Manfred Borovcnik, Klagenfurt

EXPLORATIVE DATENANALYSE - TECHNIKEN UND LEITIDEEN

A) VORBEMERKUNGEN

Wer sich in einem bestimmten Gebiet auskennen will, wer in einem Sachproblem eine Entscheidung zu treffen hat, der hat sich ausreichend zu informieren. Es genügt jedoch nicht, über viel an Information zu verfügen, man muß darüber auch einen Überblick haben. (Beschreibende) Statistik beschäftigt sich mit Information, wie sie in Zahlen erfaßbar ist: Zu einem Sachverhalt werden verschiedenste Merkmale (Variable) erfaßt, Daten (Zahlen) werden dazu erhoben. Traditionelle Beschreibende Statistik stellt Techniken bereit, mit dieser statistischen Information umzugehen: Durch Bilder (Tortendiagramme, Staffelbilder etc.) bzw. durch Kennziffern (Mittelwerte, Trends etc.) soll diese Information aufbereitet, illustriert und konzise erfaßbar gemacht werden.

In Anwendungen von Mathematik hat man immer ein grundlegendes Problem zu beachten, nämlich daß man weiß, wie Ergebnisse von Modellberechnungen in der Realität zu verstehen und nachzuvollziehen sind und wie darauf aufbauend einsichtige Entscheidungen gefällt werden können. Diesem Problem der Verschränkung von mathematischen Methoden und Anwendungen in der Realität hat man natürlich auch im Rahmen der Beschreibenden Statistik Aufmerksamkeit geschenkt. Immer jedoch hat es Strömungen gegeben, die eine "technokratische" Anwendung, eine Anwendung, die auf die spezifischen Eigenheiten der jeweiligen Situation nicht ausreichend Rücksicht nimmt, forciert haben. Begünstigt wurden solche Strömungen dadurch, daß man die mathematischen Methoden der Beschreibenden Statistik nicht immer leicht verstehen kann.

Etwa mag vordergründig klar sein, was mit dem Konzept Mittelwert $\bar{x}$ von Daten gemeint ist. Wirklich verstehen kann man den Mittelwert jedoch erst, so meine These, wenn man die Beziehungen von $\bar{x}$ zur Normalverteilung bzw. zum zentralen Grenzverteilungssatz und zum Konzept der statistischen Vertrauensintervalle hergestellt

hat (siehe dazu [2], S.206f). Kurzum, der Mittelwert wird erst verstehbar, wenn man weitere Theorien, die Wahrscheinlichkeitstheorie und die Beurteilende Statistik, gelernt hat.

Als Gegenströmung dazu hat zunächst Tukey [6] mit seiner Explorativen Datenanalyse (EDA) seit ca. 1970 einen Akzent gesetzt, der gerade auf eine enge und direkte Verschränkung von Mathematik und Realität abzielt. Diese Richtung der Beschreibenden Statistik ist in der Folge sehr populär geworden. Neben der Abänderung traditioneller Methoden wurden eine Reihe eigener Verfahren entwickelt. Ein wichtiges Ziel war dabei, die mathematischen Ergebnisse direkt, d.h. "theoriefrei", 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. Das ist natürlich eine ideelle Zielvorstellung, in der Tendenz jedoch kann man den Methoden der Explorativen Datenanalyse diese Eigenheit zusprechen.

In der Didaktik der Mathematik ist die Diskussion um die Neue Mathematik durch die Anwendungsdiskussion abgelöst worden. Verschiedenste Ansätze wurden ausformuliert, in denen die Ausbildung und Stärkung der Kompetenz des Lernenden, reale Situationen zu mathematisieren, eine wichtige Rolle einnimmt. Eine fächerunabhängige Ausprägung stellt Projektunterricht bzw. Projektartiger Unterricht dar (siehe [3]). Derart ausgeprägter Unterricht soll eine Kompetenz im vernünftigen Anwenden mathematischer Modelle auf reale Situationen ermöglichen, andererseits soll er andere Arbeitshaltungen erfahrbar machen, wie z.B. eigenständige Arbeit, Teamarbeit, Verantwortung tragen etc. In dieser Stoßrichtung ist im neuen Lehrplan der AHS-Oberstufe in der 5. Klasse ein Abschnitt projektartiger Unterricht und ein Abschnitt Analyse von Daten eingefügt worden.

