Peter Ihm

Korrespondenzanalyse und Seriation

1) Einleitung

In Nummer 5, 1983, der Archäologischen Informationen, nahm H.Ziegert in einer Arbeit über Kombinations-Statistik und Seriation zu Methode und Ergebnis der Bronzezeit-Chronologie K. Goldmanns Stellung. In diesem Beitrag möchte ich auf den mathematischen Hintergrund des Verfahrens eingehen.

Wenn man von einer kleinen Modifikation absieht, ist Goldmanns Verfahren mit der seit etwa zehn Jahren in Mode gekommenen Korrespondenzanalyse identisch. Deren Ziel läßt sich mathematisch leicht beschreiben: den Zeilen und Spalten einer Kontingenztafel sollen Koordinaten, "scores", x (Zeilen) und y (Spalten) in der Weise zugeordnet werden, daß die Korrelation zwischen x und y maximal wird. Hat die Tafel die entsprechende Struktur, wird sie dadurch diagonalisiert, d.h. so umgeordnet, daß die größeren Zeilenhäufigkeiten in oder in die Nähe der Diagonalen rücken.

Soll die Ordnung der Zeilen chronologisch sein, setzt dies die Gültigkeit bestimmter Modelle zeitlicher Variation von Merkmals- und Typenhäufigkeiten voraus. Die Realisierung derartiger Modelle wurde in vielen Fällen nachgewiesen, doch besagt die erfolgreiche Diagonalisierung einer Kontingenztafel nicht per se, daß die Ordnung chronologisch sein muß.

2) Die Kontingenztafel

Eine Kontingenztafel mit m Zeilen und n Spalten ist eine Häufigkeitstabelle. Die Zeilen können z.B. Gräbern, die Spalten Typen zugeordnet sein. Die Zellenhäufigkeit h_{ik} ist dann die Zahl der Fundstücke von k-ten Typ, die im i-ten Grab gefunden wurden.

Ein anderes Beispiel ist die Einteilung bandkeramischer Gefäße nach Randform und Stichfüllung: Die Randformen sowie die Füllungstypen werden in m bzw. n sich gegenseitig ausschließende Kategorien eingeteilt (man spricht von einer Nominalskala). h_{ik} ist die Häufigkeit der Gefäße mit der i-ten Randform und dem k-ten Stichfüllungstyp. Diese Tabelle wird meist als Kombinationstabelle bezeichnet. Für die Korrespondenzanalyse sind beide Tabellentypen gleichwertig, so daß ich hier den in der Statistik üblichen Begriff der Kontingenztafel verwenden möchte.

Ein anderer Tabellentyp entsteht, wenn die Zeilen Fundstücken, die Spalten Merkmalen dieser Fundstücke entsprechen. Meist stehen dann in den Zellen Nullen und Einsen. Betrachten wir beispielsweise eine Stichprobe von m Schwertern, bei denen n Merkmale auf An- bzw. Abwesenheit geprüft werden. Dann ist $h_{ik} = 1$, wenn das i-te Schwert das k-te Merkmal besitzt; andernfalls ist $h_{ik} = 0$. Hier spricht man nicht von einer Kontingenztafel, sondern von einer m × n-Datenmatrix. Erfreulicherweise läßt sich die Korrespondenzanalyse in ihrer Standardform auch bei Datenmatrizen anwenden, so daß wir auf diesen Fall nicht gesondert eingehen müssen und im allgemeinen von Kontingenztafeln sprechen können.

Bei einer Kontingenztafel können Zeilen-, Spalten- und die Gesamtsumme gebildet werden. Die Gesamtsumme ist

$$h_{..} = \Sigma_{i=1}^{m} \Sigma_{j=1}^{n} h_{ik} = \Sigma\Sigma h_{ik} \quad .$$

Es ist für die folgenden Ausführungen bequemer, statt mit den absoluten Zellenhäufigkeiten h_{ik} mit den relativen Zellenhäufigkeiten

$$\hat{p}_{ik} = h_{ik}/h_{..}$$

zu rechnen. Der Zirkumflex über dem p soll deutlich machen, daß es sich nicht um Wahrscheinlichkeiten, sondern um deren Schätzfunktionen aus der Stichprobe handelt. Weil aber in der vorliegenden Arbeit nicht zwischen Wahrscheinlichkeiten (p) und relativen Häufigkeiten ($\hat{p}$) unterschieden werden muß, werden wir zur Vereinfachung

