Less frequently asked questions: Nutzen und Notwendigkeit grafisch gestützter Datenanalyse

Schnell, Rainer

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Less frequently asked questions: Nutzen und Notwendigkeit grafisch gestützter Datenanalyse

Rainer Schnell

„Making sense is more important than making numbers. Making sense is even more important than making significance." (Jöreskog 2000)

1. Was ist Datenanalysegrafik?

Die grafische Darstellung von Daten oder statistischer Größen im Rahmen von Datenanalysen wird als Datenanalysegrafik bezeichnet.[1] Datenanalysegrafik wird häufig mit Präsentationsgrafik verwechselt. Präsentationsgrafik stellt die Ergebnisse einer Datenanalyse dar, meist für ein fachunkundiges Publikum. Entsprechend spielen hier kognitionspsychologische Kriterien eine dominante Rolle.[2] Datenanalysegrafik hingegen wird bei der Durchführung einer Datenanalyse eingesetzt. Die grafischen Darstellungen von Daten („Plots") dienen hier vor allem der Beurteilung der Angemessenheit der verwendeten Datenanalysemodelle.[3]

2. Warum sollte man Datenanalysegrafik anwenden?

„You can see a lot, just by looking." (Yogi Berra)

Da Plots meist weniger von statistischen Annahmen abhängig sind als zusammenfassende Statistiken (wie z. B. Mittelwerte), eignen sie sich besser als andere Techniken zur Beurteilung der Angemessenheit eines Modells. Durch die Darstellung der Beobachtungen anstelle zusammenfassender Statistiken werden häufig bemerkenswerte Strukturen und Muster in den Daten erkennbar, die sonst unbemerkt geblieben wären.[4] Plots erlauben so die Entdeckung unvermuteter Regelmäßigkeiten oder ungewöhnlicher Beobachtungen: „Graphs force us to note the unexpected" (Tukey 1977, S. 157). Datenanalysegrafik ist somit vor allem bei der Entwicklung statistischer Modelle unverzichtbar.

3. Was bedeutet „explorative Datenanalyse" (EDA)?

Häufig wird der Teil einer Datenanalyse, der sich mit der Entwicklung eines statistischen Modells befasst, als „explorative Datenanalyse" bezeichnet. Im engeren Sinne wird der Begriff „explorative Datenanalyse" (EDA) in der Regel für die von John W. Tukey entwickelten Techniken und die von ihm empfohlene Vorgehensweise bei einer Datenanalyse gebraucht. Hoaglin (1985, S. 579) nennt vier Schwerpunktthemen „explorativer" Datenanalyse:

- Robustheit
- Residuen
- Datentransformation
- grafische Darstellungen

In vielen Lehrbuchdarstellungen wird EDA auf den letzten Punkt beschränkt. Zwar sind grafische Techniken für EDA unverzichtbar, aber man kann EDA keinesfalls auf bestimmte Abbildungsformen reduzieren. Tukey (1980, S. 23) betont, dass explorative Datenanalyse keine Sammlung von Techniken, sondern eine Geisteshaltung sei. Das Grundprinzip hierbei besteht in der Flexibilität gegenüber den Daten: Ausgehend von vorläufigen Modellen werden mit verschiedenen Techniken Abweichungen von diesen Modellen geprüft und die Modelle entsprechend modifiziert: Datenanalyse besteht aus einem „model-data cycle" (Mallows/Tukey 1982, S. 113). Weder wird ein Modell von vornherein als „wahr" betrachtet, noch werden die Daten kritiklos als „gegeben" akzeptiert. Gerade diese Konzeption eines modellvermittelten Dialogs mit den Daten erklärt die Aufgeschlossenheit von Fachwissenschaftlern gegenüber diesem Vorgehen. Nicht Parameterschätzungen oder gar Signifikanztests stehen im Mittelpunkt, sondern letztlich inhaltliche Modelle und die Abweichungen der eigenen Daten von diesen Modellen.

4. Ist EDA deskriptive Statistik?

Nein. Dadurch, dass Tukey (1977) in seinem Buch „Exploratory Data Analysis" den Schwerpunkt auf einfache Techniken für Probleme mit wenigen Variablen gelegt hat, scheint sich bei vielen Statistikern die Meinung gebildet zu haben, EDA sei eine andere Bezeichnung für „deskriptive Statistik". Dies verkennt Tukeys Anliegen vollständig.

Die fehlerhafte Rezeption der Grundideen explorativer Datenanalyse ist sicherlich zum Teil auf die ungünstige Lehrbuchsituation zurückzuführen. Schon allein aufgrund seiner abenteuerlichen Terminologie dürfte Tukeys eigenes Buch (1977) wenig wirkliche Leser gefunden haben. Die spätere – vor allem

deutsche – Lehrbuchliteratur machte durch Mathematisierung aus den aufregenden „adventures of those neo-inductivists in the EDA Revolutionary Front" (Lovie 1987, S. 376) wieder die in der Statistik üblichen langweiligen Formelsammlungen. Nur sehr wenige Lehrbücher (z. B. Erickson/Nosanchuck 1977) integrieren EDA als Grundhaltung gegenüber den Daten.

5. Gibt es einen Unterschied zwischen „explorativer" und „konfirmatorischer" Datenanalyse?

Die Fehlrezeption der Grundideen der EDA hat zu einer fortwährenden Debatte über den Unterschied zwischen „explorativer" und „konfirmatorischer" Datenanalyse geführt. Ein Beispiel geben Heiler/Michels (1994, S. VII): Typisch für konfirmatorische Datenanalysen sei das Schema der testenden Statistik. Danach stehe am Anfang eine Hypothese, zu deren Überprüfung über einen geeigneten Versuchsplan eine Zufallsstichprobe durchgeführt und darauf dann ein Test angewandt werde. Explorative Verfahren gäben hingegen durch die Suche nach Auffälligkeiten bzw. nichttrivialen Strukturen Anstöße zur Bildung von Hypothesen und Modellen. Solche idealtypischen Konzeptionen sind nur möglich, wenn man den tatsächlichen Forschungsprozess aus dem Blick verloren hat.

Datenanalysen sind – wie der gesamte Forschungsprozess – immer iterativ. Daten werden unter theoretischen Perspektiven und Annahmen erhoben. Für die so gewonnenen Daten werden ausgehend von einem vorläufigen Modell Konsequenzen abgeleitet, die in einer vorläufigen Analyse überprüft werden.[5] Daran schließt sich eine empirische Kritik des vorläufigen Modells an. Diese empirische Kritik besteht häufig aus einer Form der Residuenanalyse (genau hier besitzen Datenanalysegrafiken auch ihr hauptsächliches Einsatzgebiet). Die Kritik des vorläufigen Modells führt zu einer Modifikation des Erklärungsmodells. Der Datenanalysezyklus beginnt dann erneut. Datenanalyse ist somit eine Subiteration innerhalb des gesamten Forschungsprozesses.

„Explorative" und „konfirmatorische" Phasen der Datenanalyse können daher nicht sauber getrennt werden. Die weit verbreitete Debatte um „explorative Datenanalyse" versus „konfirmatorische Datenanalyse" ist ein unnötiges Scheingefecht, das meist in Unkenntnis sowohl der Fachliteratur als auch tatsächlicher Forschungsarbeit geführt wird.

6. Ist explorative Datenanalyse ohne inhaltliche Theorie möglich?

Häufig werden statistische Grafiken als Werkzeug einer angeblich modellfreien Datenanalyse präsentiert. Insbesondere seit die Korrespondenzanalyse in den Sozialwissenschaften populär wurde, zeigen sich in vielen empirischen Arbeiten Formen „induktiven" Vorgehens, die zumindest unter Wissenschaftstheoretikern als längst überwunden angesehen wurden. Manche Verfechter scheinen jedoch davon überzeugt zu sein, mithilfe grafischer Techniken induktiv zu neuen Einsichten, wenn nicht gar zu „Gesetzen" kommen zu können.[6] Als Illustration wird von solchen Autoren häufig Tukeys Definition „explorativer Datenanalyse" (1977, S. V) zitiert: „looking at data to see what is seems to say". Der Begriff „explorative Datenanalyse" dient so zur Rechtfertigung eines angeblich theoriefreien Herangehens an die „vorhandenen" Daten. Tukey (1990, S. 332) selbst hat dies als die „tabula-rasa fallacy for display" bezeichnet: Es ist eine Illusion zu glauben, dass grafische Methoden eine automatische und verzerrungsfreie Art der Datenanalyse erlauben würden.

Die Kritik an „induktiven" Datenanalysekonzeptionen besitzt drei Ansatzpunkte:

- das Problem „theoriefreier" Messungen („gegebene" Daten);
- das Problem der Rechtfertigung induktiver Schlüsse;
- die Möglichkeit der induktiven Entdeckung von Regelmäßigkeiten.