Im folgenden werde ich einfache Techniken der EDA im Rahmen von Fallstudien vorstellen und die enge Verschränkung von Realität und Mathematik darin aufzeigen. Ich könnte mir die Erarbeitung der Techniken in der genannten Schulstufe durchaus im Zusammenhang mit Projektartigem Unterricht vorstellen, eine solche Unterweisung käme den angesprochenen Eigenheiten entgegen, jedoch ist

eine solche Bindung an diese Unterrichtsform nicht unbedingt erforderlich.

B) TECHNIKEN DER EDA

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

Beispiel: Beherbergungsbetriebe in Klagenfurt

Rohdaten

Tab.: 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 in diesem Fall nach dem Alphabet des Namens angeordnet. Aus dieser Ordnung jedoch kann man keine tiefschürfenden Einsichten gewinnen. Sie erleichtert lediglich das Aufsuchen eines bestimmten, namentlich bekannten Betriebes.

Ordnen der Daten

Einen ersten Überblick über die Bettenzahlen erhält man, wenn man die Daten (die Beherbergungsbetriebe) der Größe nach anordnet. Dieses Anordnen der Daten kann von einem Computer oder von Hand erledigt werden, als Ergebnis würde man die Daten in einer fortlaufenden Zeile angeordnet erhalten:

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 könnte man z.B. 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:

Fig.: 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

Dabei wird jede Zahl in einen sogenannten 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 (darauf wird später noch eingegangen werden). Diese Darstellung heißt Stamm-und-Blatt (St&Bl).

Das Bild zeigt den kleinsten und größten Wert, es zeigt ferner, bei welchen Bettenzahlen sich die meisten Betriebe "häufen" ((20-30),(30-40)), ferner, daß die meisten Betriebe nicht mehr als 60 Betten haben. Die Verteilung ist nicht symmetrisch. Im Gegenteil, der Umriß (Schatten) des Stamm-und-Blatt zeigt ferner, daß die Verteilung in zwei "Cluster" zerfällt, in einen Hauptcluster der "normal großen" und einen der "großen" Betriebe. Dies ist ein Anlaß, nachzudenken, was die "Ursachen" dafür sind: Welche weite-

ren 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. Eine Technik, die diesen Sachzusammenhang (welcher Art sind die großen Betriebe) besser studieren läßt, ist die Codierung der Daten. Entweder man schreibt in die betreffende Zeile des St&Bl einen Code statt der Zahl oder man beschriftet das St&Bl hinsichtlich des Clusters der größten Werte:

Fig.: 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 |30,80,80      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

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.

Sowohl der Umstand, daß in der St&Bl-Darstellung die einzelnen Daten noch identifizierbar sind, d.h., man kann den "Objektträger" ausfindig machen und beliebig anderes (auch informelles) Wissen über ihn in Erfahrung bringen oder einbringen, als auch die Technik der Codierung der Darstellung mit gut verständlichen Kurzbezeichnungen zeigt, daß man den Bezug der mathematischen Darstellung zur Realität aufrecht erhalten will, ja, daß man daraus sich gerade weiterreichende "Einsichten" erhofft.

2. Codiertes St&Bl - Zweiseitiges St&Bl

Beispiel: Tore im europäischen Fußball

Rohdaten

Tab.: 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

Die Rohdaten sind hier schon als Rangliste geführt. Die Rangliste läßt erkennen, daß die BR. Deutschland mit 3.35 Toren pro Spiel an der Spitze steht, Malta bildet mit 1.94 das Schlußlicht, Österreich nimmt in dieser Tabelle mit 2.92 Toren Rang 9 unter 30 Nationen ein. Die Verteilung der Trefferquoten wird jedoch wieder erst aus einem Histogramm oder einem St&Bl ersichtlich. Diese Darstellung soll gleich an die Untersuchung der folgenden Fragestellung gekoppelt werden: Stimmt die gängige Meinung, daß im Süden Europas weniger Tore geschossen werden als im Norden?

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. Das Ergebnis ist in der Tabelle auf der folgenden Seite enthalten.

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. Im Rahmen der Beschreibenden Statistik kann man die Differenz $\bar{x}_N - \bar{x}_S = 0.27$ dem sogenannten t-Test unterwerfen, der 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 (siehe [4]).