$$p_{ik} = h_{ik}/h_{..}$$

schreiben, wohl wissend, daß dies ein Verstoß gegen gute Sitten und Gebräuche ist. Die Randsummen sind

$$p_{i.} = \Sigma_{k=1}^{n} p_{ik} = \Sigma\, p_{ik} \quad ,$$

Summe der i-ten Zeile, und

$$p_{.k} = \Sigma_{i=1}^{m} p_{ik} = \Sigma p_{ik} \quad ,$$

Summe der k-ten Spalte. Die Gesamtsumme ist jetzt

$$p_{..} = \Sigma\Sigma p_{ik} = 1 \quad .$$

Wie man bei $p_{i.}$, $p_{.k}$ und $p_{..}$ sieht, steht ein Punkt anstelle des Index über den summiert wurde. Wir werden bei der Beschreibung des Rechenverfahrens wieder auf die absoluten Häufigkeiten zurückkommen. Dann werden die Randsummen mit $h_{i.}$, $h_{.k}$ bzw. $h_{..}$ bezeichnet.

3) Der Korrelationskoeffizient

In diesem Abschnitt behandeln wir den Korrelationskoeffizienten in einer Form, die dem späteren Verständnis der Korrespondenzanalyse dienlich sein soll. Dabei gehen wir davon aus, daß der Korrelationskoeffizient gleich der Kovarianz von Variablen ist, deren Mittelwerte und Varianzen gleich null bzw. eins sind.
Gegeben seien 10 Beobachtungspaare:

x:	1	1.5	1.5	1.5	2	2	2.5	2.5	2.5	3
y:	1	1.5	1.5	2	2	2.5	2	2	2.5	3

Statt den Korrelationskoeffizienten r nach der üblichen Formel zu berechnen, schlagen wir einen anderen Weg ein. Wir gehen davon aus, daß die Paare x,y mit der relativen Häufigkeit p=0.1 oder p=0.2 vorkommen. Die verschiedenen Werte von x und y, x_i, i = 1,2,...m, bzw. y , k = 1, 2,... n, sind zusammen mit ihren relativen Häufigkeiten in der folgenden Kontingenztafel dargestellt:

	k:	1	2	3	4	5	
	y_k:	1	1.5	2	2.5	3	$p_{i.}$
i	x_i						
1	1	0.1	.	.	.	.	0.1
2	1.5	.	0.2	0.1	.	.	0.3
3	2	.	.	0.1	0.1	.	0.2
4	2.5	.	.	0.2	0.1	.	0.3
5	3	.	.	.	.	0.1	0.1
	$p_{.k}$	0.1	0.2	0.4	0.2	0.1	1.0

Null ist zur besseren Übersichtlichkeit durch einen Punkt ersetzt. Wir berechnen nun die Mittelwerte von x und y:

$$\bar{x} = \Sigma p_{i.} x_i$$

$$\bar{y} = \Sigma p_{.k} y_k$$

und erhalten

$$\bar{x} = 0.1\times 1+0.3\times 1.5+...+0.1\times 3 = 2.0$$

$$\bar{y} = 0.1\times 1+0.2\times 1.5+...+0.1\times 3 = 2.0$$

Nun werden die x und y durch Abziehen von $\bar{x}$ bzw. $\bar{y}$ auf den Mittelwert bezogen. Dies drückt man am besten in algorithmischer Sprache aus, z.B. in BASIC:

```
FOR I=1 TO M
X(I)=X(I)-X̄
NEXT I
```

und

```
FOR K=1 TO N
Y(K)=Y(K)-Ȳ
NEXT K
```

Die Kontingenztafel ist jetzt

	k :	1	2	3	4	5	
	y_k:	-1	-0.5	0	0.5	1	$p_{i.}$
i	x_i						
1	-1	0.1	.	.	.	.	0.1
2	-0.5	.	0.2	0.1	.	.	0.3
3	0	.	.	0.1	0.1	.	0.2
4	0.5	.	.	0.2	0.1	.	0.3
5	1	.	.	.	.	0.1	0.1
	$p_{.k}$	0.1	0.2	0.4	0.2	0.1	1.0

Die Mittelwerte von x und y sind jetzt gleich null. Nun müssen die Werte noch in der Weise standardisiert werden, daß die Varianz gleich eins ist. Hierzu werden sie durch die jeweiligen Standardabweichungen, die Wurzeln aus den Varianzen, dividiert. Die Varianzen von x und y sind

$$s_x^2 = \Sigma p_{i.} x_i^2$$

$$s_y^2 = \Sigma p_{.k} y_k^2 \quad ,$$

in unserem Beispiel

$$s_x^2 = 0.1\times(-1)^2+0.3\times(-0.5)^2+\ldots+0.1\times 1^2 = 0.35$$

$$s_y^2 = 0.1\times(-1)^2+0.2\times(-0.5)^2+\ldots+0.1\times 1^2 = 0.30 \quad .$$