„Messungen" setzen stets inhaltliche Theorien voraus. Die Konstruktion neuer Theorien ist fast immer mit einer Rekonzeptualisierung eines Gebietes verbunden: „Messungen", „Daten" und „Variablen" existieren nicht unabhängig von Theorien.[7] Theoriefreie Beobachtungen sind unmöglich. Ebenso ist es unstreitig logisch unmöglich, induktive Schlüsse zu rechtfertigen. Dies wird aber auch von den Verfechtern „induktiven" Vorgehens nicht behauptet. Diese nehmen immer nur Bezug auf die heuristischen Möglichkeiten der jeweiligen Technik; und dies betrifft das dritte Problem. Die Frage, ob sich grafische Techniken zur Hypothesengenerierung eignen, lässt sich aufgrund der unklaren Fragestellung nur bejahend beantworten. Obwohl es nicht mit Sicherheit ausgeschlossen werden kann, ist die Entdeckung bedeutsamer Regelmäßigkeiten oder gar neuer Konzepte in „vorgefundenen" Daten ohne langdauernde vorherige theoretische Beschäftigung mit dem jeweiligen Forschungsgebiet extrem unwahrscheinlich – und bisher scheint sich kein einziger derartiger Fall in der Wissenschaftsgeschichte nachweisen zu lassen. Und selbst wenn es solche Fälle geben sollte, was würde das beweisen?

Die Entwicklung neuer theoretischer Konzepte scheint meistens auf die Kombination und Rekonzeptualisierung bestehender Konzepte zurückzuführen zu sein.[8] In der Wissenschaftsgeschichte lässt sich auch in den Beispielen „über-

raschender Entdeckungen" in jedem Fall zeigen, dass ein großes Ausmaß theoretischen Wissens über den Gegenstandsbereich beim Entdecker vorlag.[9] Die Wahrnehmungspsychologie zeigt, dass schon zur Interpretation visueller Stimuli grafische Schemata unentbehrlich sind (Pinker 1990, Banks/Krajicek 1991). Grafische Schemata müssen erlernt werden, d. h., die Interpretation der Plots muss gelernt werden. Da aus der Vielzahl der visuellen Stimuli stets anhand der vorhandenen Schemata ausgewählt wird (Cutting 1991, S. 45), muss Tukeys berühmter Satz „Graphs force us to note the unexpected" (Tukey 1977, S. 157) zurückhaltend interpretiert werden. Man muss vor der Betrachtung eines Plots Hypothesen über sein Aussehen besitzen, um überrascht werden zu können (vgl. Hadi 1993, S. 777). Das Interesse des Analytikers kann sich immer nur auf spezielle Aspekte beziehen.

Tukey scheint in seinen späteren Arbeiten die Missverständnisse der Prinzipien „explorativer Datenanalyse" durch die Einführung neuer Begriffe ausräumen zu wollen. Tukey (1990) unterscheidet zwischen „explorativer" und „prospektiver" Datenanalyse. Ein Prospektor weiß, wonach er sucht. Ein erfahrener Datenanalytiker verfügt über eine Reihe möglicher Modelle zur Erklärung der Daten und eine Liste von Phänomenen, nach denen er sucht: Heteroskedastizität, Nichtlinearitäten, räumliche Inhomogenitäten (Klumpungen, Löcher) usw. Ein Datensatz ist also prinzipiell nur in Hinsicht auf die Abweichungen der Daten von einem Modell von Interesse. Obwohl für grafische Residuenanalysen keine expliziten theoretischen Modelle benötigt werden (Cox/Snell 1968, S. 249), sind theoretische Erwartungen jedoch unentbehrlich.[10] Man muss wissen, was an einem Plot relevant ist, um ihn interpretieren zu können. Folglich benötigt man *vor* der Betrachtung eines Plot explizite Erwartungen darüber, wie der Plot aussehen soll. Cox/Gabriel (1982, S. 80) bezeichnen dies als „inspired inspection of irregularities". Entsprechend warnen Cook/Weisberg (1999, S. 36) vor der Verwendung von grafischen Techniken ohne statistische Theorie: „Without an underlying statistical theory, graphs can be difficult to interpret, can give misleading information, and can be more a hindrance to reasonable statistical analysis than a help."

7. Was hat grafisch gestützte Datenanalyse mit Data-Mining zu tun?

Es gibt keine notwendige Verbindung zwischen „grafisch gestützter Datenanalyse" einerseits und „Data-Mining" andererseits. Allerdings werden Verfahren grafisch gestützter Datenanalyse häufig bei „Data-Mining"-Anwendungen eingesetzt. Weiterhin gibt es natürlich Autoren, die den Einsatz von Data-Mining-Programmen für die Sozialwissenschaften propagieren..

Die Idee „automatischer" Datenanalysen scheint eine faszinierende Wirkung auf viele technisch orientierte Autoren auszuüben. In der Tat besäßen solche Verfahren einen immensen Vorteil: Rechnen ist einfacher und billiger als Nachdenken. Sobald man sich aber ernsthaft mit „Data-Mining"-Anwendungen beschäftigt, wird offensichtlich, dass dies ohne „human input" vollkommen unmöglich ist: Allein die Aufbereitung der Daten (Erstellen analysefähiger Datenstrukturen, Berechnung von Indizes, Bereinigung von Outliern und Imputation fehlender Werte) setzt neben Flexibilität und Alltagswissen halbwegs plausible Hypothesen über die Entstehung der Daten einschließlich ihrer Probleme voraus.[11] Kommerziell arbeitende Data-Mining-Unternehmen rekrutieren daher gern empirisch arbeitende Sozialwissenschaftler. Umso erstaunlicher muten daher Sozialwissenschaftler an, die durch automatische „Data-Mining"-Verfahren die sozialwissenschaftliche Theoriebildung voranzubringen gedenken.

In der Literatur wird in Hinsicht auf die Techniken häufig kaum ein Unterschied zwischen „Data-Mining" und „Statistik" gesehen. Die Besonderheiten des Data-Minings liegen dann eher in der Verwendung massiver Datensätze und im pragmatischen Verzicht auf wissenschaftliche Ansprüche der resultierenden Modelle. Ist man aber an der Entwicklung tatsächlich prädiktiver Modelle interessiert, dann gibt es keinen Unterschied in der Vorgehensweise zwischen einem „statistischen" Modell und einer „Data-Mining"-Anwendung. Entsprechend gilt für Data-Mining-Anwendungen dasselbe wie für die oben diskutierten Versuche einer theoriefreien Datenanalyse: Zwar gibt es kaum eine sinnvolle Modellentwicklung ohne die Unterstützung grafischer Techniken, aber ohne inhaltliches Modell ist auch keine sinnvolle Anwendung grafischer Analyseverfahren möglich.

8. Wie läuft eine grafisch gestützte Datenanalyse prinzipiell ab?

In der Entwicklung eines statistischen Erklärungsmodells können vier Phasen unterschieden werden (Mallows/Walley 1980, S. 11):

1. Identifikation von Regelmäßigkeiten oder Mustern in den Daten,
2. Auswahl der Form eines Modells zur Beschreibung dieser Regelmäßigkeiten,
3. Anpassung des Modells,
4. Beurteilung der Anpassung des Modells; Berechnung der Abweichungen des Modells von den Daten; Iteration zu 1.

Statistische Lehrbücher betonen vor allem die Phase der Anpassung des Modells, also die Berechnung der Parameter eines gegebenen Modells. Der inhalt-

lich interessante Teil der Arbeit findet sich eher in den anderen Prozessphasen, und bei genau diesen Schritten sind grafische Techniken meist hilfreich. Erfahrene Datenanalytiker beginnen ihre Datenanalysen mit Plots.[12]

9. Wie geht man bei einer grafisch gestützten Datenanalyse vor?

In der Regel wird zunächst in univariaten Plots nach groben Datenfehlern und Ausreißern gesucht. Hierzu eignen sich z. B. Boxplots bzw. Stem-Leaf-Plots. Die Datenfehler werden bereinigt und mögliche Ursachen für die Ausreißer untersucht (dies setzt in der Regel detaillierte inhaltliche Kenntnisse über den datengenerierenden Prozess voraus).

Anschließend sollten die univariaten Verteilungen mit Kern-Dichteschätzungen der Variablen untersucht werden. Dabei ergeben sich gelegentlich Hinweise darauf, dass die untersuchte Population aus heterogenen Subgruppen besteht. In den Plots der Dichteschätzer werden aber auf jeden Fall Abweichungen von der Symmetrie der Verteilung sichtbar. Da unter anderem viele Signifikanztests Normalverteilung voraussetzen, werden extrem schiefe Verteilungen geeigneten Datentransformationen (z. B. logarithmieren der Variablen) unterworfen.