Tab.: 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

Zweiseitiges St&Bl

Eine explorative Alternative zur Beurteilung der Unterschiede zwischen Norden und Süden hinsichtlich der Trefferzahl pro Spiel ist ein zweiseitiges St&Bl:

Fig.: 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
	1.9	4

NORDEN SÜDEN

Das "Laub" für die Gruppe Süden wird dabei rechts, jenes für die Gruppe Norden links vom Stamm angeordnet, die Zahlen werden so

in Stamm und Blatt zerlegt: 2.92 in 2.9|2. Das Laub besteht demnach gerade aus der Hundertstel-Stelle der Daten.

Zunächst erhält man aus den einzelnen Seiten der Darstellung 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 erläutert werden. Ferner gibt das zweiseitige St&Bl durch die Gegenüberstellung einen direkten Vergleich der Trefferquoten zwischen Süden und Norden. Der ausgewiesene Unterschied ist vielleicht nicht dramatisch aber deutlich sichtbar.

Codierung des St&Bl

An dieser Stelle könnte man den angestrebten Vergleich zwischen Süden und Norden abbrechen. Die St&Bl-Technik erlaubt uns jedoch, weitere, zielführende Analysen: Es wurden schon Codes zur Identifikation der Daten im St&Bl eingeführt. Werden im zweiseitigen St&Bl wenigstens die extremeren Daten mit namentlichen Codes gekennzeichnet, so erhält man folgendes Bid:

Fig.: 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

Die Codierung erlaubt, den geographischen Bezug zu den Daten aufrechtzuerhalten. Man erkennt nun deutlich, daß die hohen Trefferquoten im Süden eigentlich von typisch 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. Im folgenden wird auf die Einsicht bezüglich der Bildung der Gruppen noch eingegangen werden.

3. Vierfeldertafeln

Beispiel: Tore im europäischen Fußball (Fortsetzung)

Vierfeldertafeln mit Häufigkeiten

Die Länder wurden bereits geographisch in zwei Hälften eingeteilt. Eine weitere Unterteilung in zwei Gruppen wäre die nach der Trefferquote direkt, und zwar in Länder mit "hoher Trefferquote" und solche mit "niedriger Trefferquote". Wo legt man die Grenze zwischen diesen Gruppen? Fordert man wieder, daß die entstehenden Gruppen gleich groß sind, so erhält man den Median $\tilde{x}$ als Trennpunkt. Es kommt eigentlich jeder Wert aus dem Intervall (2.73, 2.81) in Frage, die übliche Festsetzung ist

$$\tilde{x} = (2.73+2.81)/2 = 2.77$$

Man kann sich im zweiseitigen St&Bl davon überzeugen, daß je 14 Daten über bzw. unter diesem Wert 2.77 liegen. Klassifiziert man nun die Länder nach der Zugehörigkeit zu den jeweils zwei Gruppen Norden und Süden bzw. "viele Treffer" und "wenige Treffer", so 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 folgender Matrixform notiert:

Tab.: Trefferquoten - Numerische Vierfeldertafel

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

Der Unterschied zwischen Süden und Norden, der sich aus dem zweiseitigen St&Bl angedeutet hat, ist nun überdeutlich zu sehen: Die Verhältnisse sind in N und S geradezu gegengleich.

Vierfeldertafel mit Codes

Nachteil der Vierfeldertafel ist jedoch, daß der geographische Bezug nur mehr sehr grob vorhanden ist. Die aus dem codierten zweiseitigen St&Bl gewonnene Einsicht, daß das Einteilungskriterium für Norden und Süden eigentlich nicht passend ist, daß die

Ergebnisse für S durch typisch mitteleuropäische Länder und Frankreich verfälscht werden, ist nun durch den fehlenden Bezug zur Realität verschleiert. Dieser Bezug läßt sich aber leicht wieder herstellen, indem man statt der Häufigkeiten in die Felder der Matrix Codes (leichte Varianten der internationalen Kfz-Kennzeichen) für die entsprechenden Länder einzeichnet.