Die Standardabweichungen sind $s_x = 0.5916$, $s_y = 0.5477$. Standardisierte x- und y-Werte ergeben sich jetzt aus

```
FOR I = 1 TO M
X (I) = X (I)/SX
NEXT I
```

```
FOR K = 1 TO M
Y (K) = Y (K)/SY
NEXT K
```

Die Kontingenztafel ist jetzt (gerundete Werte):

	k :	1	2	3	4	5	
	y_k:	-1.69	-0.85	0	0.85	1.69	$p_{i.}$
i	x_i						
1	-1.83	0.1	.	.	.	.	0.1
2	-0.91	.	0.2	0.1	.	.	0.3
3	0	.	.	0.1	0.1	.	0.2
4	0.91	.	.	0.2	0.1	.	0.3
5	1.83	.	.	.	.	0.1	0.1
	$p_{.k}$	0.1	0.2	0.4	0.2	0.1	1.0

Unter Verwendung der relativen Häufigkeiten p_{ik} ergibt sich der Korrelationskoeffizient, den wir diesmal mit ρ statt r bezeichnen wollen, aus der Formel

$$\rho = \Sigma\Sigma p_{ik} x_i y_k \quad ,$$

im speziellen Fall

$$\rho = 0.1x1.83x1.69+0.2x0.91x0.85+0.1x0.91x0.85+0.1x1.83x1.69$$
$$= 0.85059$$

bzw. gerundet $\rho = 0.85$. Diese relativ hohe Korrelation ist darauf zurückzuführen, daß die von Null verschiedenen Häufigkeiten vorzugsweise in der Diagonalen stehen. Der höchste Wert von $\rho = 1$ wird erreicht, wenn die Tafel quadratisch ist und nur die Diagonale von Null verschiedene Häufigkeiten enthält. Allgemein kann man sagen, daß die Korrelation umso größer ist, je mehr sich die großen Zeilenhäufigkeiten in der Diagonalen anordnen. Dieser Zusammenhang zwischen Korrelation und Diagonalisierung der Tafel wird bei der Korrespondenzanalyse genutzt.

4) Die Korrespondenzanalyse

Wir betrachten eine einfache Kontingenztafel mit m = 3 Gräbern und n = 4 Typen:

		Typ				
	1	2	3	4	5	Σ
Grab	1	1	.	3	1	5
	2	.	2	.	1	3
	3	1	.	1	.	1
		2	2	4	2	10

Ordnen wir dem i-ten Grab die Koordinate $x_i = i$, dem k-ten Typ $y_k = k$ zu, ergibt sich der Korrelationskoeffizient $\rho = -0.28$. Schreibt man die Zeilen in der Reihenfolge 2,1,3, die Spalten in der Reihenfolge 2,4,3,1, ergibt sich sukzessive

	1	2	3	4
1 (2)	.	2	.	1
2 (1)	1	.	3	1
3 (3)	1	.	1	.

und

	1 (2)	2 (4)	3 (3)	4 (1)	Σ
1 (2)	2	1	.	.	3
2 (1)	.	1	3	1	5
3 (3)	.	.	1	1	2
	2	2	4	2	10

Nun ist die Tafel in die bei Seriationsverfahren angestrebte diagonale Form gebracht worden, und der Korrelationskoeffizient der neuen Zeilen- und Spaltennummern hat den Wert $\rho = 0.78$. Nun stellen sich sofort zwei Fragen:

1. Ist dies der größte erreichte Wert von ρ ?
2. Wie verfährt man bei der Umordnung, um die Ordnung mit dem größten Wert von ρ so schnell wie möglich zu erhalten?

Hier waren einige Ordnungen einfach ausprobiert worden, was bei der kleinen Tafel möglich war, bei größeren Tafeln aber zu zeitaufwendig wäre. Gesucht wird ein Algorithmus, der es erlaubt, auf die gewünschte Lösung zielgerichtet hinzuarbeiten. Ein Verfahren wurde von Goldmann und Kammerer entwickelt und von ersterem 1972 veröffentlicht. Nach einer kleinen Modifikation erweist es sich als mit der Korrespondenzanalyse identisch. Obwohl sich diese schon in Arbeiten von Hirschfeld (1935) und Williams (1952; siehe Kendall und Stuart 1961) findet, wurde sie doch erst durch die Publikationen von Cordier (1965), Cordier-Escofier (1969) und Benzecri (1969, 1973) bekannt (Bibliographie in Benzecri 1973). Da die Korrespondenzanalyse bessere mathematische Eigenschaften aufweist als Goldmanns Ansatz, sollte man diese in den Vordergrund der Erörterungen stellen.

