Didaktik der Explorativen Datenanalyse

Bericht über einen Vortrag von Herrn Prof. Dr. H. H. Bock
vom Institut für Statistik und Wirtschaftsmathematik der RWTH Aachen, gehalten am 9.11.1984 anläßlich einer Fortbildungsveranstaltung des Vereins zur Förderung des schulischen Statistikunterrichts an der Universität Dortmund - Abteilung Statistik.

von Ingeborg Strauß, Kronberg

Der eineinhalbstündige Vortrag besaß drei Schwerpunkte:

1. Grundlegende Verfahrensweisen der Explorativen Datenanalyse (EDA),
2. Abgrenzung gegenüber/Gemeinsamkeiten mit der klassischen Statistik,
3. Relevanz der EDA für den Schulunterricht.

Den Ausführungen schloß sich eine lebhafte gut einstündige Diskussion an. Neben Inhaltlichem (s.u.) wurde auch Literatur zur EDA mitgeteilt. Interessenten dafür können sich an die Verfasserin dieses Artikels oder an Herrn König vom Zentralblatt für Didaktik der Mathematik, Karlsruhe, wenden.

-.-

Das Grundproblem der EDA lautet: Ich habe Daten, irgendwoher, sie passen auf kein bekanntes probabilistisches Modell, es beginnt die Suche nach unbekannten datenimmanenten Strukturen, z.B. einem zeitlichen Entwicklungsprozeß oder einem räumlichen Verschiebungsmuster. Das "data snooping" folgt keinen strengen Regeln, sondern lebt von der Intuition und dem Einfallsreichtum des Statistikers. Doch gibt es einige aus der Praxis erwachsene Empfehlungen für die Vorgehensweise(n), die stichwortartig und ohne Anspruch auf Vollständigkeit angegeben seien:

Man benütze/entwickle suggestive Diagramme und untersuche sie auf Zusammenhänge, Ähnlichkeiten, Auffälligkeiten ... .

Man achte auf variable Sicht- und Darstellungsweise, benütze möglichst mehrere Modelle gleichzeitig (Modellanpassung ↔ Modellfreiheit). Transformationen und Kombinationen dienen der Herausbildung wesentlicher Begriffe, die Interaktivität mit dem Computer und die Interpretation anhand der Modelle fördern neue Erkenntnisse zutage.

Man suche nach 'Outliern'. Ausreißer können Anlaß zur Modellverbesserung sein. Ursachen von Abweichungen und Auffälligkeiten können Erhebungs-

Stochastik in der Schule, Jg. 2, H. 2 (1985)

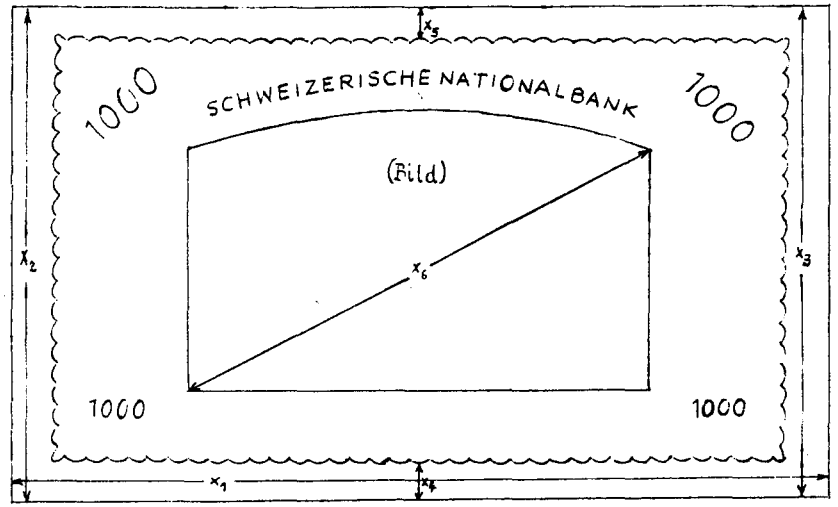
fehler, neue Merkmale etc. sein, alles Anlaß für weitergehende Fachfragen.

Leitgedanke der EDA ist das offene Konzept, eine Vorgehensweise, die bei hochdimensionalen Datensätzen am deutlichsten ihre Vorzüge zeigt.