Tab.: Trefferquoten: Codierte Vierfeldertafel

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

Man sieht ganz deutlich: Hohe Trefferquoten im Süden werden ausschließlich von nicht-südeuropäischen Ländern "verursacht". Diese sachliche Einsicht könnte zum Vergleich Mittelmeerländer gegen den Rest Europas führen. Eine entsprechende Vierfeldertafel sieht dann so aus:

Tab.: 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

Rohdaten

Tab.: 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		

Die Rohdaten sind nach geographischer Lage von Norden nach Süden angeordnet. Aus der Reihenfolge der Daten erkennt man grob die Abfolge der Klimazonen von tropisch-trocken (Wüsten) über tropisch-feucht (tropische Regenwälder) bis tropisch-trocken (Steppen und Wüsten). Die Verteilung der Niederschläge soll im Hinblick auf die Klimazonen analysiert werden. Das Klima soll stellvertretend mittels Niederschlagsdaten beschrieben werden.

Histogramm

Die Verteilungsart wird üblicherweise durch ein Histogramm untersucht. Dabei ist die Klasseneinteilung so geschickt zu wählen, daß ein flächiger graphischer Eindruck von 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.: Niederschläge in 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 <u>keinen</u> Überblick über die Verteilung der Niederschlagsdaten. Die graphische Wirkung der Darstellung wird durch folgende Techniken verdichtet:

a) Kappen der Daten auf zwei geltende Stellen: 2736 wird zu 27. Will man nicht so viel an Information verlieren, so rundet man.

b) Zusammenfassen von 2, 5 oder 10 Zeilen.

Fig.: Niederschläge in Afrika - St&Bl mit zwei geltenden Stellen (auf 100mm gekappt). Je fünf alte Zeilen zusammengefaßt

T	H
0	00011112333333344444
0.	56667778999
1	0011111222233344
1.	555555777788899
2	0337
2.	
3	44

Diese Darstellung der Daten ist zu grob. Faßt man nur zwei Zeilen aus dem ursprünglichen St&Bl zusammen, so erhält man folgendes Bild.