Sollten die Beobachtungen im Datensatz eine natürliche zeitliche Ordnung aufweisen (z. B. Daten über Geburten nach der Uhrzeit der Geburt), dann empfiehlt sich eine Untersuchung des Zusammenhangs aller Variablen mit der Zeit. Dies kann z. B. durch einen Plot der Variablen gegen die Zeit oder die Abfolge im zeitlich sortierten Datensatz geschehen. Ähnliches gilt für eine räumliche Klumpung der Beobachtungen: Z. B. werden in sozialwissenschaftlichen Befragungen meist mehrere Personen durch den gleichen Interviewer befragt oder mehrere Interviews oder Messungen finden innerhalb der gleichen organisatorischen Einheit (Schulklassen, Krankenhäuser) statt. Mögliche Effekte der räumlichen Klumpung lassen sich z. B. durch Boxplots aller abhängigen Variablen gegen die durch die Interviewernummer gebildeten Gruppen prüfen. Sollten sich entweder bei der Untersuchung der zeitlichen Abfolge oder der räumlichen Klumpung Effekte zeigen, dann ist die für fast alle statistischen Verfahren notwendige Unabhängigkeitsannahme der Beobachtungen verletzt. In diesem Fall müssen dann Untersuchungen der Ursache für die Abhängigkeit der Beobachtungen durchgeführt und Analyseverfahren verwendet werden, die die Berücksichtigung dieser Abhängigkeiten erlauben.[13]

Anschließend werden die paarweisen Scatterplots der Variablen eines Datensatzes betrachtet. Es empfiehlt sich, in diese Scatterplotmatrizen die Kurven nichtparametrischer Regressionen (so genannte Scatterplot-Smoother, z. B.

„lowess" bzw. „loess") einzuzeichnen. Diese Plots erlauben die Identifikation bivariater Ausreißer und das Erkennen nicht-linearer Zusammenhänge.

Diese ersten Schritte einer Datenanalyse werden häufig durch einen Plot der Daten im Raum der ersten zwei oder drei Hauptkomponenten abgeschlossen. Hier zeigen sich häufig multivariate Ausreißer und Cluster ähnlicher Beobachtungen.

Erst hieran schließt sich die eigentliche Konstruktion und Beurteilung der vorläufigen Datenanalysemodelle, z. B. eines Regressionsmodells, an.

10. Welche Rolle spielen die Abweichungen von einem Modell in der Datenanalyse?

Die Abweichungen der Daten von einem vorläufigen Datenanalysemodell sind bei der Entwicklung eines statischen Modells von zentraler Bedeutung. Die Abweichungen eines Modells von den Daten werden als „Residuen" bezeichnet. Tukey (1977, S. 208) hat das Grundprinzip mit seinen beiden „Gleichungen"

$$\text{data} = \text{fit} + \text{residuals}$$

bzw.

$$\text{data} = \text{smooth} + \text{rough}$$

zusammengefasst.

Wenn der „fit" die bedeutsamsten Aspekte der Daten erfasst, sollten die Residuen keine Struktur mehr erkennen lassen, sie sollten „reasonably irregular" (Tukey 1977, S. 549) aussehen. Um dies zu prüfen, ist es häufig sinnvoll zu versuchen, Gemeinsamkeiten der Beobachtungen mit gleich großen Residuen zu finden. Dazu werden die Residuen gegen eine große Zahl anderer Variablen geplottet.

Residuenanalyse ist in der Datenanalysepraxis wenig verbreitet, obwohl die Analyse der Residuen eines Modells eines der wichtigsten Forschungswerkzeuge sein kann. Tukey/Wilk (1970, S. 387) betrachten die Koeffizienten einer linearen Regression in den seltensten Fällen als von eigenständigem Interesse. Im Allgemeinen sei eine multiple lineare Regression lediglich nützlich als Generator für Residuen und als Lieferant für eine empirische Beschreibung der Daten.[14] Die Analyse der Residuen kann zur Entdeckung bisher im Modell unberücksichtigter Variablen oder anderer Formen des Zusammenhangs führen. Entsprechend neu spezifizierte Modelle ergeben neue Residuen. Man kann von einer „diagnosegeleiteten Fit-Revision" sprechen (Mallows/Tukey 1982, S. 126).

Die Grundprinzipien einer grafisch gestützten Datenanalyse lassen sich somit kurz zusammenfassen (Friendly 1991, S. 35–41):
1. Berechnung der Residuen unter einem Modell,
2. Hervorhebung systematischer Tendenzen,
3. schrittweise Verbesserung des Modells.

Grafisch gestützte Datenanalyse ist daher immer interaktiv und iterativ. Da kein Plot alle möglichen interessanten Aspekte der Daten zeigen kann, empfiehlt sich meistens eine Betrachtung der Daten aus vielen verschiedenen Perspektiven. Unterschiedliche Typen von Plots heben immer andere Strukturaspekte in den Daten hervor. Häufig gibt ein Plot Anlass dazu, einen anderen Plot zu erstellen, der dann andere Eigenheiten der Daten betont. Einem Plot folgen in der Regel weitere Analysen oder Datentransformationen, denen neue Plots folgen usw. Lubinsky/Pregibon (1988, S. 247) haben für diese Art der Datenanalyse den Begriff „Display/Action cycle" geprägt.[15] Letztlich ist Residuenanalyse daher der wichtigste Bestandteil grafisch gestützter Datenanalysen.

11. Welche grafischen Datenanalysemöglichkeiten gibt es?

In der Fachliteratur findet sich eine große und stetig wachsende Zahl spezialisierter Plots.[16] Viele der spezialisierten Plots besitzen einen gemeinsamen Nachteil: Die Interpretation des Plots muss erst mühsam an vielen Beispielen erlernt werden, die Interpretation ist aber auch nach einigem Training nicht eindeutig.[17] Damit verletzen diese Plots den wichtigsten Grundsatz für Plots, den Tukey (1990) in einer Überschrift als „Impact, not archaelogy" formulierte. Die meisten „erfolgreichen" Plots sind simpel und genügen in vielen Fällen einer von zwei einfachen Interpretationsregeln:
- Im Plot nahe beieinander liegende Punkte sind sich ähnlicher als entfernt liegende Punkte.
- Wenn die Daten einem zugrunde liegenden Modell entsprechen, dann liegen die Plotpunkte auf einer Geraden.

Entsprechend konzentriert sich die Lehrbuchliteratur auf vergleichsweise wenige Grundtechniken, die in immer neuen Kombinationen sinnvolle Plots ermöglichen. Ein modernes grafisch gestütztes Datenanalysesystem sollte mindestens über folgende Möglichkeiten verfügen[18]:

Univariate Verteilungen:
- Boxplots
- nichtparametrische Dichteschätzungen
- eindimensionale Scatterplots
- Dotplots

- Symmetrieplots
- P-Plots
- Q-Plots

Bivariate Verteilungen:
- Scatterplots und Scatterplots-Smoothing
- Level-and-Spread-Plots
- nichtparametrische Dichteschätzungen
- QQ-Plots

Mehrdimensionale Verteilungen:
- Q-Plots gegen Gamma-Verteilungen (zur Prüfung auf multivariate Normalverteilung)
- 3D-Scatterplots
- Scatterplotmatrizen
- Coplots
- Parallelkoordinatenplots

Multivariate Verfahren:
- Hauptkomponentenanalyse
- Biplots
- Korrespondenzanalyse
- Multidimensionale Skalierung und „nonlinear mapping"
- Clusterplots
- ANOVA-Plots (Effektplots, Residuendiagnostik, Interaktionsplots)
- Prokrustes-Analysen
- Dynamic Graphics

Der Katalog ist keineswegs vollständig und spiegelt sicherlich subjektive Präferenzen wieder. Über die Details und Begründungen eine solchen Katalogs zu debattieren, ist aber solange müßig, wie diese Techniken nur vereinzelt in einigen wenigen Programmen auftauchen: Viele dieser Plots sind mit Standardprogrammen kaum oder nicht möglich.

