EINE EINFÜHRUNG IN DIE EXPLORATIVE DATENANALYSE

von Manfred Borovcnik, Klagenfurt

Kurzfassung: Techniken der Explorativen Datenanalyse (EDA) werden, in Fallstudien eingebettet, dargestellt. Diese Darlegung trifft den Kern der EDA selbst. Es handelt sich hiebei nämlich um eine besondere Art, mit Anwendungen von Statistik umzugehen. Dies zu vermitteln, ist Hauptanliegen des Autors.

Explorative Datenanalyse (EDA) ist seit 1970 sehr populär geworden. Neben der Abänderung traditioneller Methoden der Beschreibenden Statistik wurden eine Reihe eigener Verfahren entwickelt. Ein wichtiges Ziel dabei war es, die mathematischen Ergebnisse "theoriefrei", d.h. direkt verstehbar zu machen. Daß man also, gerade weil man keine weitere Theorie dazu lernen muß, die Ergebnisse von der Sache, von der Realität her, verstehen kann. Dies sollte in erster Linie Anwendern helfen, ihre Probleme sinnvoll zu bearbeiten. Im folgenden werde ich einfache Techniken der EDA im Rahmen von Fallstudien vorstellen und die enge Verschränkung von Realität und Mathematik darin aufzeigen.

1. Stamm-und-Blatt (St&Bl)

Beispiel: Beherbergungsbetriebe in Klagenfurt

Rohdaten

Tab. 1: Beherbergungsbetriebe in Klagenfurt nach Bettenzahl

Name des Betriebs	Betten	Name des Betriebs	Betten
Aragia	115	Bozener Weinstube	22
Blumenstöckl	37	Geyer	50
Dermuth	80	Jenull	4
Europapark	60	Kärntner Hamatle	10
Flughafenhotel	24	Klepp	21
Goldener Brunnen	50	Kollmann "Roko-hof"	100
Hopf	50	Lindenkeller	40
Janach	40	Marktl	16
Kurhotel Carinthia	42	Müller	30
Löwenkeller	30	Mozarthof	37
Mondschein	64	Plattenwirt	58
Moser-Verdino	140	Ratzmann	6
Musil	29	Schloßwirt	36
Porcia	80	Schweizerhaus	6
Sandwirt	80	Seeblick	18
Wörthersee	60	St. Primus (Egger)	9
Waidmannsdorferhof	44	Strauß	24
Zentral	25	Wachau	30
Zlami	50	Wadler	20
ÖJHV-Jugendherberge	140	Waldwirt	31

Die Rohdaten sind alphabetisch angeordnet. Das erleichtert das Aufsuchen eines bestimmten, namentlich bekannten Betriebes. Tiefschürfende Einsichten sind daraus jedoch nicht zu erwarten.

Ordnen der Daten

Einen ersten Überblick über die Bettenzahlen erhält man, wenn man die Daten (die Beherbergungsbetriebe) der Größe nach anordnet: 4,6,6,9,10,16,18,20,21,22,24,24,25,25,29,30,30,30,31,36,37,37,40, 40,42,44,50,50,50,56,60,64,80,80,100,115,140,140.
Daraus kann man unmittelbar den größten und kleinsten Wert mit 140 bzw. 4 ablesen. Im Sachzusammenhang etwa ist es auffällig, daß eine kleine Stadt wie Klagenfurt über zwei Betriebe mit 140 Betten verfügt, andererseits, daß Betriebe mit unter zehn Betten auch in dieser Liste geführt werden.

Ordnen im Stamm-und-Blatt

In der EDA verwendet man einen anderen Algorithmus zum Ordnen der Daten. Dieser liefert selbst schon ein Bild, ein sogenanntes Stamm-und-Blatt (St&Bl). Dabei wird jede Zahl in einen Stamm und ein Blatt zerlegt, z.B.: 37 in 3|7. Durch das fortlaufende Anschreiben der Zahlen entsteht ein histogrammartiges Gebilde von der Verteilung der Daten. Um diesen Verteilungscharakter hervorzuheben, ist es manchmal von Vorteil, in der Ausgangsliste zwei oder fünf oder gar zehn Zeilen zu einer neuen zusammenzufassen.