-.-

Als graphische und rechnerische Techniken werden solche bevorzugt, die r o b u s t sind, ein Stichwort, das die ganze EDA wie ein roter Faden durchzieht. So spielen die Mediane eine weit größere Rolle als in der klassischen deskriptiven Statistik. 'Rahmen-Antennen-Bilder', die den Median, die Quartile und Semiquartile sowie Ausreißer, Maximum und Minimum zeigen (o o o ⊢---┤□├---⊣ o o), sind neben den 'Stengel und Blätter-Darstellungen' die wohl geläufigsten einfachen graphischen Hilfsmittel, die am Beginn einer Untersuchung stehen (können). Auch die sogenannten, aus für den Datensatz charakteristischen Zahlenwerten bestehenden, 'Standardzusammenfassungen' gehören ihrer Art der Präsentation wegen in den Bereich der beschreibenden Statistik. Der Modellcharakter der Standardzusammenfassungen besteht darin, daß viele der Strukturdetails, die etwa im Stengel und Blätter-Schaubild sichtbar sind, unterdrückt werden, während bestimmte Strukturmerkmale komplementär dazu hervorgehoben und präzisiert werden. Beliebt ist in der EDA das 'Spielen', 'Experimentieren' mit Transformationen der Graphiken, um symmetrische Verteilungen zu erhalten, was hilft, den Verteilungstyp einer empirischen Verteilungsfunktion beschreibend zu erkennen. Unterstützt werden diese Maßnahmen durch Verteilungsnetze und Spezialpapiere. Als weitere Untersuchungsmöglichkeit bietet sich beispielsweise, typisch für Zeitreihen, die Glättung, das 'Smoothing', an mit dem Ziel, Zufallsschwankungen zu unterdrücken, aber andererseits Periodizitäten und Strukturbrüche sichtbar zu machen. Professor Bock führte all dies anhand kleinerer Beispiele aus. Danach wies er auf die besondere Rolle der Residuen, etwa bei Regressionen, hin und verdeutlichte auch dies exemplarisch. Die Residuen spielen in der EDA eine wesentlich eigenständige Rolle, so bei der 'Sensitivitätsanalyse'. Die Vorgehensweise bei der 'Hauptkomponentenanalyse' wurde anhand eines beeindruckenden realen Problems in ihren Grundgedanken erläutert (nach B. Flury/H. Riedwyl: Angewandte multivariate Statistik. Computergestützte Analyse mehrdimensionaler Daten, G. Fischer Verlag Stuttgart 1983):

Es handelt sich um eine Erhebung, die an je 100 echten und gefälschten Schweizer Tausend-Franken-Noten durchgeführt worden ist. Gemessen wurden auf einer Seite der Banknote sechs Längenmaße, die die Größe von Druckbild und Papierformat sowie die Lage des Druckbildes auf dem Papier einzufangen vermögen:



Eine grundlegende Methode zur Vereinfachung vieler multivariater Probleme besteht nun darin, statt der gemessenen Variablen geeignete normierte Linearkombinationen derselben zu betrachten. Für den zweidimensionalen Fall kann man die geometrische Anschauung zu Hilfe nehmen. Die folgende Abbildung zeigt den Punktschwarm aller 200 Noten in den Variablen X_4 (= unten) und X_5 (= oben). Entsprechend vergleicht man die beiden Kovarianzmatrizen zweier Variablen und erhält so Kriterien mit maximaler Trennschärfe für die beiden Geldscheingruppen 'echt' und 'falsch'.

Immer wieder im Verlaufe seiner Ausführungen warf Professor Bock einen Seitenblick auf die Schule, verwies gezielt an geeigneten, unproblematischen Stellen auf die Möglichkeiten eines Einsatzes dort und warnte an anderen Stellen vor einem solchen. Seine These/Frage, die letztendlich auch die ganze anschließende Diskussion durchzog, lautete: EDA kann und soll nicht Ersatz für den bisherigen Stochastik-Unterricht sein (auch nicht für den Teilbereich der deskriptiven Statistik). Andererseits wird man ihr nicht ausweichen dürfen/können. Wo also soll sie im Unterricht ihren Platz finden? - Eine allseits befriedigende Antwort auf diese Frage