12. Was ist multivariate Grafik?

Jeder Versuch, Daten für mehr als zwei Variablen in irgendeiner Form grafisch darzustellen, wird als „multivariate Grafik" bezeichnet. Aufgrund der Begrenzung der Wahrnehmung auf maximal drei Dimensionen müssen bei multivaria-

ter Grafik in irgendeiner Weise die „überzähligen" Variablen in den zweidimensionalen Raum einer Abbildung projiziert werden. Hierzu wurde eine große Zahl von Techniken entwickelt, von denen sich aber nur wenige in der Praxis bewähren konnten.[19]

Hierzu gehören vor allem die Symbolplots verschiedenster Art (z. B. die berühmten Chernoff-Faces) oder der Andrews-Plot, bei dem die Summe der Sinus- bzw. Cosinuswerte der Variablen jeder einzelnen Beobachtung geplottet werden. Keiner dieser Plots hat sich in der Forschungspraxis bewährt.[20]

Die mittlerweile bekannteste Technik multivariater Grafik dürften Scatterplot-Matrizen („Sploms") sein, bei denen alle paarweisen Scatterplots einer Variablenmenge erstellt und in einer Matrix angeordnet werden. Eine extrem nützliche Form von Plots für drei oder vier Variablen sind Coplots, bei denen ein Scatterplot zweier Variablen in Abhängigkeit von der Ausprägung einer dritten und vierten Variablen dargestellt wird (Cleveland 1993, S. 276–281). Eine weitere nützliche und etwas weniger bekannte Technik ist die Verwendung von Parallel-Koordinatenplots, bei denen jede Beobachtung als horizontale Linie in einem Plot dargestellt wird, dessen X-Achse durch eine Abfolge der Variablen gebildet wird.[21]

Insbesondere für die Ausbildung, die Identifikation von Ausreißern und zur Erkennung von Subgruppen eignen sich die Techniken der „dynamischen Grafik". Hierbei werden mehrere grafische Darstellungen der gleichen Daten verbunden („linked-plots"). So kann man z. B. in einer Scatterplot-Matrix innerhalb eines Scatterplots mit einer Maus einen Bereich von Beobachtungen markieren, die dann in allen anderen Scatterplots hervorgehoben werden („brushing").[22]

Zur multivariaten Grafik gehören auch alle Plots der dimensionsreduzierenden Verfahren wie z. B. der Hauptkomponentenanalyse (einschließlich des Biplots), der Korrespondenzanalyse und den verschiedenen Formen der multidimensionalen Skalierung.[23] Interessanterweise gibt es eine große Zahl grafischer Hilfsmittel in der Clusteranalyse, die weder in den meisten Programmen noch in den Standardlehrbüchern zur Clusteranalyse enthalten sind.[24]

Die Analyse der Residuen einfacher statistischer Modelle scheint die interessanteste und erfolgversprechendste Anwendung grafischer Techniken in der Datenanalyse zu sein. In den meisten Fällen bestehen „multivariate Grafiken" dann aus einfachen zweidimensionalen Scatterplots. Lediglich die Variablen, die die Achsen definieren, sind das Resultat „multivariater Statistik". In diesem Sinne bestehen multivariate Plots häufig aus der Anwendung spezieller Formen der Residuenanalyse in Standardverfahren wie z. B. Faktorenanalyse und Varianzanalyse.

13. Wie kann man grafisch gestützte Datenanalyse erlernen?

Zunächst sollte man die Grundlagen statistischer Grafik erlernen. Hierzu gibt es zu dem in jeder Hinsicht hervorragenden Buch von Cleveland (1994) keine Alternative. Weitere Möglichkeiten insbesondere bei der Anwendung multivariater Verfahren finden sich bei Schnell (1994a). Die Lektüre solcher Lehrbücher ist aber lediglich die notwendige Vorbedingung für den eigentlichen Lernprozess: Datenanalyse kann man nur dadurch erlernen, dass man Datenanalysen durchführt. Dies gilt auch für grafisch gestützte Datenanalysen. Falls man diese Techniken wirklich beherrschen will, ist es ratsam, einen eigenen Datensatz zu analysieren. Dies sollte ein „interessanter" Datensatz sein, d. h. ein Datensatz, der zur Beantwortung einer bestimmten Fragestellung erstellt wurde. Die in der Lehre in den Sozialwissenschaften üblichen Mammut-Datensätze (ALLBUS, SOEP) eignen sich hingegen kaum. Man braucht eine inhaltliche Fragestellung und detaillierte Kenntnisse über die Details der Datenerhebung. Neben einem Datensatz und einem inhaltlichen Problem benötigt man ein oder mehrere Analyseprogramme. Hinweise auf geeignete Programme finden sich im nächsten Abschnitt. Mit einem der dort genannten Programme sollten zunächst die Basisplots ausprobiert werden, dann sollte eine Entwicklung eines tatsächlichen Erklärungsmodells erfolgen. Der Schwerpunkt sollte dabei auf den Möglichkeiten der Beurteilung der Modellabweichungen liegen. Unterstützend kann dabei auf die Bücher von Cook (1998) und Cook/Weisberg (1999) zurückgegriffen werden.

14. Mit welcher Software kann man grafisch gestützte Datenanalysen durchführen?

Obwohl die Software-Entwicklung in den letzten fünf Jahren beachtliche Fortschritte erzielt hat, existiert bis heute keine für Sozialwissenschaftler akzeptable Software für grafisch gestützte Datenanalysen.[25] So haben die Standardpakete SPSS und SAS zwar bedeutende neue Features erhalten, wie z. B. Kerndichteschätzer und Scatterplot-Smoother, trotzdem fehlen essentielle Werkzeuge, wie z. B. bedingte Scatterplots, Linked-Plots, Biplots usw.[26] Weiterhin ist die Adaption einzelner Plots an die Bedürfnisse der Nutzer (z. B. Overlays unabhängiger Plots zu einem neuen Plot) schwierig oder unmöglich.

Das in vieler Hinsicht vorbildliche SYSTAT ist bei praktischen Anwendungen „großer" Datensätze (mehr als ca. 2000 Beobachtungen und mehr als 500 Variablen) unakzeptabel langsam und instabil. Noch störender ist die Tatsache, dass es SYSTAT seit mehreren Versionen nicht immer gelingt, einen Bildschirmplot auch als Datei fehlerfrei wiederzugeben.

In den letzten Jahren wurde STATA unter Sozialwissenschaftlern zunehmend populärer (Kreuter/Kohler 2001). Hier müssen allerdings viele Plots erst programmiert werden. Immerhin bietet STATA aber hierzu problemlos alle notwendigen Werkzeuge. Die Architektur von STATA verhindert jedoch leider dynamische Grafik aller Art.[27]

Auf den ersten Blick scheint STATISTICA ein sehr grafisch orientiertes Programm zu sein. Dafür fehlen viele für Sozialwissenschaftler bedeutsam gewordene Analyseverfahren, z. B. Poisson-Regressionen und Ordered-Logit-Modelle. Am gravierendsten ist die Unmöglichkeit, das System um neue Analyse-Plots zu erweitern, wie z. B. ANOVA-Effekt-Plots (Schnell 1994a, S. 279–283).

Als „echte" EDA-Software können wohl nur SAS-JMP und DATADESK betrachtet werden. Für die Analyse großer Datensätze mit Standardverfahren eignen sich beide nur bedingt. Dies gilt ebenso für die beiden bei Statistikern beliebten Programme S-Plus (bzw. die Public Domain Version „R") und LISP-STAT.[28] Ohne Programmierkenntnisse in „S" bzw. LISP sind wirkliche Datenanalysen kaum möglich. Beide eignen sich eher für die Entwicklung neuer Verfahren, weniger für die Analyse tatsächlicher Datensätze. Ähnliches gilt für die Programme, die auf LISP-STAT aufbauen, wie z. B. „Arc" (Cook/Weisberg 1999) oder „ViSta". Beide Programme eignen sich vor allem für die universitäre Lehre.[29] Während Arc für die Regressionsdiagnostik geschrieben wurde, stellt ViSta ein außergewöhnlich leistungsfähiges Programm zur dynamischen Grafik dar. ViSta enthält u. a. Spinplots, Scatterplots, Scatterplotmatrizen, Histogramme, Boxplots, Parallel-Koordinatenplots, Dotplots und Biplots. Die Plots können untereinander verbunden werden, sodass „Brushing" möglich wird. Das Programm enthält Prozeduren zur ANOVA und multiplen Regression, Hauptkomponentenanalyse, multidimensionalen Skalierung und Korrespondenzanalyse. Da aber schon das Datenhandling sozialwissenschaftlicher Datensätze (unterschiedliche Missing-Value-Codes, Änderung der Analyseeinheit [Person – Haushalt]) etc. mit solchen Programmen nahezu unmöglich ist, eignen sie sich nicht für den Datenanalysealltag.

Damit verbleiben z. Z. nur pragmatische Lösungen durch die Anwendung mehrerer Programme. So lässt sich z. B. SPSS mit S-Plus kombinieren (S-Plus arbeitet dann als „Plug-In" für SPSS). Diese Kombination ist ohne Zweifel sehr mächtig, hat allerdings einen in doppelter Hinsicht hohen Preis: Zwei Lizenzen, zwei verschiedene Syntaxformen. Eine nicht ganz so komfortable, dafür aber für sehr große Datensätze (z. B. SOEP) erprobte Lösung ist die Kombination von STATA mit SYSTAT oder AXUM als Grafikprogramm.