Fig. 1: Bettenzahl Klagenfurter Betriebe - Nach Größe sortiert

0	4,6,6,9		4,6,6,9
10	10,16,18		10,16,18
20	24,29,25,22,25,21,24,20		20,21,22,24,24,25,25,29
30	37,30,30,37,36,30,31		30,30,30,31,36,37,37
40	40,42,44,40		40,40,42,44
50	50,50,50,56,		50,50,50,56
60	60,64,60		60,60,64
70		sortiert ->	
80	80,80,80		80,80,80
90			
100	100		100
110	115		115
120			
130			
140	140,140		140,140

Das Bild zeigt: Bei den Bettenzahlen (20-30),(30-40) "häufen" sich die meisten Betriebe; wenige Betriebe haben mehr als 60 Betten; die Verteilung ist nicht symmetrisch; sie zerfällt in zwei "Cluster", die "normal großen" und die "großen" Betriebe. Dies ist ein Anlaß, nachzudenken, was die "Ursachen" dafür sind: Welche weiteren Eigenschaften sind den großen Beherbergungsbetrieben noch eigen, die den kleineren Betrieben nicht zukommen?

Hier erweist es sich von Vorteil, daß man im St&Bl, im Gegensatz zum gewöhnlichen Histogramm, die einzelnen Daten noch kennt. Daher kann man noch den Beherbergungsbetrieb ausfindig machen, der dazu gehört. Dies wird noch erleichtert, wenn man in die betreffende Zeile des St&Bl einen Code statt der Zahl schreibt oder wenn man das St&Bl hinsichtlich des Clusters der größten Werte beschriftet. Dann kann man z.B. die interessante Feststellung machen, daß alle "großen" Betriebe mit Ausnahme der Jugendherberge zu den Klagenfurter Traditionsbetrieben zählen. Diese sind zu einer Zeit gegründet worden, als es vermutlich üblich war, die Betriebe größer auszulegen.

Fig. 2: St&Bl mit Codes für den Cluster der größten Werte

0	4,6,6,9		
10	10,16,18		
20	24,29,25,22,25,21,24,20		
30	37,30,30,37,36,30,31		
40	40,42,44,40		
50	50,50,50,56,		
60	60,64,60		
70			
80	80,80,80	DERM,PORC,SAND	
90			
100	100	ROKO	
110	115	ARAG	
120			
130			
140	140,140	MOVE,ÖHJV	

DERM	Dermuth
PORC	Porcia
SAND	Sandwirt
ROKO	Rokohof
ARAG	Aragia
MOVE	Moser/Verdino
ÖHJV	Jugendherberge

Im St&Bl sind die einzelnen Daten identifizierbar, d.h., man kann den "Objektträger" ausfindig machen. Daher kann man beliebig anderes (auch informelles) Wissen über ihn in Erfahrung bringen oder einbringen. Die Technik der Codierung der Darstellung mit gut verständlichen Kurzbezeichnungen unterstützt diese Identifizierung. All das unterstreicht, daß man sich gerade aus dem Bezug der mathematischen Darstellung zur Realität weiterreichende "Einsichten" erhofft.

2. Variationen des St&Bl

Beispiel: Tore im europäischen Fußball
Stimmt es, daß im Süden Europas weniger Tore geschossen werden als im Norden? Die BRD steht mit 3.35 Toren pro Spiel an der Spitze, Österreich nimmt mit 2.92 Toren Rang 9 unter 30 Nationen ein (Tab. 2). Die Verteilung der Trefferquoten wird erst aus einem Histogramm oder einem St&Bl ersichtlich.

Rohdaten

Tab. 2: Treffer pro Spiel im Fußball in Europa 1982/83

Rang	Land	Trefferquote	Rang	Land	Trefferquote
1	BRD	3.35	16	England	2.73
2	Finnland	3.32	17	Norwegen	2.68
3	DDR	3.25	18	Bulgarien	2.66
4	Schweiz	3.23	19	Jugoslawien	2.65
5	Luxemburg	3.16	20	Schweden	2.62
6	Niederlande	3.15	21	Spanien	2.54
7	Ungarn	3.12	22	UdSSR	2.54
8	Nordirland	3.02	23	Griechenland	2.39
9	Österreich	2.92	24	Zypern	2.39
10	CSSR	2.88	25	Portugal	2.37
11	Schottland	2.88	26	Polen	2.29
12	Frankreich	2.87	27	Island	2.16
13	Eire	2.83	28	Italien	2.10
14	Dänemark	2.82	29	Türkei	2.09
15	Belgien	2.81	30	Malta	1.94

Hälften der Daten - Vergleich: Süden-Norden

Die Länder sollen (künstlich) in zwei Hälften eingeteilt werden, je nach geographischer Lage soll ein Land einer der Gruppen Süden bzw. Norden zugeordnet werden. Die Sowjetunion scheint hiebei nicht zuordenbar, sie und zum Ausgleich noch Luxemburg werden aus der Wertung genommen.