Es handelt sich also darum, Zeilen und Spalten der Tafel solange umzuordnen, bis der Korrelationskoeffizient ρ zu einem Maximum wird. In unserem Beispiel haben wir hierzu $x_i = i$, $y_k = k$ gesetzt, also die Zeilen- und Spaltennummern miteinander korreliert. Tatsächlich kommt man aber zu einem brauchbaren Algorithmus, wenn man stattdessen zuläßt, daß die x_i und y_k, d.h. die Zeilen und Spalten zugeordneten Koordinaten, auch andere reelle Zahlenwerte annehmen können als 1,2,... . Das bedeutet, daß benachbarte Zeilen und Spalten dann im allgemeinen verschieden weit auseinanderrücken oder sogar zusammenfallen können. Die Lösung ergibt sich aus folgendem theoretischen Ansatz:

$$\Sigma\Sigma p_{ik} x_i y_k = \rho = \text{Maximum}$$

mit den Nebenbedingungen

4.1 $$\Sigma p_{i.} x_i = 0 = \Sigma p_{.k} y_k$$

4.2 $$\Sigma p_{i.} x_i^2 = 1 = \Sigma p_{.k} y_k^2 \quad .$$

Diese Forderungen führen zu den Gleichungen

4.3 $$\Sigma p_{ik} x_i / p_{.k} = y_k \quad , \quad k=1,2,...n$$

4.4 $$\Sigma p_{ik} y_k / p_{i.} = x_i \quad , \quad i=1,2,...m \quad .$$

Wegen der Division durch $p_{i.}$ bzw. $p_{.k}$ kann man auch mit den absoluten Häufigkeiten h_{ik} arbeiten, wonach sich die Gleichungen (4.3) und (4.4) zu

4.5 $\quad \Sigma h_{ik} x_i / h_{.k} = y_k \quad ,$

4.6 $\quad \Sigma h_{ik} y_k / h_{i.} = x_i$

umformen lassen. Sie lassen sich iterativ lösen, wie wir an einem Beispiel sehen werden. Wir verwenden die ungeordnete Tafel dieses Abschnitts und beginnen mit $y_K = k$, d.h. $y_1 = 1$, $y_2 = 2$, $y_3 = 3$, $y_4 = 4$. Dann sind x_1 , x_2 , x_3 Mittelwerte für die erste, zweite bzw. dritte Zeile:

y:	1	2	3	4	Σ	
	1	.	3	1	5	$x_1=(1\times1+3\times3+1\times4)/5=2.80$
	.	2	.	1	3	$x_2=(2\times2+1\times4)/3=2.67$
	1	.	1	.	2	$x_3=(1\times1+1\times3)/2=2.00$

Mit diesen x-Werten würde nun weitergerechnet, um daraus Spaltenmittel y zu erhalten. Nun sind aber die Bedingungen (4.1) und (4.2) nicht erfüllt. Achteten wir nicht darauf, erhielten wir alsbald die triviale Lösung $x_i = 1$ und $y_K = 1$ für alle i und k. Die y-Werte müssen wie in Abschnitt 3 standardisiert werden. Dieses Verfahren kann man sich etwas vereinfachen, indem man von den y_K den kleinsten Wert, y_{min} , abzieht, durch die Spannweite $y_{max} - y_{min}$ dividiert und mit den so erhaltenen neuen Werten weiterrechnet. Für $y_{min} = 2.00$, $y_{max} = 2.80$ erhalten wir

$$y_1 = (2.80-2.00)/(2.80-2.00) = 1$$
$$y_2 = (2.67-2.00)/(2.80-2.00) = 0.84$$
$$y_3 = (2.00-2.00)/(2.80-2.00) = 0$$