15. Warum werden Datenanalysegrafiken in den Sozialwissenschaften so selten verwendet?

Betrachtet man moderne Lehrbücher zu Datenanalysetechniken oder blättert durch die führenden (englischsprachigen) Statistikzeitschriften, so sind die Konsequenzen der ungeheuren Ausweitung der verfügbaren „computing power" der PCs unübersehbar: Die Ersetzung der klassischen Annahmen (z. B. Normalverteilung, Varianzhomogenität, Linearität etc.) durch computerintensive Techniken wie Randomisierung, Bootstrap und generalisierte additive Modelle einerseits und die starke Ausbreitung grafisch gestützter Techniken andererseits. Beide Tendenzen zusammen haben zu ungemein flexibleren und komplexeren Modellen in der Datenanalyse geführt.[30]

Umso bemerkenswerter ist die nahezu vollständige Abwesenheit dieser Techniken in der sozialwissenschaftlichen Forschungsliteratur.[31] Mag dies für die computerintensiven Techniken aufgrund ihrer relativen technischen Schwierigkeiten noch verständlich erscheinen, so ist die weit gehende Abwesenheit grafischer Techniken in den Sozialwissenschaften (vgl. Cleveland 1984) zunächst rätselhaft.

Zwei verschiedene Ursachen erscheinen plausibel:

- Erstens die übliche Ausbildung in Statistik für Sozialwissenschaftler,
- zweitens das Ziel der Datenanalyse.

In der Ausbildung von Mathematikern und vielen Statistikern spielen tatsächliche Datenanalysen kaum eine Rolle. Da andererseits die mathematischen Kenntnisse bei Fachwissenschaftlern meist geringer sind, fällt die Ausbildung von Studenten in Datenanalyse (und das Verfassen der Lehrbücher) häufig an die anscheinend besser qualifizierten Experten für stochastische Prozesse oder lineare Algebra. An inhaltlichen Problemen (und damit an „echten Datensätzen") sind Statistiker aber häufig nicht interessiert. Für die immanenten Probleme der Stochastik oder der linearen Algebra benötigt man aber keine Plots. Diese werden daher auch kaum gelehrt. Aus dem gleichen Grund ist Datenanalysegrafik zumindest im deutschsprachigen Raum kein populäres Forschungsgebiet der Statistik, was sich nicht nur in den Lehrbüchern, sondern auch in den statistischen Zeitschriften zeigt. Da im Gegensatz zu Statistikern Fachwissenschaftler Daten analysieren, um ein inhaltliches Problem zu klären und nicht um eine Technik zu demonstrieren, ist eine Unterscheidung zwischen Statistikern und Datenanalytikern nützlich.[32] Datenanalytiker haben ein Interesse an der Analyse „tatsächlicher" Datensätze (mit vielen Beobachtungen, mit fehlenden Werten, mit Ausreißern, mit unsauberen Messungen). Bei einer technisch korrekten Datenanalyse realer Datensätze zeigt sich die prinzipielle Unverzichtbarkeit von Plots hingegen meist rasch.

Datenanalyseplots sind als Hilfsmittel für den Umgang mit Daten natürlich nur dann unersetzlich, wenn das Ziel einer Datenanalyse darin besteht, etwas über den datengenerierenden Prozess zu lernen. Dies ist keineswegs selbstverständlich. Die sozialen und institutionellen Bedingungen vieler Datenanalysen in der Praxis (in der Medizin, Soziologie, Psychologie etc.) zwingen viele Datenanalytiker dazu, „erwartete" Ergebnisse zu produzieren. Die Betreiber von „Normalwissenschaft" in diesem Sinne, wie sie sich z. B. in Diplomarbeiten, Dissertationen und Forschungsberichten findet, sind nicht an der Gewinnung neuer Einsichten interessiert, sondern an der Demonstration ihrer persönlichen Fähigkeit zur Produktion „signifikanter" Parameterschätzungen oder „interpretierbarer Ergebnisse". Die völlige Vernachlässigung der Residuenanalyse oder gar der Datenbereinigung ist die einzige rationale Wahl von Akteuren unter solchen Produktionsbedingungen. Weiterhin gibt es eine beträchtliche Zahl von Zeitschriftenpublikationen, die zwar oberflächlich wie eine empirische Arbeit aussehen, aber eigentlich ein Tutorial für die Durchführung einer – für Soziologen – neuen Datenanalysetechnik sind. Diese Arbeiten zeichnen sich meist durch die Abwesenheit theoretischer Überlegungen und das Ignorieren der Probleme der Datenbasis aus: Erklärtes Ziel ist eben keine Datenanalyse, sondern „Lehre" bzw. die Demonstration der Kompetenz der Autoren. Unter diesen Bedingungen spielt die Güte einer Modellanpassung keine Rolle.

Es ist daher wenig erstaunlich, dass ein großer Teil der sozialwissenschaftlichen Zeitschriftenliteratur kaum eine Diskussion der Güte der Modelle enthält. Entsprechend zeigt die Auszählung aller zwischen 1990 und 1999 erschienenen Artikel der „Kölner Zeitschrift für Soziologie und Sozialpsychologie" und der „Zeitschrift für Soziologie" bei Schnell (2000), dass von den ca. 240 Arbeiten, die eine quantitative Datenanalyse durchführten, ca. die Hälfte der Arbeiten keine Beurteilung des Modellfits enthielten.

Die derzeit populärsten multivariaten Datenanalyseverfahren neben der multiplen Regression sind die Ereignisdatenanalyse (insbesondere in Form der Cox-Regression) sowie die logistische Regression. Bei beiden neueren Techniken ist das häufigste Vorgehen identisch: Nach der Wahl einer abhängigen Variablen werden eine mehr oder weniger beliebige – und fast nie vorher theoretisch hergeleitete – Menge von Variablen in das Erklärungsmodell aufgenommen und der entsprechende Parameter gegen Null getestet. Bei einigen Arbeiten werden verschiedene Modelle anhand diverser Likelihood-Ratio-basierter Statistiken verglichen, z. B. anhand von AIC oder BIC. Bis auf einzelne Ausnahmen wird weder die Untersuchung der Residuen noch die Wiederholung der Modellanpassung an einer neuen Stichprobe erwähnt. Es scheint kaum besonders gewagt, wenn man vermutet, dass in der überwiegenden Mehrzahl der Fälle keine solchen detaillierten Goodness-of-Fit-Analysen erfolgten.[33]

Diese Art von Datenanalyse besitzt einen – je nach Perspektive – bemerkenswerten Vor- bzw. Nachteil: Sie kann nicht scheitern. Sobald die Daten nicht wirklich aus einem Pseudo-Zufallszahlengenerator stammen und die Stichprobe hinreichend groß ist, sind signifikante Ergebnisse garantiert. Da eine triviale Struktur bei sozialwissenschaftlichen Individualdatensätze schon aufgrund der soziodemografischen Variablen (z. B. als Indikator für Stellung im Lebenszyklus) erwartbar ist, sollten auch die Variablen im „Modell" die „richtigen" Vorzeichen besitzen.

Dabei wird von den Autoren eine einfache Tatsache vergessen: Dass die Daten einem Modell nicht widersprechen, ist kein Beweis für die Gültigkeit des Modells. Dieser kann nur durch die erfolgreiche Anwendung des Modells auf neue Daten erfolgen. Der prognostische Erfolg eines Modells erscheint aber eher gegeben, wenn die Voraussetzungen des Modells erfüllt sind. Und diese lassen sich ohne grafische Techniken kaum beurteilen.

16. Was hat Datenanalysegrafik mit Inferenzstatistik zu tun?

Erstens unterliegen wie alle Stichprobenstatistiken auch die Ergebnisse grafisch gestützter Datenanalyse Stichprobenfehlern. Zweitens werden von einigen Autoren Datenanalysegrafiken als Ersatz für Signifikanztests empfohlen. Beides muss etwas näher erläutert werden.

Da bei wissenschaftlichen Fragestellungen die Daten fast immer nur Stichproben aus der eigentlich interessierenden Population darstellen, liefern die grafischen Darstellungen nur Ergebnisse für die jeweilige Stichprobe. Folglich sind die Ergebnisse einer grafisch gestützten Datenanalyse auch mit Stichprobenfehlern behaftet. Bei Interpretationen der Plots multivariater Grafik, insbesondere derjenigen der dimensionsreduzierenden Verfahren, wird dies häufig vergessen.

Ein schönes Beispiel liefert die Interpretation einer Korrespondenzanalyse. Korrespondenzanalysen sind in einigen Teilbereichen der Soziologie relativ populär. Bei einer Korrespondenzanalyse werden die Residuen eines Chi-Quadrat-Unabhängigkeitsmodells zumeist durch einen Plot der beiden ersten Eigenvektoren grafisch dargestellt. Bei der üblichen Interpretation werden entweder nahe beieinander liegende Objekte im Plot als ähnlich interpretiert oder Objekte nach ihrer relativen Lage zu den Achsen des Plots beurteilt. Beide Varianten setzten voraus, dass sich die Lage der Objekte im Raum nicht allein durch Stichprobenschwankungen wesentlich ändern kann. Nun kann man zwar bei einer Korrespondenzanalyse keine Konfidenzintervalle berechnen, man kann aber durch Resampling-Techniken die Analyse anhand derselben Stichprobe mehrfach wiederholen. Die durch die Stichprobenziehung bedingte Unsicherheit

lässt sich mit einem Plot, in dem die bei nach Resampling wiederholten Analysen berechneten Koordinaten eines Objekts gemeinsam dargestellt werden, leicht beurteilen.[34] In der Praxis der Anwendung der Korrespondenzanalysen finden sich solche Beurteilungen des Effekts von Stichprobenschwankungen kaum.

