Falles droht abrechnungsmäßig und somit wirtschaftlich einen ganz anderen Stellenwert zu bekommen als vordem. Die Diskussionswellen schlagen so hoch, dass jeder einzelne Krankenhausarzt von diesem Prozess und dem politisch angestrebten Ziel wissen kann. Vielleicht sollte man hier dies aus den Erklärungsversuchen nicht ausschließen.

Anhand dieser Beispiele wurde gezeigt, dass das Kompositionsabweichungsdiagramm die Interpretation von höherdimensionalen Kompositionsdaten über einer (Einfluss-) Variablen deutlich besser unterstützt, als das die bislang verfügbaren Alternativen können.

7 Weitere grafische Darstellung von Kompositionsdaten

Für große Datenmatrizen, d. h., Datensätze mit sowohl vielen Merkmalen (Spalten) als auch vielen (unabhängigen) Beobachtungen (Zeilen) entwickelte Gabriel (1971) das Biplot, ursprünglich zur Visualisierung von Hauptkomponenten-Analysen. Aitchison and Greenacre (2002) adaptieren Biplots für Log-ratio-Analysen (und somit für Kompositionsdaten).

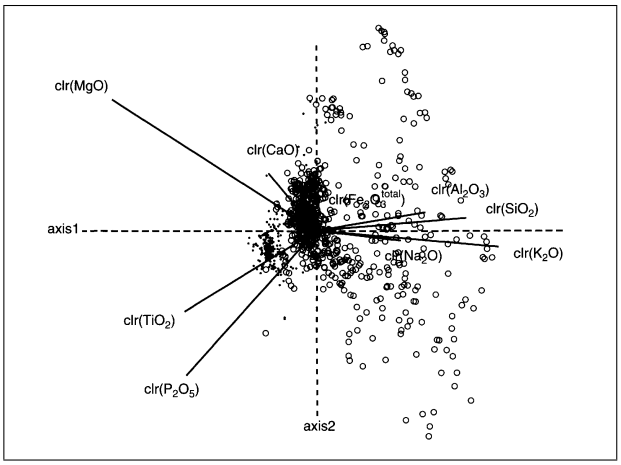


Abbildung 12: Biplot **clr**-transformierter Kompositionsdaten aus S^9

Abbildung 12 (Martín-Fernández and Thío-Henestrosa, 2006, Fig. 2) zeigt das Biplot für Vulkangesteinskomponenten (9 Komponenten: Al_2O_3 bis TiO_2), die Kompositionen selbst wurden zuvor aus dem Simplex S^9 mittels **clr**-Tranformation in eine 8-dimensionale Hyperebene des $\mathbb{R}^9$ transformiert. Das Ergebnis kann dazu beitragen, zwischen verschiedenen Gruppen (hier als Punkte bzw. Kreise dargestellt) in den Daten zu diskriminieren und die Bedeutung der einzelnen Komponenten für die Diskrimination zu bewerten, also ein Klassifikationsproblem zu lösen. Hier wird ebenfalls von der interpretatorischen Nähe der transformierten zu den Daten im Simplex Gebrauch gemacht: alle Komponenten bleiben sichtbar, alle Komponenten werden gleich behandelt.

Literatur

Aguilar Zuil, L., C. Barceló-Vidal, and J. M. Larrosa (2007). Compositional Time Series Analysis: A Review. `http://ima.udg.edu/~barcelo/index_archivos/ISI2007_Aguilar.pdf` (13. 10. 2010).

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society B (Statistical Methodology) 44*, 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data (Reprint with additional material)*. Caldwell, NJ: The Blackburn Press.

Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Reply to letter to the editor by S. Rehder and U. Zier on «Logratio analysis and compositional distance» by J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn. *Mathematical Geology 33*, 849–860.

Aitchison, J. and J. J. Egozcue (2005). Compositional data analysis: Where are we and where should we be heading. *Mathematical Geology 37*, 829–850.

Aitchison, J. and M. Greenacre (2002). Biplots of compositional data. *Journal of the Royal Statistical Society C (Applied Statistics) 51*, 375–392.

Buccianti, A. and V. Pawlowsky-Glahn (2005). New perspectives on water chemistry and compositional data analysis. *Mathematical Geology 37*, 703–727.

Chayes, F. (1960). On correlation of variables of constant sum. *Journal of Geophysical Research 65*, 4185–4193.

Eaton, M. L. (2007). *Multivariate Statistics: A Vector Space Approach*. Lecture Notes – Monograph Series # 53. Beachwood, OH: Institute of Mathematical Statistics.

Egozcue, J. J. and V. Pawlowsky-Glahn (2006). Simplicial geometry for compositional data. In A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn (Eds.), *Compositional Data Analysis in the Geosciences: From Theory to Praxis*, Volume 264 of *Geological Society Special Publication*, pp. 145–159. London: Geological Society.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*, 279–300.

Filzmoser, P., K. Hron, and C. Reimann (2009). Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment 407*, 6100–6108.

Fry, J. M., T. R. L. Fry, K. R. McLaren, and T. N. Smith (2000). Compositional data analysis and zeroes in microdata. *Applied Economics 32*, 953–959.