Fig.: Niederschläge in Afrika - St&Bl mit zwei geltenden Stellen (auf 100mm 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
```

Ein verdichtetes St&Bl ist ein Histogramm mit speziellen Klassengrenzen, das um 90° gedreht wird (in der letzten Darstellung sind das die Klassen 0 bis unter 200, 200 bis unter 400 usf.) Im gewöhnlichen 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 jetzt auch die Identifikation erschwert, aber nicht prinzipiell unmöglich gemacht. Die extremen Daten sind in der Darstellung nach wie vor namentlich gekennzeichnet. Dieses St&Bl bzw. sein Umriß (Schatten) 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.

Beispiel: Vergleich der Niederschläge in Afrika, Südamerika und Australien

Rohdaten

Tab.: 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
Derby	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	2964	Hobart	636
Christchurch	658	Dunedin	791		

Tab.: 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				

Angestrebt ist ein Vergleich der Klimate in den drei Erdteilen der Südhalbkugel.

Vergleich von Histogrammen und Mittelwerten

In der Beschreibenden Statistik ist es üblich, Unterschiede in Verteilungen durch Histogramme oder durch Mittelwerte zu vergleichen. Der Nachweis, ob Unterschiede in den Mittelwerten bestehen, kann formal durch die sogenannte 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.: Vergleich der Niederschläge in Afrika, Südamerika und Australien mit verdichteten St&Bl-Darstellungen

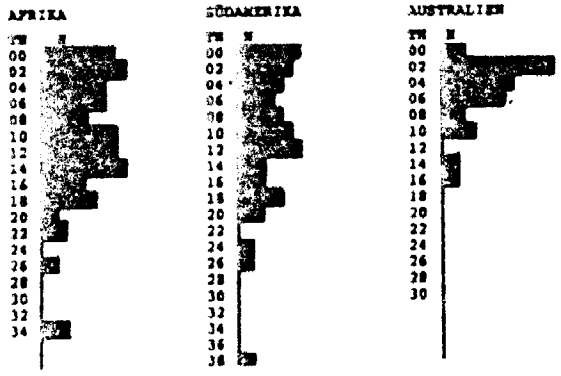
AFRIKA		SÜDAMERIKA		AUSTRALIEN	
TH	H	TH	H	TH	H
00	0001111	00	000111	00	11
02	23333333	02	22333	02	22222223333
04	444445	04	4455	04	4444555
06	666777	06	777	06	666666677
08	8999	08	9999	08	8999
10	0011111	10	00011	10	001
12	2222333	12	222333	12	22
14	44555555	14	55	14	4
16	7777	16	66	16	6
18	88899	18	8999	18	
20	0	20	01	20	
22	33	22		22	
24		24	4	24	
26	7	26	7	26	
28		28		28	8
30		30		30	
32		32			
34	44	34			
		36			
		38	9		

In der EDA vergleicht man St&Bl-Darstellungen anstelle von Histogrammen. Die Unterschiede in den Klimazonen bezüglich der Niederschlagsdaten können demnach in zwei Hauptrichtungen beurteilt werden:

a) 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.

b) Wie sind die Ausreißer zu interpretieren? Beschriften der extremen Werte ist, wie gehabt, 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.: Vergleich der Niederschlagsdaten von Afrika, Südamerika und Australien: Umrisse der St&Bl - Australien ohne Neuseeland



5. Zahlenzusammenfassungen

Zur Beschreibung von Daten sind manchesmal Zahlen besser geeignet 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 ausgewiesen. Es wurde schon darauf hin-

gewiesen, daß es durchaus schwierig sein kann, den daraus gewonnenen Zahlenwerten eine Interpretation abzugewinnen. Eine interessante 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 recht 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 $\bar{x}$ - 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.