Tab. 3: Trefferquoten - Länder nach geographischer Lage geordnet

"von Norden"			"von Süden"		
Rang	Staat	Quote	Rang	Staat	Quote
27	Island	2.16	24	Zypern	2.39
2	Finnland	3.32	29	Türkei	2.09
17	Norwegen	2.68	30	Malta	1.94
20	Schweden	2.62	23	Griechenland	2.39
11	Schottland	2.88	25	Portugal	2.37
14	Dänemark	2.82	21	Spanien	2.54
8	Nordirland	3.02	28	Italien	2.10
13	Eire	2.83	18	Bulgarien	2.66
16	England	2.73	19	Jugoslawien	2.65
6	Niederlande	3.15	7	Ungarn	3.12
15	Belgien	2.81	4	Schweiz	3.23
3	DDR	3.25	9	Österreich	2.92
26	Polen	2.29	12	Frankreich	2.87
1	BRD	3.35	10	CSSR	2.88

Mittelwerte

Die übliche Beurteilung der Unterschiede basiert auf Mittelwerten, die mittlere Torquote im Süden beträgt $\bar{x}_S = 2.58$, im Norden $\bar{x}_N = 2.85$. Der Unterschied beträgt demnach 0.27 Tore pro Spiel. Die Beurteilung, ob der Unterschied in den Mittelwerten groß oder klein ist, ist eigentlich schwierig. Das hängt damit zusammen, daß gar nicht so leicht zu verstehen ist, was der Mittelwert für eine Datenliste überhaupt bedeutet. Man kann auch mit Fußballfachwissen den Unterschied von 0.27 nicht besser erklären. In der Beurteilenden Statistik wird die Differenz $\bar{x}_N - \bar{x}_S = 0.27$ dem sogenannten t-Test unterworfen. Dieser prüft, ob der Wert 0.27 "signifikant" von 0 verschieden ist, ob man statistisch es als gesichert betrachten kann, daß es einen Unterschied in der Trefferquote zwischen Norden und Süden gibt, oder nicht. Eine andere Art der Prüfung der Unterschiede in den Trefferquoten besteht in der Anwendung des Wilcoxon-Tests.

Zweiseitiges St&Bl

Eine explorative Alternative zur Beurteilung der Unterschiede zwischen Norden und Süden hinsichtlich der Trefferzahl pro Spiel ist ein zweiseitiges St&Bl. Das "Laub" für die Gruppe Norden wird dabei links, jenes für die Gruppe Süden rechts vom Stamm angeordnet. Es besteht gerade aus der Hundertstel-Stelle der Daten.

Fig. 3: Zweiseitiges St&Bl: Süden - Norden

	25	3.3	
	5	3.2	3
	5	3.1	2
	2	3.0	
		2.9	2
	1238	2.8	78
	3	2.7	
	28	2.6	65
		2.5	4
		2.4	
		2.3	997
	9	2.2	
	6	2.1	0
		2.0	9
NORDEN		1.9	4 SÜDEN

Aus den einzelnen Seiten der Darstellung erhält man einen Eindruck von der Verteilung in der zugehörigen Gruppe. Hierzu müßte man allerdings noch je zwei Zeilen zusammenfassen, um die Darstellung kompakter zu gestalten. Diese Technik wird später erläu-

tert werden. Ferner gibt das zweiseitige St&Bl durch die Gegenüberstellung einen direkten Vergleich der Trefferquoten zwischen Süden und Norden. Der Unterschied ist deutlich sichtbar.

Codiertes St&Bl

Werden im zweiseitigen St&Bl wenigstens die extremeren Daten mit namentlichen Codes gekennzeichnet, so bleibt der geographische Bezug zu den Daten aufrecht. Man erkennt dann deutlich, daß die hohen Trefferquoten im Süden eigentlich von mitteleuropäischen Ländern und von Frankreich stammen. Die Zuordnung zu den Gruppen Süden und Norden scheint der ursprünglichen Fragestellung nicht angemessen zu sein. Dieses Feedback ist ein Lohn für aufrechterhaltene Verbindung zwischen der mathematischen Darstellung der Daten und der Identität der Daten.

Fig. 4: Codiertes zweiseitiges St&Bl - Süden-Norden

Finnland, BRD	25	3.3		
DDR	5	3.2	3	Schweiz
Niederlande	5	3.1	2	Ungarn
	2	3.0		
		2.9	2	Österreich
	1238	2.8	78	Frankreich, CSSR
	3	2.7		
	28	2.6	65	
		2.5	4	
		2.4		
		2.3	997	
Polen	9	2.2		
Island	6	2.1	0	Italien
		2.0	9	Türkei
		1.9	4	Malta
NORDEN				SÜDEN

3. Vierfeldertafel

Beispiel: Tore im europäischen Fußball (Fortsetzung)