Fry, J. M., T. R. L. Fry, K. R. McLaren, and T. N. Smith (2001). Modelling zeroes in microdata. *Applied Economics 33*, 383–392.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika 58*, 453–467.

Graf, M. (2006). *Swiss Earnings Structure Survey 2002-2004 – Compositional data in a stratified two-stage sample: Analysis and precision assessment of wage components.* Swiss Statistics Methodology Report 338-0038. Neuchâtel: Swiss Federal Statistical Office.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective.* New York: Springer.

Martín-Fernández, J. A. and S. Thío-Henestrosa (2006). Rounded zeros: some practical aspects for compositional data. In A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn (Eds.), *Compositional Data Analysis in the Geosciences: From Theory to Praxis*, Volume 264 of *Geological Society Special Publication*, pp. 191–201. London: Geological Society.

Mills, T. C. (2009). Forecasting obesity trends in England. *Journal of the Royal Statistical Society A (Statistics in Society) 172*, 107–117.

Mills, T. C. (2010). Forecasting compositional time series. *Quality & Quantity 44*, 673–690.

Pawlowsky-Glahn, V. and J. J. Egozcue (2007). Statistische Analyse von Kompositionsdaten. In *58. Berg- und Hüttenmännischer Tag 2007 – Tagungsband Geologische Modellierung*, Freiberg, pp. 253–260.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2007). Lecture Notes on Compositional Data Analysis. `http://dugi-doc.udg.edu/bitstream/10256/297/1/CoDa-book.pdf` (13. 06. 2008).

Pearson, K. (1897). On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London 60*, 489–498.

Ravishanker, N., D. K. Dey, and M. Iyengar (2001). Compositional time series analysis of mortality. *Communications in Statistics – Theory and Methods 30*, 2281–2291.

Rönz, B. and H. G. Strohe (Eds.) (1994). *Lexikon Statistik.* Wiesbaden: Gabler.

Srivastava, D. K., J. M. Boyett, C. W. Jackson, and X. Tong (2007). A comparison of permutation Hotelling's t^2 test and log-ratio test for analyzing compositional data. *Communications in Statistics – Theory and Methods 36*, 415–431.

van den Boogaart, K. G. (2008). Arbeits- und Interessengebiete. `http://www.stat.boogaart.de/interests_de.php` (07. 05. 2008).

van den Boogaart, K. G. (2009). Arbeits- und Interessengebiete. `http://www.stat.boogaart.de/interests_de.php` (21. 03. 2009).

(LaTeX 2ε-Version: 2. November 2010, 16:09)

UNIVERSITÄT POTSDAM

Wirtschafts- und Sozialwissenschaftliche Fakultät

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 1	1995	Strohe, Hans Gerhard: Dynamic Latent Variables Path Models - An Alternative PLS Estimation -
Nr. 2	1996	Kempe, Wolfram. Das Arbeitsangebot verheirateter Frauen in den neuen und alten Bundesländern - Eine semiparametrische Regressionsanalyse
Nr. 3	1996	Strohe, Hans Gerhard: Statistik im DDR-Wirtschaftsstudium zwischen Ideologie und Wissenschaft
Nr. 4	1996	Berger, Ursula: Die Landwirtschaft in den drei neuen EU-Mitgliedsstaaten Finnland, Schweden und Österreich - Ein statistischer Überblick
Nr. 5	1996	Betzin, Jörg: Ein korrespondenzanalytischer Ansatz für Pfadmodelle mit kategorialen Daten
Nr. 6	1996	Berger, Ursula: Die Methoden der EU zur Messung der Einkommenssituation in der Landwirtschaft - Am Beispiel der Bundesrepublik Deutschland
Nr. 7	1997	Strohe, Hans Gerhard / Geppert, Frank: Algorithmus und Computerprogramm für dynamische Partial Least Squares Modelle
Nr. 8	1997	Rambert, Laurence / Strohe, Hans Gerhard: Statistische Darstellung transformationsbedingter Veränderungen der Wirtschafts- und Beschäftigungsstruktur in Ostdeutschland
Nr. 9	1997	Faber, Cathleen: Die Statistik der Verbraucherpreise in Rußland - Am Beispiel der Erhebung für die Stadt St. Petersburg -
Nr. 10	1998	Nosova, Olga: The Atractiveness of Foreign Direct Investment in Russia and Ukraine - A Statistical Analysis
Nr. 11	1999	Gelaschwili, Simon: Anwendung der Spieltheorie bei der Prognose von Marktprozessen
Nr. 12	1999	Strohe, Hans Gerhard / Faber, Cathleen: Statistik der Transformation - Transformation der Statistik. - Preisstatistik in Ostdeutschland und Rußland -
Nr. 13	1999	Müller, Claus: Kleine und mittelgroße Unternehmen in einer hoch konzentrierten Branche am Beispiel der Elektrotechnik - Eine statistische Langzeitanalyse der Gewerbezählungen seit 1882 -
Nr. 14	1999	Faber, Cathleen: The Measurement and Development of Georgian Consumer Prices
Nr. 15	1999	Geppert, Frank / Hübner, Roland: Korrelation oder Kointegration – Eignung für Portfoliostrategien am Beispiel verbriefter Immobilenanlagen
Nr. 16	2000	Achsani, Noer Azam / Strohe, Hans Gerhard: Statistischer Überblick über die indonesische Wirtschaft
Nr. 17	2000	Bartels, Knut: Testen der Spezifikation von multinominalen Logit-Modellen
Nr. 18	2002	Achsani, Noer Azam / Strohe, Hans Gerhard: Dynamische Zusammenhänge zwischen den Kapitalmärkten der Region Pazifisches Becken vor und nach der Asiatischen Krise 1997
Nr. 19	2002	Nosova, Olga: Modellierung der ausländischen Investitionstätigkeit in der Ukraine
Nr. 20	2003	Gelaschwili, Simon / Kurtanidse, Zurab: Statistische Analyse des Handels zwischen Georgien und Deutschland
Nr. 21	2004	Nastansky, Andreas: Kurz- und langfristiger statistischer Zusammenhang zwischen Geldmengen- und Preisentwicklung: Analyse einer kointegrierenden Beziehung
Nr. 22	2006	Kauffmann, Albrecht / Nastansky, Andreas: Ein kubischer Spline zur temporalen Disaggregation von Stromgrößen und seine Anwendbarkeit auf Immobilienindizes