Mit diesen Werten wird weitergerechnet:

x				
1	1	.	3	1
0.84	.	2	.	1
0	1	.	1	.
Σ	2	2	4	2
y:	0.50	0.84	0.75	0.92

Hier wird nun in gleicher Weise verfahren. Mit $y_{min} = 0.50$, $y_{max} - y = 0.42$ gilt

$$y_1 = 0$$
$$y_2 = 0.81$$
$$y_3 = 0.60$$
$$y_4 = 1$$

und wir erhalten nun x-Werte aus

y:	0	0.81	0.60	1	Σ	x
	1	.	3	1	5	0.56
	.	2	.	1	3	0.87
	1	.	1	.	2	0.30

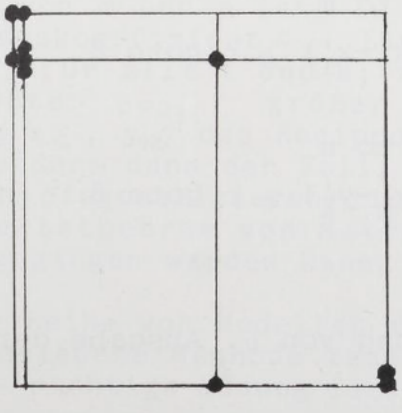
Dieses setzt sich nun so fort. Man ahnt jetzt schon, daß die erste und zweite Zeile miteinander vertauscht werden. Nach einer geringen Zahl von Iterationsschritten erhalten wir

x \ y	0	1	0.03	0.54	Σ
0.13	1	.	3	1	5
1	.	2	.	1	3
0	1	.	1	.	2
Σ	2	2	4	2	10

und die Lösung bleibt stabil. Die Zeilen und Spalten sind dann in den Reihenfolgen 3,1,2 bzw. 1,3,4,2 zu schreiben:

		y				
		0	0.03	0.54	1	Σ
	0	1	1	.	.	2
x	0.13	1	3	1	.	5
	1	.	.	1	2	3
	Σ	2	4	2	2	10

Dies ist die Ordnung, die durch Probieren erhalten wurde, abgesehen davon, daß Zeilen und Spalten in umgekehrter Reihenfolge stehen. Diese mögliche gleichzeitige Umkehrung der Reihenfolge von Zeilen und Spalten ist verfahrensinhärent, doch ist bis auf mögliche verschiedene Normierung die Lösung eindeutig. Benutzt man die Koordinaten x und y selbst zur Darstellung, ergibt sich



Der Korrelationskoeffizient, jetzt für die x- und y-Werte berechnet, ist $\rho = 0.9476$. Er ist deutlich größer als der für die Zeilen- und Spaltennummern erhaltene Wert von $\rho = 0.78$.

Nun stelle ich das Verfahren allgemein dar, wobei ich mich eines Gemisches der BASIC- und der bisher benutzten Symbolik bediene. Ich schreibe die Anweisungen daher mit kleinen Buchstaben.

1 for i=1 to m
$x'_i = 1$
next i
for k=1 to n
$y'_k = k$
next k

2 for i=1 to m
$x_i = \Sigma p_{ik} y'_k / p_{i.}$
next i

3 for i=1 to m
$x_i = (x_i - x_{min})/(x_{max} - x_{min})$
next i

4 for k=1 to n
$y_k = \Sigma p_{ik} x_i / p_{.k}$
next k

5 for k=1 to n
$y_k = (y_k - y_{min})/(y_{max} - y_{min})$
next k

6 for i=1 to m
if $abs(x_i - x'_i) \geq \varepsilon$ goto 8
next i
for k=1 to n
if $abs(y_k - y'_k) \geq \varepsilon$ goto 8
next k

7 Berechnung von ρ, Ausgabe der Ergebnisse
end

```
8       for i=1 to m
        x'i = xi
        next i
        for k=1 to n
        y'k = yk
        next k
        goto 2
```

In 3 und 5 kann man auch $x_i = (x_i - \bar{x})/s_x$ bzw. $y_K = (y_K - \bar{y})/s_y$ verwenden. Da man zur Umordnung der Tabelle am Ende aber ohnehin ein Programm zum Ordnen der x_i und y_K benötigt, ist das nicht unbedingt vorteilhafter. Für die Handrechnung ist die ursprüngliche Version in 3 und 5 bequemer. Goldmanns Algorithmus unterscheidet sich von dem hier dargestellten dadurch, daß in 3 und 4 x_i = Rang (x_i) bzw. y_K = Rang (y_K) verwendet wird, wobei der Rang die Ordnungszahl ist, wenn man die Zahlen der Größe nach ordnet.