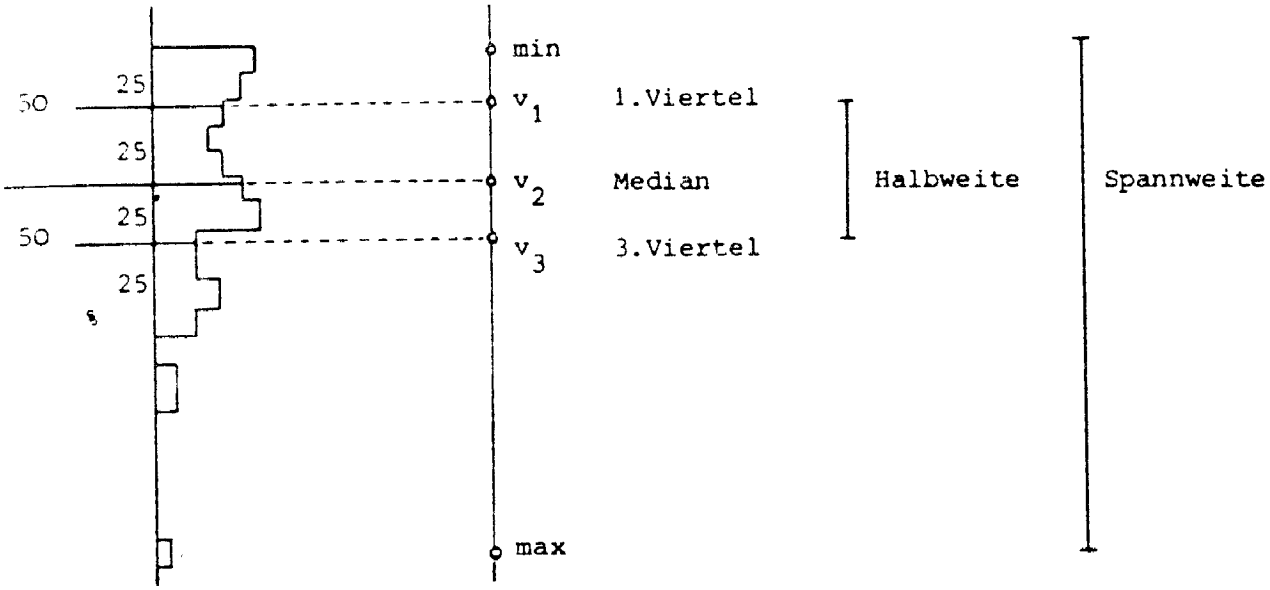
Bestimmen der Trennpunkte

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

Halbiert man den Haufen von 46 Daten (siehe die Abb. auf der folgenden Seite), 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 24.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.: 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.

Fünfzahlenzusammenfassung

Folgende Kennziffern, die beim fortlaufenden Haufenhalbieren anfallen, und die Lage und Breite einer Verteilung kennzeichnen sollen, gibt man übersichtlich in einer Fünfzahlenzusammenfassung an:

A		
	$\tilde{x}$	
A/2	v_1 v_3	v_3-v_1
A/4	min max	max-min

Rechts vom Kasten schreibt man meist die Kennziffern für die Breite, die Halbweite und die Spannweite, an. Diese Zahlenzusammenfassungen geben ganz gezielte Vergleichsmöglichkeiten zwischen den Verteilungen an.

Fig.: Fünfzahlenzusammenfassungen der Niederschläge (in cm)

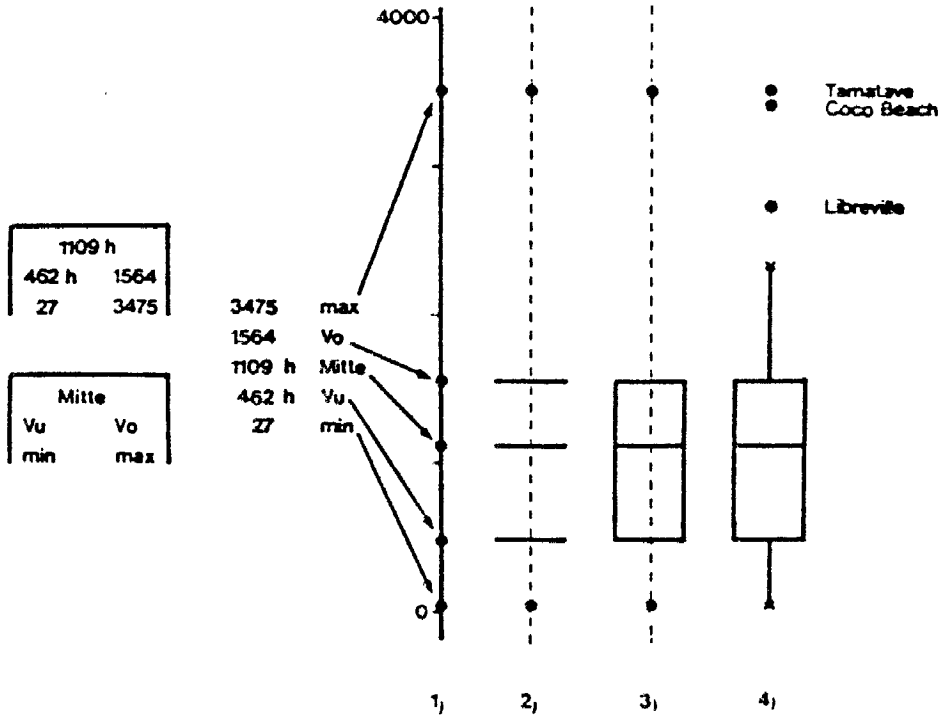
68	Afrika			46	Südamerika			41	Australien		
34	110			23	103			20h	60		
17	46	156	110	11h	43	158	115	10	31h	93	61h
1	2	347	345	1	00	392	392	1	11	286	275

Diese Zahlenkasten werden hier nicht näher interpretiert, weil sie die Basis der im folgenden dargestellten Kastenschaubilder bilden.

6. Kastenschaubilder

Die Kennziffern aus der Fünfzahlenzusammenfassung bieten viel Information für einen gezielten Vergleich mehrerer Verteilungen. Die zahlenmäßige Information ist jedoch leichter zugänglich, wenn sie graphisch aufbereitet wird.

Fig.: Schrittweise Konstruktion eines Kastenschaubildes aus der Zahlenzusammenfassung - Niederschlagsdaten für Afrika

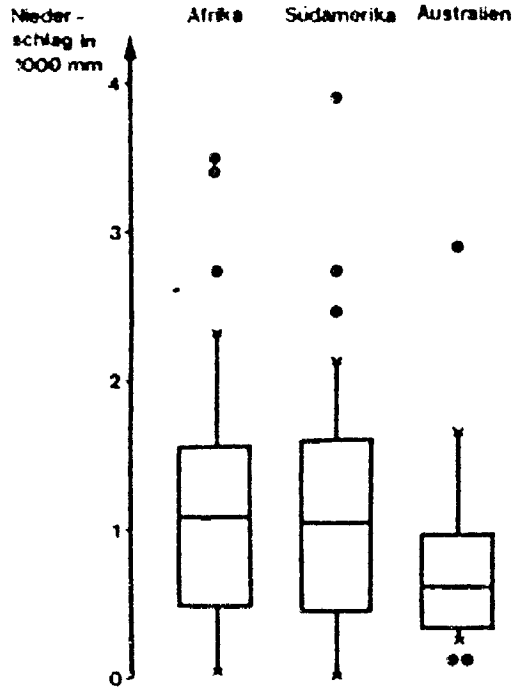


Der zentrale Kasten in der vorhergehenden Abbildung umfaßt die mittlere Hälfte der Daten. Für das Einzeichnen der Ausläufer gibt es unterschiedliche Regeln, z.B.:

- Ausläufer für jenen Bereich, in dem die Verteilung einen geschlossenen Eindruck macht.
- 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.

Beispiel: Niederschläge in Afrika, Südamerika und Australien-Kastenschaubilder zum Vergleich



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

C) EIGENHEITEN UND HINTERGRUND VON EDA-TECHNIKEN

Die Beispiele in Teil B wurden aus [2] entnommen. Für weitere Techniken, insbesondere für die Analyse zweidimensionaler Daten sowie für weitere Details sei ebenfalls auf [2] verwiesen. Dort ist auch der Hintergrund der Verfahren ausführlich dargestellt, weshalb hier lediglich überblicksartig einige Eigenheiten dargestellt werden sollen.

1. Vergleich der Techniken: EDA - Beschreibende Statistik

	Stamm&Blatt	Histogramm
o	Datenbearbeiter erstellt und interpretiert selbst.	Wird von anderen erstellt.
o	Entsteht beim einfachen Ordnen der Daten.	Eigene Prozedur
o	Ist Histogramm mit pragmatischen Klassengrenzen.	Feinheiten der Wahl der Klassengrenzen.
o	Darstellung absoluter Häufigkeiten.	Relative Häufigkeiten.
o	Beim Erstellen: Trend in den Daten?	
o	Vollständige Identifizierbarkeit einzelner Daten.	Kein Bezug der Daten zu Objekten.