UNIVERSITÄT POTSDAM

Wirtschafts- und Sozialwissenschaftliche Fakultät

STATISTISCHE DISKUSSIONSBEITRÄGE

Herausgeber: Hans Gerhard Strohe

Nr. 23	2006	Mangelsdorf, Stefan: Empirische Analyse der Investitions- und Exportentwicklung des Verarbeitenden Gewerbes in Berlin und Brandenburg
Nr. 24	2006	Reilich, Julia: Return to Schooling in Germany
Nr. 25	2006	Nosova, Olga / Bartels, Knut: Statistical Analysis of the Corporate Governance System in the Ukraine: Problems and Development Perspectives
Nr. 26	2007	Gelaschwili, Simon: Einführung in die Statistische Modellierung und Prognose
Nr. 27	2007	Nastansky, Andreas: Modellierung und Schätzung von Vermögenseffekten im Konsum
Nr. 28	2008	Nastansky, Andreas: Schätzung vermögenspreisinduzierter Investitionseffekte in Deutschland
Nr. 29	2008	Ruge, Marcus / Strohe, Hans Gerhard: Analyse von Erwartungen in der Volkswirtschaft mit Partial-Least-Squares-Modellen
Nr. 30	2009	Newiak, Monique: Prüfungsurteile mit Dollar Unit Sampling – Ein Vergleich von Fehlerschätzmethoden für Zwecke der Wirtschaftsprüfung: Praxis, Theorie, Simulation –
Nr. 31	2009	Ruge, Marcus: Modellierung von Stimmungen und Erwartungen in der deutschen Wirtschaft
Nr. 32	2009	Nosova, Olga: Statistical Analysis of Regional Integration Effects
Nr. 33	2009	Mangelsdorf, Stefan: Persistenz im Exportverhalten – Kann punktuelle Exportförderung langfristige Auswirkungen haben? -
Nr. 34	2009	Kbiladze, David: Einige historische und gesetzgeberische Faktoren der Reformierung der georgischen Statistik
Nr. 35	2009	Nastansky, Andreas / Strohe, Hans Gerhard: Die Ursachen der Finanz- und Bankenkrise im Lichte der Statistik
Nr. 36	2009	Gelaschwili, Simon / Nastansky, Andreas: Development of the Banking Sector in Georgia
Nr. 37	2010	Kunze, Karl-Kuno / Strohe, Hans Gerhard: Time Varying Persistence in the German Stock Market
Nr. 38	2010	Nastansky, Andreas / Strohe, Hans Gerhard: The Impact of Changes in Asset Prices on Real Economic Activity - A Cointegration Analysis for Germany -
Nr. 39	2010	Kunze, Karl-Kuno / Strohe, Hans Gerhard: Antipersistence in German Stock Returns
Nr. 40	2010	Dietrich, Irina / Strohe, Hans Gerhard: Die Vielfalt öffentlicher Unternehmen aus der Sicht der Statistik - Ein Versuch, das Unstrukturierte zu strukturieren -
Nr. 41	2010	Nastansky, Andreas / Lanz, Ramona: Bonuszahlungen in der Kreditwirtschaft: Analyse, Regulierung und Entwicklungstendenzen
Nr. 42	2010	Dietrich, Irina / Strohe, Hans Gerhard: Die Vermögenslage öffentlicher Unternehmen in Deutschland - Statistische Analyse anhand von amtlichen Mikrodaten der Jahresabschlüsse.
Nr. 43	2010	Ulbrich, Hannes-Friedrich: Höherdimensionale Kompositionsdaten – Gedanken zur grafischen Darstellung und Analyse -

ISSN 0949-068X