Das hier beschriebene Verfahren ist unter dem englischen Namen Reciprocal Averaging bekannt geworden. Es stand zunächst für sich selbst, bis Hill (1973, 1974) den Zusammenhang mit der Korrespondenzanalyse entdeckte. Tatsächlich leistet die Korrespondenzanalyse mehr als die Diagonalisation einer Kontingenztafel, doch möchte ich hierauf in einem späteren Beitrag eingehen.

5) Anwendungsmöglichkeiten

Zunächst ist festzustellen, daß die Korrespondenzanalyse in ihrer hier beschriebenen Form des Reciprocal Averaging nur diejenige Zeilen- und Spaltenanordnung liefert, die den Korrelationskoeffizienten maximisiert und damit die beste diagonale Anordnung der Tafel gibt. Das Problem ist nun, was man damit macht. Bekanntlich wird die Korrespondenzanalyse als Hilfsmittel zur chronologischen Ordnung von Funden oder Fundkomplexen verwendet, wobei es nicht an Kritik gefehlt hat. Diese ist zum Teil berechtigt, zum Teil unberechtigt.

Zunächst könnte man argumentieren, die Lösung sei nicht eindeutig. Tatsächlich gibt es insgesamt r Lösungsmöglichkeiten, wobei r das kleinere von m und n ist. Zu jeder dieser Lösungen gehört ein Korrelationskoeffizient ρ . Eine Lösung ist trivial mit $x_i = 1$ und $y_K = 1$ für alle i und k; zu ihr gehört $\rho=\rho_0=1$. 1. Ist das nächstgrößte $\rho=\rho_1$ größer als die anderen, ist die zugehörige Lösung x_{1i} , y_{1k} des Reciprocal Averaging eindeutig. Dies ist insbesondere dann der Fall, wenn sich die Matrix einigermaßen deutlich diagonalisieren ließ. Warum das so ist, geht aus der Eigenwertstheorie von Matrizen hervor, auf die hier nicht näher eingegangen werden kann.

Man kann nun für eine Reihe von Modellen zeitlicher Variationen zeigen, daß die beschriebene Methode tatsächlich die - bis auf Zufallsschwankungen - richtige Lösung liefert. Bei einem Modell

wird davon ausgegangen, daß pro Zeiteinheit ein Typ verschwindet und ein neuer hinzukommt. Ein Beispiel ist die Parallelogramm-Matrix

```
1  1  1  1  .  .  .
.  1  1  1  1  .  .
.  .  1  1  1  .  .
.  .  .  1  1  1  1
         ...
```

(statt der Null wurde der Übersichtlichkeit wegen ein Punkt gemacht).

Ein anderes Beispiel ist die Zeitabhängigkeit von Zunahme und Wiederabnahme der Typenwahrscheinlichkeit (-häufigkeit) in Form einer Glockenkurve. Daß es so etwas in der Praxis gibt, zeigen die Abbildungen 1 und 2.