o	Codes zur visuellen Unterstützung zum Auffinden von Besonderheiten.	
o	Muster der Verteilung + Besonderheiten (zerfallen Daten in Gruppen, Ausreißer?)	Generelles Muster.
o	Daten vieldimensional im Bearbeiter, keine vollständige Trennung vom Sachbezug.	Daten eindimensional.

	Kastenschaubild	Histogramm
o	Verteilung durch wenige markante Punkte.	Verteilung in vielen Einzelheiten.
o	Ziel: direktes Hinführen auf Muster + Besonderheiten.	Verteilungsmodell als generelles Muster ($\nearrow$, $\searrow$) oder Reduktion auf $\bar{x}$, s.
o	Gewöhnlicher Streubereich extra markiert.	
o	Besondere Werte eigens codiert.	Besondere Werte fallen durch Klasseneinteilung durch den Rost.
o	Visuelle Beziehungen zwischen den Kennziffern.	
o	Vergleich von Verteilungen: sehr effizient durch Reduktion auf wenige Details, keinerlei Voraussetzungen.	Vergleich erst auf Basis von Tests, mächtiges Instrument, das nur unter spezifischen Voraussetzungen erlaubt ist.

	Median	Mittelwert
o	Faßt Muster zu einer Zahl, ist Mitte der zentralen Box.	Steht für die gesamte Verteilung, berücksichtigt jedes Einzeldatum.
o	Ist robust.	Ist empfindlich gegen Ausreißer.
o	Hat triviale Deutung.	Ist oft schwer zu deuten.

2. Zum Stil der EDA

Folgende kursorische Aufzählung sollte sich aus den behandelten Beispielen in Teil B ergeben, für nähere Details verweise ich wiederum auf [2]. Verfahren der Explorativen Datenanalyse zeichnen sich aus durch:

- o Besonderheiten und Muster
  Man sucht nicht nur nach einem speziellen Muster, das die Verteilung der Daten ausreichend beschreiben läßt, z.B., die Daten sind normalverteilt. Vielmehr erwartet man sich gerade aus der Interpretation von Besonderheiten (Ausreißer, Daten zerfallen in mehrere Untergruppen etc.) sachliche Aufschlüsse.
- o Flexibilität der Verfahren
  Die Verfahren können nach Bedarf abgewandelt werden. Ziel ist es, neue Einsichten zu gewinnen, die Frage nach dem "richtigen" Einsatz wird dagegen nachrangig.
- o Robustheit
  Solche Verfahren, die auf Einzelwerte empfindlich reagieren, erfordern es, daß ausreißerverdächtige Werte bereits vor der Analyse ausgeschieden werden. Gerade von einer sachgerechten Interpretation solcher Werte erwartet man sich Aufschluß. Es werden daher solche Verfahren bevorzugt, die robust gegen Ausreißer sind (Median z.B.).
- o Geringer Aufwand
  Die Verfahren sollen leicht verständlich und leicht durchzuführen sein: Abzählen statt Rechnen, Vergleich von Kastenschaubildern statt Varianzanalyse, Trendgeraden per Augenmaß statt Regressionsgerade nach der Methode der kleinsten Qua-

drate usf.

- o Praktische Ausrichtung
  Die Verfahren leben eigentlich davon, daß ihre Zwischenergebnisse am realen Bezugsfeld interpretiert werden, damit man dann direkt entscheiden kann, wie die Untersuchung weiter gehen kann. Die Verfahren sind daher nicht auf ihren rein mathematischen Gehalt, der gering ist, zu reduzieren.
- o Visuelles Arbeiten