Vierfeldertafel mit Häufigkeiten

Die Länder wurden geographisch in zwei Hälften eingeteilt. Eine weitere Unterteilung wäre die nach der Trefferquote, und zwar in Länder mit "hoher" und solche mit "niedriger Trefferquote". Fordert man wieder, daß die entstehenden Gruppen gleich groß sind, so erhält man den Median, $\tilde{x} = (2.73+2.81)/2 = 2.77$, als Trennpunkt. Durch diese doppelte Klassifizierung erhält man vier Gruppen. Österreich z.B. zählt zur Gruppe S - "viele Tr.". Aus dem zweiseitigen St&Bl zählt man direkt ab, wie viele Länder in die jeweilige Gruppe fallen. Das Ergebnis wird in Matrixform notiert (Tab. 4). Der Unterschied zwischen Süden und Norden, der sich aus dem zweiseitigen St&Bl angedeutet hat, ist nun überdeutlich.

Tab. 4: Trefferquoten - Numerische Vierfeldertafel

	viele Tr.	wenige Tr.
NORDEN	9	5
SÜDEN	5	9

Vierfeldertafel mit Codes

Nachteil der Vierfeldertafel ist, daß der geographische Bezug nur sehr grob ist. Die aus dem codierten zweiseitigen St&Bl gewonnene Einsicht, daß die Ergebnisse für S durch mitteleuropäische Länder und Frankreich verfälscht werden, ist nun verschleiert. Dieser Bezug zur Realität läßt sich aber leicht wieder herstellen, indem man statt der Häufigkeiten in die Felder der Matrix Codes für die Länder einträgt. Man sieht deutlich (Tab. 5): Hohe Trefferquoten im Süden werden ausschließlich von nicht-südeuropäischen Ländern "verursacht". Diese sachliche Einsicht führt zum Vergleich Mittelmeerländer gegen den Rest Europas (Tab. 6).

Tab. 5: Trefferquoten: Codierte Vierfeldertafel

	viele Treffer			wenige Treffer		
NORDEN	D	SF	DDR	S	ENG	N
	NL	IRL	NIRL			
	DK	SCO	B	PL	IS	
SÜDEN	CH	H	A	BG	YU	E
				GR	CY	P
	CS	F		I	TR	M

Tab. 6: Trefferquoten: Vergleich Mittelmeerländer - andere Länder

	viele Tr.	wenige Tr.	SUMME
Mittelmeer	0	9	9
Rest	14	5	19
SUMME	14	14	28

Eine ganz wesentliche Triebkraft der vorangehenden Analyse bestand darin, daß man Zwischenresultate der mathematischen Behandlung sofort an der Sache, an der Realität, überprüft und die weitere Analyse danach modifiziert hat. Dies war insbesondere dadurch möglich, daß die Daten z.T. auch während der mathematischen Bearbeitung noch identifizierbar waren, sodaß man je nach Bedarf darüber bestimmtes Wissen in die Analyse einbringen konnte.

4. Verdichtetes Stamm&Blatt

Beispiel: Niederschläge in Afrika

Die Stationen sind von Norden nach Süden angeordnet. Die Reihenfolge läßt grob die Klimazonen erkennen, von tropisch-trocken (Wüsten) über tropisch-feucht (Regenwälder) bis tropisch-trocken (Steppen und Wüsten). Die Verteilung der Niederschläge soll im Hinblick auf diese Klimazonen analysiert werden.

Rohdaten

Tab. 7: Niederschläge in Afrika in mm

Port Sudan	110	Karima	27	Faya Largeau	22
Chartum	177	Kassala	345	Mao	312
Abecher	494	Ati	469	El Obeid	369
Mongo	1059	Fort Lamy	642	Am-Timan	870
Tamale	1104	Fort Archambault	1143	Moundou	1270
N'Dele	1231	Wau	1115	Kumasi	1482
Bria	1563	Bouca	1494	Bouar	1382
Accra	728	Bangassou	1744	Bangui	1535
Berberati	1506	Bitam	1891	Ouesso	1567
Impfondo	1772	Coco Beach	3422	Mitzic	1853
Libreville	2736	Entebbe	1143	Port Geatil	1900
Lambarene	2039	Franceville	1851	Mouila	2301
Gamboma	1787	Djambola	1938	M'Pouyo	1129
Mayumba	1719	Dolisic	1373	Brazzaville	1394
Mahe	2322	Pointe Noire	1228	Luanda	358
Diego-Suarez	963	Dzaoudsi	1081	Majunga	1521
Maintirano	962	Tamatave	3475	Tananarive	1291
Maun	438	Windhoek	354	Tulear	355
Pietersburg	490	Fort Dauphin	1565	Pretoria	753
Jan Smuts	690	Keetmanshoop	134	Upington	189
Alexander Bay	46	Kimberley	381	Bloemfontein	555
Durban	975	Beaufort West	238	East London	791
D. F. Malan	456	Port Elizabeth	626		

Histogramm

Die Verteilungsart wird üblicherweise durch ein Histogramm untersucht. Dabei ist die Klasseneinteilung so geschickt zu wählen, daß ein flächiger Eindruck der Verteilung der Daten entsteht.