Inferenzstatistik wird in den Sozialwissenschaften überwiegend in Form von Signifikanztests angewendet. Die exzessive Anwendung von Signifikanztests ist in den Sozialwissenschaften seit Jahrzehnten umstritten.[35] Die Hauptkritikpunkte sind:[36]

1. Signifikanztests beantworten die Frage, wie häufig der beobachtete Effekt auftreten würde, wenn die Nullhypothese korrekt wäre (also $p(D|H_0)$). Wissenschaftler interessieren sich aber eher für die Wahrscheinlichkeit der Forschungshypothese bei gegebenen Daten (also $p(H_1|D)$).
2. Bereits vor den Tests ist in der Regel bekannt, dass die üblichen Nullhypothesen falsch sind. Entsprechend wenig lernt man aus ihrer Ablehnung.
3. Die Ja-Nein-Antwort eines Signifikanztests bei fixem Alpha ist keine angemessene Art der Quantifizierung der Unsicherheit über die Größe eines Parameters.

Diese und weitere Kritikpunkte an der Theorie und vor allem an der Praxis der Signifikanztests haben einige Autoren dazu geführt, Signifikanztests völlig abzulehnen und stattdessen Effektstärkemaße zu berechnen, Meta-Analysen durchzuführen und/oder Daten grafisch darzustellen.[37] Sicherlich kann die Debatte um den Sinn der Signifikanztests hier nicht entschieden werden.[38] Trotzdem kann festgehalten werden, dass bei vielen inhaltlichen Problemen andere Techniken sinnvoller sind als Signifikanztests.

Hierzu gehören sicherlich meist Fragen nach der Effektstärke und der Replizierbarkeit des Effekts. Obwohl sich Effektstärken fast immer problemlos berechnen lassen, werden sie weitaus seltener als Signifikanzniveaus berichtet. Will man die tatsächliche Stärke eines Effekts beurteilen, sind zumindest zusätzliche Plots fast immer erforderlich. Plots der Rohdaten, getrennt nach experimentellen Bedingungen bzw. Residuen- und Fitplots, geben meist weit interessantere Aufschlüsse als Signifikanztests. Viele Debatten um die „Existenz" von Effekten hätten sich vermeiden lassen, hätte man zusätzlich zu den Signifikanztests einige Plots der Daten, z. B. getrennt nach experimentellen Bedingungen, veröffentlicht, denen die Effektstärken, die mögliche Existenz von Ausreißern und „ungewöhnlichen" Verteilungen (extreme Schiefe, Multimodalität) zu entnehmen gewesen wären.

Trotzdem erwecken viele Statistiklehrbücher den Eindruck, dass das Ziel einer Datenanalyse immer eine Parameterschätzung und/oder ein Signifikanztest ist. Dies ist in der Forschungspraxis eher selten. Hier sind Datenanalyseplots fast immer nützlicher als Signifikanztests – wenn man etwas über die datengenerierenden Prozesse lernen möchte.

17. Die Rolle grafisch gestützter Datenanalyse für die Entwicklung empirisch bewährter Modelle

„Tools matter." (Cleveland 1993, S. 340)

Sowohl in der Theorieentwicklung als auch in den praktischen Anwendungen (Policy-Forschung, Marktforschung etc.) stehen die Ergebnisse der Forschung von Soziologen und Politologen in einem immer stärkeren Wettbewerb mit Modellen, die von Ökonomen entwickelt wurden. Sollten Soziologen und Politologen daran interessiert sein, diesen Wettbewerb zu bestehen, dann werden sie langfristig gezwungen sein, nachweisbar prognosefähige Modelle für sozialwissenschaftlich relevante Probleme zu entwickeln. Hierzu sind die Techniken grafisch gestützter Datenanalyse vor allem in der Residuendiagnose unverzichtbar. Die Notwendigkeit grafischer Techniken wird auch von Vertretern anderer Fachgebiete gesehen (vgl. Loftus 1996). Behrens (1997, S. 154–155) fordert für die Psychologie unter anderem eine verstärkte Berücksichtigung der EDA bei der Veröffentlichung empirischer Arbeiten in Zeitschriften sowie bei der Ausbildung der Graduierten. Soziologen und Politologen scheinen hingegen zu großen Teilen immer noch zu glauben, lediglich mit Fallstudien oder einer ritualisierten Sozialforschung, für die empirische Ergebnisse letztlich keine Bedeutung für die „allgemeine Theorie" hat, auskommen zu können. Sollte das Ziel sozialwissenschaftlicher Forschung hingegen in der Entwicklung empirisch fundierter und prognosefähiger Theorien bestehen, dann scheint dieses ohne eine grundlegende Veränderung der universitären Ausbildung und der Praxis der Datenanalyse kaum erreichbar. Grafisch gestützte Modellentwicklung ist dabei unverzichtbar.

Anmerkungen

1 Dieser Artikel enthält die Kernthesen und deren Weiterentwicklung meiner 1994 erschienen Monographie (Schnell 1994a).

2 Zu den Regeln der Präsentationsgraphik gehören z. B. die Ausstattung eines Diagramms mit ausführlichen Legenden und einsichtigen Skalierungen. Solche Regeln finden sich z. B. bei Wainer (1984) und Burn (1993). Zu den kognitionspsychologischen Grundlagen vgl. Kosslyn (1994).

3 Aus diesem Grund sind viele der Regeln, die für Präsentationsgraphiken unverzichtbar sind, für Datenanalysegraphiken kaum sinnvoll: Die meisten Datenanalyseplots sind „Wegwerf-Plots", die niemand außer dem Datenanalytiker je sehen wird.

4 Dies führt einige Datenanalytiker zu der Forderung, dass jede Interpretation einer Teststatistik von Plots begleitet werden sollte (vgl. z. B. Hadi 1993, S. 775).

5 Dieser Abschnitt folgt der Darstellung bei Box (1976, S. 793 und S. 796). Tukeys Äußerungen sind stets mit dieser Konzeption verträglich, vgl. vor allem Tukey/Wilk (1970, S. 372 und S. 385), Mallows/Tukey (1982) und Tukey (1977, S. vii, 1980, 1990).

6 Die hier zugrunde liegende Haltung findet sich in vielen Arbeiten, z. B. bei Weihs/Schmidli (1990) oder Wegman/Carr (1993, S. 919). So schreiben Young/Kent/Kuhfeld (1988, S. 422): „We are very optimistic that highly integrated, highly interactive, appropriately interfaced MEDA/VEDA systems will become very useful tools for exploring, understanding, and forming hypotheses about the structure of multivariate data". Entsprechend behaupten Young/Faldowski/McFarlane (1993, S. 959) „(. . .) scientific exploration leads to scientific hypotheses". Levkovits (1991, S. 60) geht noch weiter: „Exploratory visualization is required when data is so complex that the scientist does not necessarily understand *what* (kursiv im Original, R. S.) needs to be displayed."

7 Kuhn (1978, S. 293) hat dies anhand vieler historischer Beispiele belegt: „*Der Weg vom wissenschaftlichen Gesetz zur wissenschaftlichen Messung lässt sich nur selten in umgekehrter Richtung gehen* (kursiv im Original, R. S.). Um quantitative Gesetzmäßigkeiten zu entdecken, muss man gewöhnlich wissen, was für eine Gesetzmäßigkeit man sucht, und die Instrumente müssen dementsprechend konstruiert sein; und selbst dann liefert die Natur nicht immer kampflos konsistente oder verallgemeinerungsfähige Ergebnisse."

8 Holland u. a. (1987, S. 326). Hierzu werden mentale Repräsentationen der Objekte benötigt. Die graphische Darstellung von Daten kann als Verfahren zur Erhöhung der Menge gleichzeitig verarbeitbarer Informationen aufgefasst werden (Faust 1984, S. 112). Langley u. a. (1987, S. 329) schreiben hierzu: „The evidence suggests that processing information in a drawing or a chart and processing it in the ‚mind's eye' have much in common. That is to say, the kinds of inferences that can be retained readily in the two cases are highly similar. More information can be retained reliably in the display on paper than in the limited memory capacity of the ‚mind's eye', but this seems to be the principal difference between the two representations."

9 Dies zeigen die Arbeiten zum Problem induktiver Generalisierungen der „cognitive science" an Beispielen der Wissenschaftsgeschichte sowie auch anhand von Computersimulationen des Entdeckungsprozesses, vgl. Langley (1987) sowie Holland u. a. (1987).