  Explorative Techniken haben einen stark visuellen Charakter. Dies ist gerade für den direkten Arbeitsfortgang von entscheidender Bedeutung. In irgendeiner visuellen Projektion der Daten sollen ihre inneliegenden "Strukturen" aufgedeckt und damit sachliche Einsichten geweckt werden.
- o Einfache Konzepte
  Damit die weitere Datenbearbeitung voranschreiten kann, ist es nötig, die Zwischenergebnisse zu verstehen. Dies wird entscheidend dadurch gestützt, daß die verwendeten Konzepte trivial, d.h. ohne Bezug auf eine weitere Theorie zu verstehen sind.

3. Eigenständigkeit und Kreativität

Abschließend seien noch einige Aspekte angedeutet, die im Rahmen einer wie hier verstandenen Explorativen Datenanalyse für den Unterricht eine wichtige Rolle in diesem Gebiet der Statistik einnehmen. Diese können als Chance aber auch als Schwierigkeit begriffen werden:

- o Im explorativen Gebrauch visueller Darstellungen liegt eine gewisse Offenheit in den verwendeten Techniken, diese sind nicht festgelegt, ja sie können fallweise überhaupt abgeändert oder neu erfunden werden, falls das Problem es erfordert.
- o Die subjektive Komponente von Wissen und der Umgang damit ist auf zwei Ebenen zu beachten: Erkenntnisse können nicht immer eindeutig "abgeleitet" werden, ja die gewonnenen Einsichten könnten durchaus widersprüchlich sein. Es gibt keine optimale Technik, die in einer ganz bestimmten Situation angewendet werden müßte.

- Der Sachverstand lenkt die Interpretation der Zwischenergebnisse und damit die weitere Bearbeitung. Man muß über das Umfeld, aus dem das Problem stammt, Bescheid wissen. Man kann nicht zielführend Methoden formal probieren, ohne sie wirklich verstanden zu haben.
- Die Bearbeitungsmethode ist interaktiv zwischen dem Bearbeiter und dessen Wissen über das Umfeld und den mathematischen Zwischenergebnissen. Das erfordert eine experimentelle Arbeitshaltung, oft verbunden mit Risikobereitschaft, Wege zu gehen, die wahrscheinlich zu keiner befriedigenden Einsicht führen.
- Ein differenziertes Verständnis der Ergebnisse beeinflußt sowohl den Fortgang der Bearbeitung als auch das Anerkennen bestimmter Interpretationen als ein brauchbares Ergebnis. Man muß mit den Ergebnissen seiner Arbeit viel direkter und kritischer umgehen. Man kann dies nicht einem anderen, etwa dem Lehrer, überlassen.

Literatur:

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

[2] Manfred Borovcnik u. Günther Ossimitz: Materialien zur Beschreibenden Statistik und Explorativen Datenanalyse.- Wien/Stuttgart: Hölder-Pichler-Tempsky/Teubner 1987.

[3] Roland Fischer u. Günther Malle: Mensch und Mathematik - Eine Einführung in didaktisches Denken und Handeln. Mannheim: Bibliographisches Institut 1985.

[4] Johann Pfanzagl: Allgemeine Methodenlehre der Statistik. Bd.2. Berlin - New York: de Gruyter 1974.

[5] Wolfgang Polasek: Explorative Datenanalyse. Einführung in die deskriptive Statistik. Berlin-Heidelberg-New York: Springer 1988.

[6] John W. Tukey: Exploratory Data Analysis. Reading: Addison-Wesley 1977.