Verdichten des St&Bl

Dem Histogramm entspricht die EDA-Technik des St&Bl. Im folgenden wird erläutert, wie das St&Bl so abgeändert werden kann, daß ein entsprechend flächiger Eindruck von der Verteilung der Daten entsteht. Die Daten liegen etwa zwischen 0 und 3500 mm Niederschlag, die Blätter müssen demnach aus Zehner- und Einerstelle der Daten gebildet werden, z.B.: 2736 wird in 27|36 zerlegt.

Fig. 5: Niederschläge: Afrika - St&Bl mit zweistelligen Blättern

```
 H  ZE
 0 |27,22,46              Karima, Faya Largeau, Alexander Bay
 1 |10,77,34,89
 2 |38
 3 |45,12,69,58,54,55,81
 4 |94,69,38,90,56
 5 |55
 6 |42,90,26
 7 |28,53,91
 8 |70
 9 |63,62,75
10 |59,81
11 |04,43,15,43,29
12 |70,31,28,91,
13 |82,73,94
14 |82,94
15 |63,35,06,67,21,65
16 |
17 |44,72,87,19
18 |91,53,51
19 |00,38
20 |39
21 |
22 |
23 |01,22
24 |
25 |
26 |
27 |36                     Libreville
 : |
 : |
34 |22,75                  Coco Beach, Tamatave
```

Die Punkte im Stamm deuten an, daß etwas fehlt. Das St&Bl hat 34 Zeilen und bietet keinen Überblick über die Verteilung der Niederschlagsdaten. Die Wirkung der Darstellung wird verdichtet durch Kappen der Daten auf zwei geltende Stellen (2736 wird zu 27), oder durch Runden, oder durch Zusammenfassen von 2, 5 oder 10 Zeilen. Faßt man je fünf Zeilen zusammen, so wird die Darstellung zu grobklotzig, faßt man nur zwei zusammen, so erhält man ein geeignet flächiges Bild (Fig. 6).

Ein verdichtetes St&Bl ist ein Histogramm mit speziellen Klassengrenzen, das um 90^{o} gedreht wird. Im Histogramm wird jedes Datum

durch eine bestimmte Fläche dargestellt, die einzelnen Daten sind nicht identifizierbar. Durch das Vergröbern der Daten auf zwei geltende Stellen wird zwar die Identifikation auch erschwert, aber nicht prinzipiell unmöglich gemacht. Die extremen Daten sind in der Darstellung nach wie vor namentlich gekennzeichnet. Der Umriß dieses St&Bl ist U-förmig, mit einem schwachen, aber langgezogenen Ausläufer nach oben. In den zwei Gipfeln kann man die beiden Hauptklimate des Kontinents (tropisch-trocken bzw. tropisch-feucht) deutlich wiedererkennen.

Fig. 6: Niederschläge in Afrika - St&Bl auf zwei geltende Stellen gekappt. Je zwei alte Zeilen zusammengefaßt

```
TH  H
00 |0001111
02 |23333333
04 |444445
06 |666777
08 |8999
10 |0011111
12 |2222333
14 |44555555
16 |7777
18 |88899
20 |0
22 |33
24 |
26 |7
28 |
30 |
32 |
34 |44
```

5. St&Bl zum Vergleich von Verteilungen

Beispiel: Niederschläge in Afrika, Südamerika und Australien
Angestrebt ist ein Vergleich der Klimate in diesen Erdteilen.

Rohdaten

Tab. 8: Niederschläge in Australien in mm

Thursday Island	1691	Darwin	1492	Daly Waters	638
Broome	584	Halls Creek	477	Townsville	1010
Cloncurry	428	Onslow	238	Nullagine	314
Rockhampton	950	Longreach	394	Alice Springs	255
Carnarvon	230	Charleville	414	Meekatharra	230
Dalby	637	Brisbane	1020	Oodnadatta	116
Marree	144	Bourke	298	Kalgoorlie	242
Perth	915	Broken Hill	233	Ceduna	321
Dubbo	553	Kattanning	494	Sydney	1139
Adelaide	583	Canberra	608	Deniliquin	392
Auckland	1281	Mt. Gambier	683	Melbourne	657
Napier	800	Nelson	995	Wellington	1268
Western Junction	727	Hokitika	2864	Hobart	636
Christchurch	658	Dunedin	791		