10 Tukey/Wilk (1970, S. 272): „Some prior presumed structure, some guidance, some objectives, in short some ideas of a model, are virtually essential, yet these must not be taken too seriously. Models must be used but must never be believed."

11 Vgl. hierzu Hand (1998, S. 113–114), Westphal/Blaxton (1998, S. 75–121), Berry/Linoff (2000, S. 50–53 und S. 177–181).

12 Vgl. hierzu Chatfield (1993, S. 22–47) und Wilkinson (1999).

13 Dies sind z. B. hierarchische lineare Modelle, vgl. einführend z. B. Snijders/Bosker (1999).

14 Interessant ist dies auch in Hinsicht auf die von Freedman (1985, 1987, 1991) belebte Debatte um die sinnvolle Anwendung von Regressionsverfahren in den Sozialwissenschaften.

15 Eine solche Art der Datenanalyse setzt neben leistungsfähiger Hardware vor allem Software voraus, die die problemlose und schnelle Erstellung einer Vielzahl verschiedener Datenanalyseplots unterstützt. Einzelprogramme, die neben einem Standardpaket verwendet werden müssen, sind daher für praktische Datenanalysen dieses Typs meistens sinnlos.

16 So erscheinen vor allem im „Journal of the American Statistical Association", „The American Statistician" und in dem seit 1992 erscheinenden „Journal of Computational and Graphical Statistics" regelmäßig neue Formen graphischer Darstellungen.
17 Zu den Prinzipien und experimentellen Befunden zur Wahrnehmung multivariater Graphiken vgl. Yu (1995).
18 Einzelheiten zu allen genannten Verfahren finden sich bei Schnell (1994a).
19 Vgl. hierzu einführend Jacoby (1998). Vertiefende und sich ergänzende Darstellungen finden sich bei Cleveland (1993, S. 272–340) und Schnell (1994, S. 125–215).
20 Dies gilt besonders für die statistisch interessanten Varianten des „Projection Pursuit", bei der diejenigen zweidimensionalen Projektionen eines hochdimensionalen Raums geplottet werden, die „interessante" Abweichungen von den Projektionen „zufälliger" Verteilungen ergeben, vgl. hierzu ausführlich Klinke (1997, S. 102–166). Ein PP-Programm findet sich bei *„http://www.research.att.com/areas/stat/xgobi"*.
21 Da das Aussehen dieses Plots stark von der Abfolge der Variablen auf der X-Achse abhängt, erlauben einige Programme eine automatische Permutation der Reihenfolge der Variablen („tour").
22 Diese einfache Form dynamischer Graphik findet sich mittlerweile in einigen populären Statistik-Programmen wie z. B. Statistica. „Linked-Plots" für verschiedene Darstellungsformen der Daten finden sich aber nur in Programmen wie Data-Desk und ViSta, vgl. hierzu Frage 14.
23 Zu den verschiedenen Formen des Biplots vgl. Gower/Hand (1996), zur MDS vgl. Cox/Cox (1994). Man kann die Plots verschiedener Projektionstechniken besser miteinander vergleichen, wenn sie für einen Vergleich optimal rotiert und skaliert werden (so genannte Prokrustes-Analyse, vgl. Schnell 1994, S. 209–212). Auch diese Technik ist in kaum einem Programm vorhanden.
24 Gerade diese Plots können aber die Interpretation der Ergebnisse einer Clusteranalyse wesentlich erleichtern, z. B. Thresholdplots, Silhouettenplots und Plots des „Minimum Spanning Trees". Zu all diesen und weiteren Plots in der Clusteranalyse vgl. Schnell (1994a, S. 291–326).
25 Zur den Kriterien und der Kritik früherer Implementierungen vgl. Schnell (1994b) und Schnell/Matschinger (1994).
26 Die Internet-Adressen der Hersteller aller genannten Programme finden sich unter *„http://www.stata.com/links/stat_software.html"*.
27 Die Version 7 erlaubt nun auch die Durchführung von Clusteranalysen. Damit fehlt dem Programm von den klassischen Techniken multivariater Graphik nur noch die MDS.
28 Zu R, S und S-Plus vgl. einführend Spector (1994) sowie Venables/Ripley (1997) für statistische Anwendungen und Venables/Ripley (2000) zur Programmierung. Zu LISP-Stat vgl. Tierney (1990).
29 Beide Programme sind kostenlos erhältlich, vgl. für Arc *„http://www.wiley.com"* den Link für Cook/Weisberg (1999) „Applied Regression"; für ViSta *„http://www.visualstats.org"*.
30 Zur Illustration eignet sich jedes beliebige Datenanalysegebiet, als Beispiel sei der Vergleich älterer Lehrbücher zur multivariaten Analyse mit modernen Lehrbüchern (wie Hamilton 1993 oder Cook/Weisberg 1999) empfohlen.
31 Dass sich Beispiele in den Lehrbüchern und den Methodenzeitschriften finden, widerspricht dieser These nicht.
32 Vgl. Tukey (1962) und Cooley/Lohnes (1971, S. v).

33 Generell scheinen die Diagnosetechniken bei Sozialwissenschaftlern weitgehend unbekannt zu sein. Die meisten angewandten Lehrbücher für Sozialwissenschaftler diskutieren diese Techniken kaum; die entsprechende Literatur (z. B. Collett 1991, S. 120–187, Collett 1994, S. 149–198, Pötter 1988, Wu 1990) wird kaum zitiert.

34 Beispiele finden sich bei Greenacre (1993, S. 174–177) und Schnell (1994, S. 196–197). Ein entsprechendes Gauss-Programm findet sich unter „*http://www.uni-konstanz.de/FuF/Verwiss/Schnell/prog.htm*"

35 Zu den wesentlichen Argumenten vgl. z. B. die Arbeiten in dem von Harlow u. a. 1997 herausgegebenen Band sowie dessen kritische Besprechung bei Krantz (1999).

36 Die Darstellung folgt hier Judd u. a. (1995, S. 437–438).

37 Vgl. hierzu unter anderem Loftus (1996) sowie Schmidt/Hunter (1997).

38 Vgl. hierzu Abelson (1997) und Meehl (1997).

Literatur

Abelson, R. P.: A Retrospective on the Significance Test Ban of 1999; in: Harlow, L. L./Mulaik, S. A./Steiger, J. H. (eds.): What if there were no significance tests?, Mahwah/N. J. 1997, S. 117–141.

Banks, W. P./Krajicek, D.: Perception; in: Annual Review of Psychology, 42, 1991, S. 305–331.

Behrens, J. T.: Principles and Procedures of Exploratory Data Analysis; in: Psychological Methods, 2, 2, 1997, S. 131–160.

Berry, M. J. A./Linoff, G. S.: Mastering Data Mining, New York 2000.

Box, G. E. P.: Science and Statistics; in: Journal of the American Statistical Association, 71, 356, 1976, S. 791–799.

Burn, D. A.: Designing Effective Statistical Graphs; in: Rao, C. R. (ed.): Handbook of Statistics, Vol. 9, New York 1993, S. 745–773.

Chatfield, C.: Problem Solving. A Statistician's Guide, London 1993.

Cleveland, W. S.: Graphs in Scientific Publications; in: The American Statistician, 38, 4, 1984, S. 261–269.

Cleveland, W. S.: The Elements of Graphing Data, Murray Hill 1993.

Cleveland, W. S.: Visualizing Data, Murray Hill 1994.

Collett, D.: Modelling Binary Data, London. 1991.

Collett, D.: Modelling Survival Data in Medical Research, London 1994.

Cooley, W. W./Lohnes, P. R.: Multivariate Data Analysis, New York 1971.

Cook, R. D.: Regression Graphics, New York 1998.

Cook, R. D./Weisberg, S.: Graphs in Statistical Analysis: Is the Medium the Message? In: The American Statistician, 53, 1, 1999, S. 29–37.

Cook, R. D./Weisberg, S.: Applied Regression including Computing and Graphics, New York 1999.

Cox, C./Gabriel, K. R.: Some Comparisons of Biplot Display and Pencil and Paper E. D. A. Methods; in: Launer, R. L./Siegel, A. F. (eds.): Modern Data Analysis, New York 1982, S. 45–82.

Cox, D. R./Snell, E. J.: A General Definition of Residuals; in: Journal of the Royal Statistical Society, Series B, 30, 2, 1986, S. 248–265.