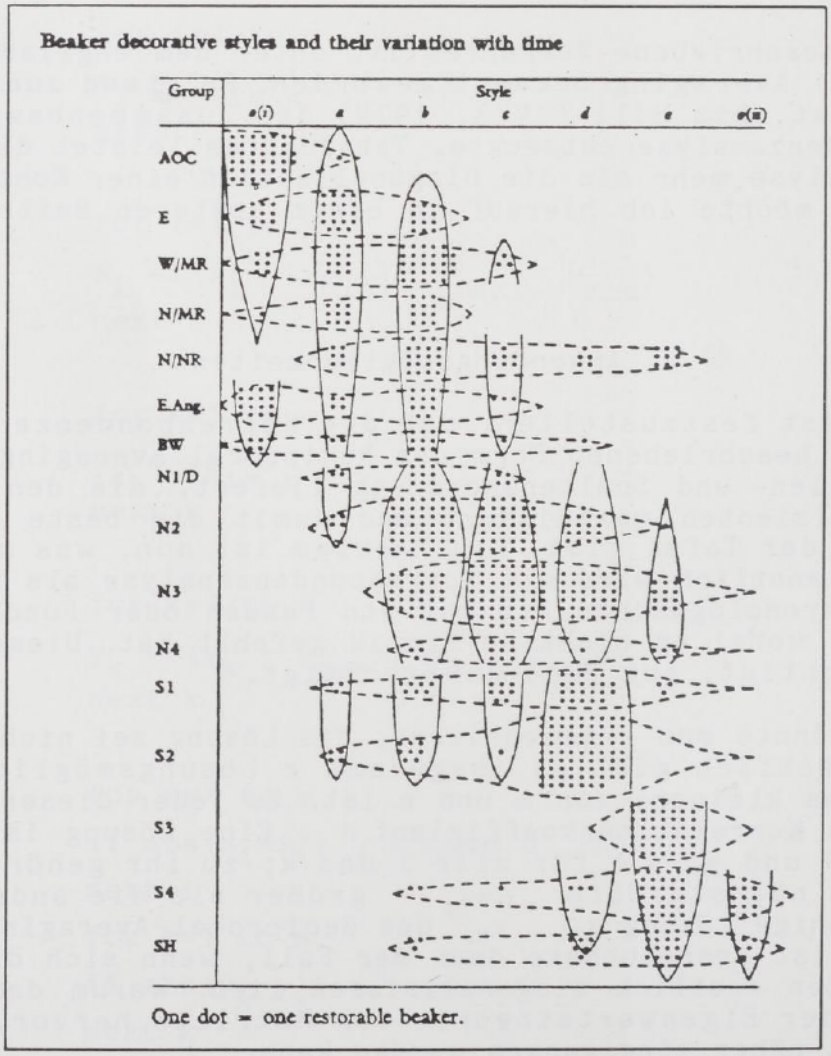


Abb. 1

Chronologische Ordnung von Glockenbechern; aus Clarke (1970).

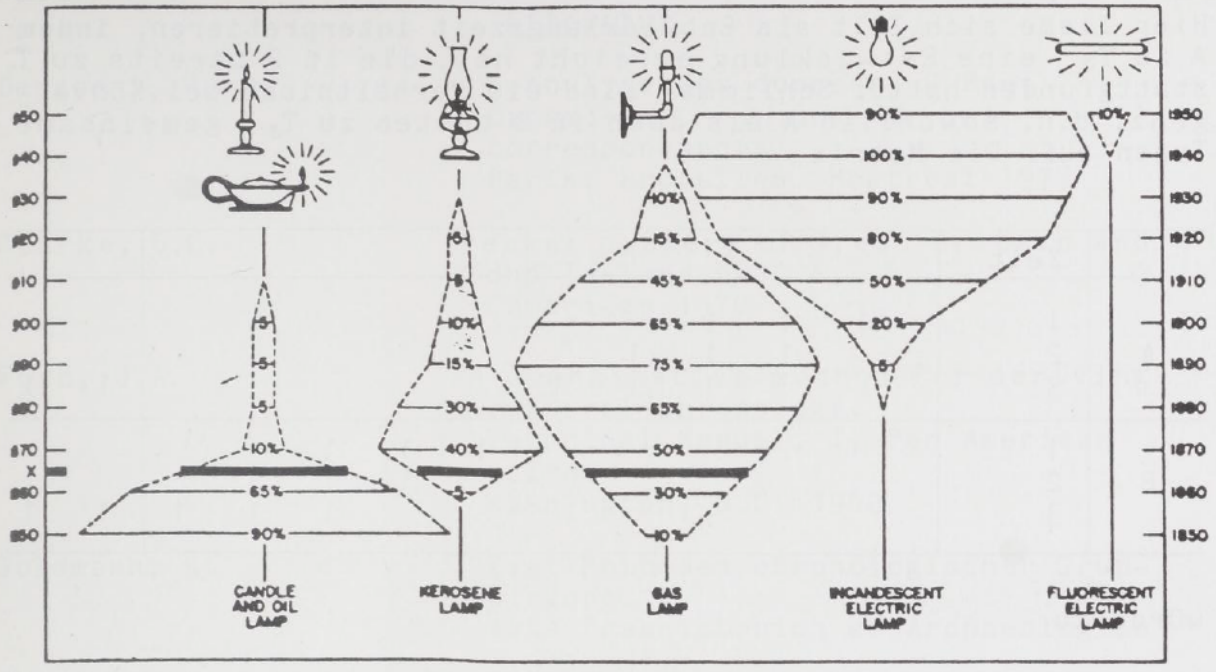


Abb.2 Der Anteil verschiedener Beleuchtungsvorrichtungen in Pennsylvania,U.S.A.von 1850 bis 1950; aus Ford (1962)

Ein Beispiel dafür gibt auch Ziegert (1983, Abb. 2). Ziegert weist aber mit Recht darauf hin, daß man Beispiele konstruieren kann, bei denen eine mehr oder weniger vollkommene Diagonalisierung nicht zu der gewünschten chronologischen Ordnung führt. Wir nehmen einmal je eine Parallelogramm-Matrix für eine Region A und eine getrennte Region B an, deren Zeilen die gleiche astronomische Zeitspanne von T_1 bis T_3 wiedergeben, deren Spalten aber keine gemeinsamen Typen aufweisen. Wir erhielten dann z.B. als gemeinsame Matrix

	Zeit											
	1	1	1	1								
A	2		1	1	1							
	3			1	1	1						
	1						1	1	1			
B	2							1	1	1		
	3								1	1	1	

Hier wird tatsächlich eine falsche Chronologie vorgetäuscht, aber ich glaube, daß man das erkennen kann. Problematischer wäre das Auftreten einer Brücke zwischen den Blöcken, z.B.