Tab. 9: Niederschläge in Südamerika in mm

Maracaibo	596	Maracay	969	Merida	1936
Ciudad Bolivar	1044	San Fernando	1302	Cayenne	3922
Belem	2733	Turiacu	2193	Olinda	1601
Porto Nacional	1814	Aracaju	1218	Salvador	1914
Utiariti	2059	Cuiaba	1269	Arica	0
Belo Horizonte	1561	La Quiaca	296	Antofagasta	0
Rio de Janeiro	1050	Sao Paolo	1323	Salta	712
Curitiba	1364	Tucuman	974	Posadas	1589
Catamarca	367	La Rioja	321	Alegrete	1683
La Serena	124	Porto Alegre	1298	Cordoba	714
Concordia	1021	San Juan	94	Mendoza	210
Rosario	960	Valparaiso	459	San Luis	513
Santiago	370	Sto. Vitoria	1179	Mar del Plata	751
Juan Fernandez	980	Valdivia	2455	Puerto Monte	1900
Trelew	167	Isla Guafo	1175	Sarmiento	135
Punta Arenas	431				

Vergleich von Histogrammen und Mittelwerten

In der Beschreibenden Statistik ist es üblich, Unterschiede in Verteilungen durch Histogramme oder durch Mittelwerte zu erfassen. Der Nachweis, ob Unterschiede in den Mittelwerten bestehen, kann formal durch die Varianzanalyse erfolgen. Die Voraussetzungen für dieses Verfahren (zufällige Daten aus einer Normalverteilung) sind hier nicht erfüllt. Darüberhinaus ist die mittlere Niederschlagsmenge eine unwesentliche Kennziffer, sodaß ein Vergleich der Daten, der darauf basiert, nur wenig informativ sein kann. Es geht ja um einen Vergleich von Klimaprofilen.

Vergleich von St&Bl

Fig. 7: Vergleich der Niederschläge in Afrika, Südamerika und Australien mit verdichteten St&Bl-Darstellungen

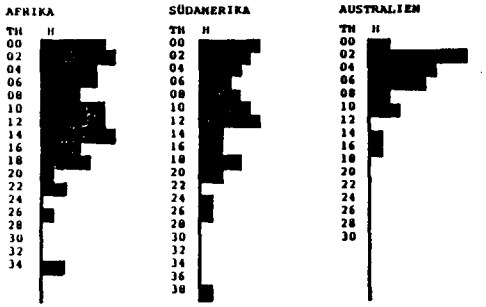
```
AFRIKA
TH  H
00 |0001111
02 |23333333
04 |444445
06 |666777
08 |8999
10 |0011111
12 |2222333
14 |44555555
16 |7777
18 |88899
20 |0
22 |33
24 |
26 |7
28 |
30 |
32 |
34 |44
```

```
SÜDAMERIKA
TH  H
00 |000111
02 |22333
04 |4455
06 |777
08 |9999
10 |00011
12 |222333
14 |55
16 |66
18 |8999
20 |01
22 |
24 |4
26 |7
28 |
30 |
32 |
34 |
36 |
38 |9
```

```
AUSTRALIEN
TH  H
00 |11
02 |22222223333
04 |4444555
06 |666666677
08 |8999
10 |001
12 |22
14 |4
16 |6
18 |
20 |
22 |
24 |
26 |
28 |8
30 |
```

In der EDA vergleicht man St&Bl-Darstellungen anstelle von Histogrammen. Die Unterschiede in den Niederschlagsdaten können demnach in zwei Hauptrichtungen beurteilt werden:
i) Wie unterscheiden sich die Niederschlagsverteilungen in den Hauptclustern: breiter, schmäler, U-förmig, eingipfelig etc? Welche Klimazonen finden sich wo? Wo nicht? In welchem Umfang? Dazu ist es von Vorteil, nur die Umrisse der St&Bl zu vergleichen.
ii) Wie sind die Ausreißer zu interpretieren? Beschriften der extremen Werte ist auch hier von Vorteil, z.B. sind die hohen Niederschlagsdaten in Australien durch Stationen in Neuseeland "verursacht".
Dieses Feedback kann dazu führen, daß man alle neuseeländischen Daten eliminiert. Der Vergleich dann zeigt noch deutlicher, wie sehr die australischen Daten auf Wüsten- und Steppenklima konzentriert sind. Auch hier führt die Rückkoppelung erster Ergebnisse an der Realität eventuell zu einer Modifikation der Analyse.

Fig. 8: Vergleich der Niederschlagsdaten von Afrika, Südamerika und Australien - ohne Neuseeland: Umrisse der St&Bl



6. Kennziffern

Zur Beschreibung von Daten sind Zahlen manchmal besser als Bilder.