Cutting, J. E.: Why our Stimuli Look as They do; in: Lockhead, G. R./Pomerantz, J. R. (eds.): The perception of structure, Washington 1991, S. 41–52.
Cox, T. F./Cox, M. A. A.: Multidimensional Scaling, London 1994.
Erickson, B. H./Nosanchuck, T. A.: Understanding Data, Toronto 1977.
Faust, D.: The Limits of Scientific Reasoning, Minneapolis 1984.
Freedman, D. A.: Statistics and the Scientific Method; in: Mason, W. M./Fienberg, S. E. (eds.): Cohort Analysis in Social Research, New York 1985, S. 343–390 (mit Diskussion).
Freedman, D. A.: As Others See Us: A Case Study in Path Analysis; in: Journal of Educational Statistics, 12, 2, 1987, S. 101–128.
Freedman, D. A.: Statistical Models and Shoe Leather; in: Marsden, P. V. (ed.): Sociological Methodology 1991, vol. 21, 1991, S. 291–313.
Friendly, M.: SAS System for Statistical Graphics, Cary/NC 1991.
Gower, J. C./Hand, D. J.: Biplots. London 1996.
Greenacre, M. J.: Correspondence Analysis in Practice, London 1993.
Hadi, A. S.: Graphical Methods for Linear Models; in: Rao, C. R. (ed.): Handbook of Statistics, Vol. 9, New York 1993, S. 775–802.
Hamilton, L. C.: Regression with Graphics. Pacific Grove 1992.
Hand, D. J.: Data Mining: Statistics and more?; in: The American Statistician, 52, 2, 1998, S. 112–118.
Harlow, L. L./Mulaik, S. A./Steiger, J. H. (eds.): What if there were no significance tests?, Mahwah/N. J. 1997.
Heiler, S./Michels, P.: Deskriptive und Explorative Datenanalyse, München 1994.
Hoaglin, D. C.: Exploratory Data Analysis; in: Kotz, S./Johnson, N. L.(eds.): Encyclopedia of Statistical Sciences, New York 1985, Vol. 2, S. 579–583.
Holland, J../Holyoak, K. J./Nisbett, R. E./Thagard, P. R.: Induction. Processes of Inference, Learning and Discovery, London 1987, 2. Auflage.
Jacoby, W. G.: Statistical graphics for visualizing multivariate data, Thousand Oaks 1998.
Jöreskog, K.: From Factor Analysis to Advanced Structural Equation Modelling: Perspectives, Visions and Trends, Vortrag am 4. 10. 2000 anlässlich der 5. „International Conference on Logic and Methodology" in Köln.
Judd, C. M./McClelland, G. H./Culhane, S. E.: Data Analysis: Continuing Issues in the Everyday Analysis of Psychological Data; in: Annual Review of Psychology, 46, 1995, S. 433–465.
Klinke, S.: Data Structures for Computational Statistics, Heidelberg 1997.
Kohler, U./Kreuter, F.: Datenanalyse mit Stata. Allgemeine Konzepte und ihre praktische Anwendung. München 2001.
Kosslyn, S. M.: Elements of Graph Design, New York 1994.
Kuhn, T. S.: Die Funktion des Messens in der Entwicklung der physikalischen Wissenschaften; in: Kuhn, T. S.: Die Entstehung des Neuen: Studien zur Struktur der Wissenschaftsgeschichte, Frankfurt 1978, S. 254–307.
Krantz, D. H.: The Null Hypothesis Testing Controversy in Psychology; in: Journal of the American Statistical Association, 44, 448, 1999, S. 1372–1381.
Langley, P./Simon, H. A./Bradshaw, G. L./Zytkow, J. M.: Scientific Discovery. Computational Explorations of the Creative Processes, Cambridge, Mass. 1987.
Levkowitz, H.: Exploratory Data Visualization: The Human Visual System should be the Main Design Consideration; in: ASA Proceedings of the Section on Statistical Graphics, 1991, S. 60–63.

Loftus, G. R.: Psychology Will be a Much Better Science When We Change the Way We Analyze Data; in: Current Directions in Psychological Science, 5, 6, 1996, S. 161–171.

Lovie, P.: Rezension von Cleveland (1985): The Elements of Graphing Data; in: Applied Statistics, 1987, 36, S. 376.

Lubinsky, D./Pregibon, D.: Data Analysis as Search; in: Journal of Econometrics, 38, 1988, S. 247–268.

Mallows, C. L./Walley, P.: A Theory of Data Analysis? in: ASA Proceedings of the Business and Economic Statistics Section, 1980, S. 8–14.

Mallows, C. L./Tukey, J. W.: An Overview of Techniques of Data Analysis, Emphasizing its Exploratory Aspects; in: Tiago de Oliveira, J./Epstein, B. (eds.): Some Recent Advances in Statistics, London 1982, S. 111–172.

Meehl, P. E.: The Problem is Epistemiology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions; in: Harlow, L. L./Mulaik, S. A./Steiger, J. H. (eds.): What if there were no significance tests?, Mahwah/N. J. 1997, S. 395–425.

Pinker, S.: A Theory of Graph Comprehension; in: Freedle, R. (ed.): Artificial Intelligence and the Future of Testing, Hillsdale 1990, S. 73–126.

Pötter, U.: Graphische Diagnoseverfahren für Cox-Modelle, Arbeitspapier Nr.278, SFB 3, Frankfurt/Mannheim 1988.

Schmidt, F. L./Hunter, J. E.: Eight Common but false Objections to the Discontinuation of Significance Testing in the Analysis of Research Data; in: Harlow, L. L./Mulaik, S. A./Steiger, J. H. (eds.): What if there were no significance tests?, Mahwah/N. J. 1997, S. 37–64.

Schnell, R.: Graphisch gestützte Datenanalyse, München 1994.

Schnell, R.: Requirements of statistical graphic systems and currently available software; in: Faulbaum, F. (ed.): SoftStat'93: Advances in Statistical Software 4, Stuttgart 1994, S. 311–316.

Schnell, R.: Ritualisierte Sozialforschung. Zum Verhältnis von sozialwissenschaftlicher Theorie und der Praxis der Datenanalyse der letzten Dekade; unveröffentlichtes Vortragsmanuskript, Konstanz 2000.

Schnell, R./Matschinger, M.: Multivariate Graphics: Current Use and Implementations in the Social Science; in: Dirschedl, P./Ostermann, R.(eds.): Computational Statistics, Heidelberg 1994, S. 275–294

Snijders, T./Bosker, R.: Multilevel Analysis, London 1999.

Spector, P.: An Introduction to S and S-Plus, Belmont 1994.

Tierney, L.: Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics. New York 1990.

Tukey, J. W.: The Future of Data Analysis; in: The Annals of Mathematical Statistics, 33, 1962, S. 1–67.

Tukey, J. W.: Exploratory Data Analysis. Reading, Mass. 1977.

Tukey, J. W.: We Need Both Exploratory and Confirmatory; in: American Statistician, 34, 1980, S. 23–25.

Tukey, J. W.: Data-Based Graphics: Visual Display in the Decades to Come; in: Statistical Science, 5, 3, 1990, S. 327–339.

Tukey, J. W./Wilk, M. B.: Data Analysis and Statistics: Techniques and Approaches; in: Tufte, E. R. (ed.): The Quantitative Analysis of Social Problems, Reading/Mass. 1970, S. 370–390 (Original 1965).

Venables, W. N./Ripley, B. D.: Modern Applied Statistics with S-Plus, 3. ed., New York. 1999.

Venables, W. N./Ripley, B. D.: S Programming, New York 2000.

Wainer, H.: How to Display Data Badly; in: The American Statistician, 38, 2, 1984, S. 137–147.

Wegman, E. J./Carr, D. B.: Statistical Graphics and Visualization; in: Rao, C. R. (ed.): Handbook of Statistics, Vol. 9, New York 1993, S. 857–958.

Weihs, C./Schmidli, H.: OMEGA (Online Multivariate Exploratory Graphical Analysis) Routine Searching for Structure; in: Statistical Science, 5, 2, 1990, S. 175–226 (mit Diskussion).

Wilkinson, L.: Graphs for Research in Counseling Psychology; in: The Counseling Psychologist, 27, 3, 1999, S. 384–407.

Westphal, C./Blaxton, T.: Data Mining Solutions. Methods and Tools for Solving Real-World Problems, New York 1998.

Wu, L. L.: Simple Graphical Goodness-of-Fit Tests for Hazard Rate Models; in: Mayer, K. U./Tuma, N. B. (eds.): Event History Analysis in Life Course Research, Madison 1990, S. 184–199.

Young, F. W./Kent, D. P./Kuhfeld, W. F.: Dynamic Graphics for Exploring Multivariate Data; in: Cleveland, W. S./McGill, M. E. (eds.): Dynamic Graphics for Statistics, Belmont 1988, S. 391–324.

Young, F. W./Faldowski, R. A./Mcfarlane, M. M.: Multivariate Statistical Visualization; in: Rao, C. R. (ed.): Handbook of Statistics, Vol. 9, New York 1993, S. 959–998.

Yu, C. H.: The interaction of research goal, data type, and graphical format in multivariate visualization. Unpublished dissertation. Arizona State University 1995.