	Zeit											
	1	1	1	1								
A	2		1	1	1							
	3			1	1	1	1					
	1					1	1	1	1			
B	2							1	1	1		
	3								1	1	1	

Hier ließe sich Zeit als Entwicklungszeit interpretieren, indem A zu T_3 eine Entwicklung erreicht hat, die in B bereits zu T_1 stattgfunden hatte. Schlimmer sind die Verhältnisse bei Konvergenz, d.h. sowohl in A als auch in B treten zu T_3 gemeinsame Typen auf. Die Matrix

	Zeit														
A	1		1	1	1										
	2			1	1	1									
	3				1	1	1							1	1
B	1							1	1	1					
	2								1	1	1				
	3									1	1	1		1	1

würde zu

	Zeit														
A	1		1	1	1										
	2			1	1	1									
	3				1	1	1	1	1						
B	3							1	1	1	1	1			
	2										1	1	1		
	1											1	1	1	

umgeordnet. Bei kritischer Betrachtung des Ergebnisses dürfte man aufgrund allgemeiner Kenntnisse aber erkennen können, daß die zeitliche Ordnung hier gegeneinander läuft, und auf das Phänomen der Konvergenz aufmerksam werden.

Welche Schlußfolgerungen sollen aus diesen Beispielen gezogen werden? Sicher nicht: "Das Kind mit dem Bade ausschütten" und die Korrespondenzanalyse ablehnen. Sie ist zunächst ein brauchbares Hilfsmittel zur Diagonalisierung einer Kontingenztafel bzw. Datenmatrix und erhält ihren Wert aufgrund des Gewinnes an Übersichtlichtkeit. Ob eine erhaltene Ordnung chronologisch ist, folgt nicht allein aus einer möglicherweise eindrucksvollen Diagonalisierung. Es müssen vielmehr Argumente dafür gefunden werden, daß tatsächlich ein chronologisches Variationsmodell im oben genannten Sinne erwartet werden kann. Dies kann dann der Fall sein, wenn die Relevanz der Typen oder Merkmale für die Chronologie bekannt ist. Wer eine hervorstechende Diagonalisierung einer Tabelle erhalten hat, ohne derartige Vorkenntnisse zu besitzen, kann nichts anderes tun als die näheren Umstände prüfen, die zu dieser Anordnung geführt haben. Ist es nicht die Zeit - was sonst? Hier beginnt die Suche nach Erkenntnis.

Literatur

Benzecri, J.-P. — L'analyse des données. 2 Bde: Lataxinomie. L'analyse des correspondances. Paris, Bruxelles, Montreal 1973

Clarke, D.L. — Beaker pottery of Great Britain and and Ireland, Bd. 1. Cambridge 1970

Ford, J.A. — A quantitative method for deriving cultural chronology. Technical Manual, I, Pan American Union. Washington, D.C. 1962

Goldmann, K. — Zwei Methoden chronologischer Gruppierung. Acta Praehistorica et Archaeologica 3, 1 - 34.

Hill, M.O. — Reciprocal averaging: an eigenvector method of ordination. J. Ecology 61, 237-249 (1973)

Hill, M.O. — Correspondence analysis: a neglected multivariate method. Applied Statistics 23, 240-354 (1974)

Hirschfeld, H.O. — A connection between correlation and contingency. Proc. Cambr. Phil. Soc. 31, 520-524 (1935).

Kendall, M.G. und Stuart, A. — The advanced theory of statistics. Bd. 2. Inference and relationship. London 1961.

Ziegert, H. — "Kombinations-Statistik" und "Seriation". Zu Methode und Ergebnis der Bronzezeit-Chronologie K. Goldmanns. Archäologische Informationen 5, 21-52 (1983).

Prof. Dr. Peter Ihm
Institut für medizinisch-biologische Statistik
und Dokumentation der Philips-Universität
Ernst-Giller-Str. 20, 3550 Marburg