Mittelwert und Standardabweichung

Üblicherweise gibt man die Lage bzw. das Zentrum einer Verteilung durch den Mittelwert $\bar{x}$ an, die Breite einer Verteilung wird durch die Standardabweichung s angegeben. Es wurde schon darauf hingewiesen, daß es durchaus schwierig sein kann, den daraus gewonnenen Zahlenwerten eine Interpretation abzugewinnen. Eine Hilfe bieten dabei die sogenannten s-Regeln:

s-Intervall	ca.Anteil an Daten darin
$[\bar{x}-s,\bar{x}+s]$	67%
$[\bar{x}-2s,\bar{x}+2s]$	95%
$[\bar{x}-3s,\bar{x}+3s]$	99%

Das bedeutet, daß ca. 95% der Daten im 2s-Intervall $[\bar{x}-2s,\bar{x}+2s]$ liegen. Ein Wert außerhalb kann demnach als extrem eingestuft werden. Der Haken an der Sache ist: Dies ist eine Faustregel, die idealerweise für die Normalverteilung zutrifft, für ausgeprägt eingipfelige Verteilungen brauchbare Aussagen liefert, in anderen Fällen (mehrgipfelig, Ausreißer etc.) jedoch versagen kann.

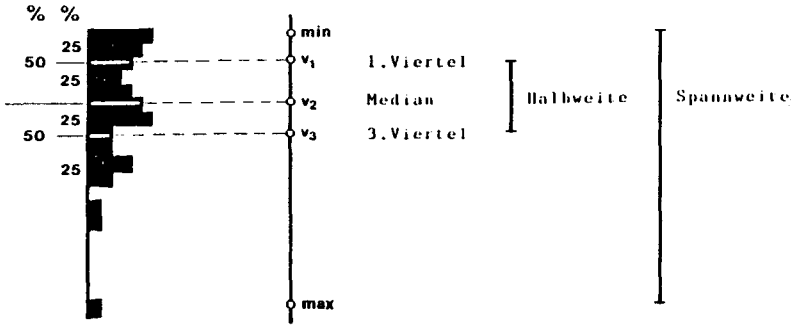
Median - Viertelpunkte

In der EDA ist die Rückkoppelung der mathematischen Ergebnisse an die Realität ein vordringliches Ziel, deshalb ist es wünschenswert, für Lage und Breite einer Verteilung Kennziffern parat zu haben, die unmittelbar zu verstehen sind. Folgender Algorithmus des Haufenhalbierens führt zu solchen Kennziffern:
Die Daten werden der Größe nach angeordnet und dann in eine untere Hälfte U und eine obere Hälfte O halbiert. Der Trennpunkt zwischen U und O heißt Median $\tilde{x}$ - er halbiert die Verteilung und zählt als ein Wert, der für alle Daten steht. Der untere Haufen U bzw. der obere Haufen O wird derselben Prozedur unterworfen. Man erhält damit das sogenannte 1.Viertel v_1 bzw. das 3.Viertel v_3. Zwischen v_1 und v_3 liegt die zentrale Hälfte der Daten, je ein Viertel liegt unter v_1 bzw. über v_3.

Bestimmen der Trennpunkte:	Anzahl der Daten	"Tiefe" des Trennpunktes
	A=46	
	A/2=23	23h
	A/4=11h	12

Halbiert man den Haufen von 46 Daten (Fig. 9), so verbleiben in den Teilhaufen U und O je 23 Daten, der Trennpunkt hat eine Tiefe von 23h, d.h. er liegt zwischen dem 23. und [illegible]größten Datum, dieser Trennpunkt heißt Median. Halbiert man den unteren Haufen U von 23 Daten, so fallen jedem Teilhaufen 11h (11.5) Daten zu, der Teilungspunkt, das 1.Viertel, ist der 12.größte Wert.

Fig. 9: Kennziffern von Lage und Breite einer Verteilung durch fortgesetztes Haufenhalbieren



Ermittelt man die Trennpunkte aus dem verdichteten St&Bl für Südamerika, so erhält man: $v_1 = 400$, $v_2 = \tilde{x} = 1000$, $v_3 = 1500$. Die genaueren Werte 431, 1031 und 1589 erhält man aus dem St&Bl mit zweiziffrigen Blättern.

7. Kastenschaubilder

Kennziffern für Lage und Breite einer Verteilung bieten viel Information für einen gezielten Vergleich mehrerer Verteilungen. Sie ist leichter zugänglich, wenn sie graphisch aufbereitet wird.

Der zentrale Kasten in Fig. 10 umfaßt die mittlere Hälfte der Daten. Für das Einzeichnen der Ausläufer gibt es unterschiedliche Regeln, z.B. Ausläufer für jenen Bereich, in dem die Verteilung einen geschlossenen Eindruck macht, oder Ausläufer so, daß unterhalb bzw. oberhalb dieser je 10% der Daten liegen. Die Daten jenseits der Ausläufer werden durch Punkte einzeln markiert und durch Beschriftung namentlich gekennzeichnet.

Die Kastenschaubilder (Fig. 11) zeigen deutlich, daß die Niederschläge für Afrika und Südamerika ziemlich ähnlich verteilt sind, während sich die Daten von Australien davon deutlich abheben: Die zentrale Box für Australien ist tiefer angesetzt und ist schmä-

ler, der Ausläufer nach oben ist kürzer. Dies als Ausdruck dafür, daß Australien einen wesentlich höheren Anteil am trockenen Steppen- und Wüstenklima hat. Man beachte: Die Daten für Neuseeland sind hierbei nicht eliminiert worden.

Fig. 10: Schrittweise Konstruktion eines Kastenschaubildes - Daten für Afrika

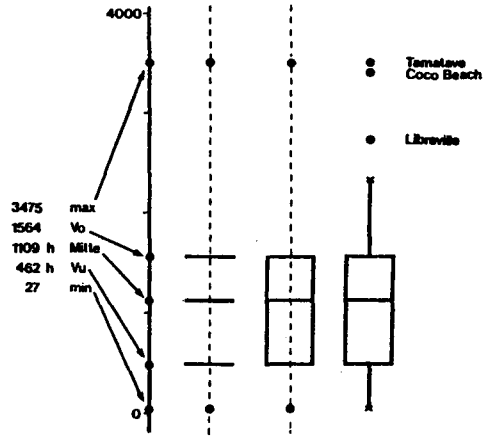
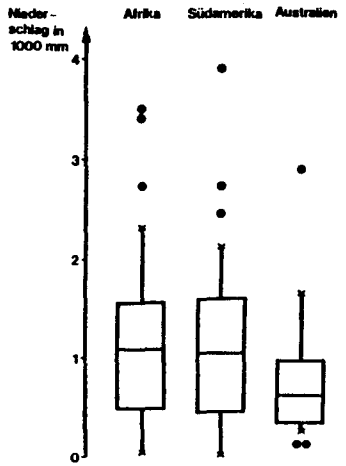


Fig. 11: Niederschläge in Afrika, Südamerika und Australien - Kastenschaubilder zum Vergleich



8. Abschließende Bemerkungen

Der Hintergrund der Verfahren ist ausführlich in Borovcnik/Ossimitz dargestellt. Die Beispiele sollen vermitteln, daß man in der EDA eine besondere Art pflegt, an Anwendungen von Statistik heranzugehen. Welche Verfahren in einem realen Problem angewendet werden, hängt nicht allein von einer übergeordneten Theorie ab, die vorgefertigte Modelle bereit stellt. Verfahren, die robust sind in bezug auf Annahmen, die üblichen Modellen zugrunde liegen, werden in der EDA bevorzugt oder neu entwickelt. Eine bedeutende Rolle spielt, daß der Anwender erst aus der Interpretation seiner Zwischenergebnisse interaktiv die weitere Bearbeitung des Problems bestimmt. Das erfordert einen kritischen Umgang mit den Ergebnissen. Damit die erforderlichen Interpretationen leichter fallen, sind die Konzepte in der EDA eigentlich trivial, man kann sie, ohne auf eine weitere Theorie Bezug nehmen zu müssen, direkt interpretieren. Durch visuelle Projektionen der Daten soll ihre "innenliegende" Struktur aufgedeckt und sachliche Einsichten erleichtert werden.

Für den Unterricht ergeben sich Chancen aber auch kritische Punkte: Techniken sind nicht eindeutig festgelegt, sie können nach "Notwendigkeit" verändert werden. Diese Offenheit muß erst bewältigt werden. Ebenso ist die Einsicht, daß Wissen oder der Umgang damit immer auch subjektiv ist, erst zu verdauen. EDA fördert keine formale, sondern eine sachverständige Auseinandersetzung mit einem Problem mit einfachen mathematischen Methoden.

Literatur:

Biehler, R.: Explorative Datenanalyse - Eine Untersuchung aus der Perspektive einer deskriptiv-empirischen Wissenschaftstheorie. Materialien und Studien Bd.24. Bielefeld: Insitut für Didaktik der Mathematik 1982.

Borovcnik, M. u. G. Ossimitz: Materialien zur Beschreibenden Statistik und Explorativen Datenanalyse.Wien/Stuttgart: Hölder-Pichler-Tempsky/Teubner 1987.

Polasek, W.: Explorative Datenanalyse. Einführung in die deskriptive Statistik. Berlin-Heidelberg-New York: Springer 1988.

Tukey, J.W.: Exploratory Data Analysis. Reading: Addison-Wesley 1977.