Ein einfaches Verfahren zur Identifikation von Ausreißern bei multivariaten Daten

Günter Buttler

Friedrich-Alexander-Universität Erlangen-Nürnberg
Wirtschafts- und Sozialwissenschaftliche Fakultät
Lehrstuhl für Statistik und empirische Wirtschaftsforschung
Lange Gasse 20
90403 Nürnberg
Germany

Abstract

Statistical analysis is often disturbed by objects which are extremely different from the rest of the data. Those outliers can be due to different causes. Therefore it is always recommended to examine them separately.

Outliers in one or two-dimensional cases are easily recognized in a frequency distribution. In multidimensional data they can be identified by sub-ordering. It is proposed to do this by calculating pairwise distances. The necessary standardization of the variables can be done by using the sum of all pairwise distances.

Proceeding this way possible outliers can easily be identified in tabular as well as in graphical form. It can also be demonstrated which dimension, that is which variables are contributing to the outlier status. As it is quite simple to remove any object or variable, one can see what is happening to the rest of the data without those extreme values.

1. Einführung

Statistische Analysen werden immer wieder durch das Auftreten von Werten gestört, die erheblich von der Masse der Daten abweichen. Wenn man diesen Ausreißern keine besondere Aufmerksamkeit widmet, beeinflussen sie in den traditionellen statistischen Verfahren die Ergebnisse mehr oder minder stark, d. h. die berechneten Maßzahlen charakterisieren nicht mehr unbedingt die Situation der Datenmehrheit. In der schließenden Statistik werden Schätz- und Testergebnisse verzerrt, die konfirmatorische Statistik gibt Anlaß zu falschen, zumindest aber zu ausgefallenen Hypothesen, und selbst in der deskriptiven Statistik werden unplausible Maßzahlen ermittelt.

Die eleganteste Art und Weise, mit Ausreißern umzugehen, bietet das große Gebiet der robusten Statistik. Die robusten Verfahren sind grundsätzlich so ausgerichtet, daß sie das Gewicht bzw. die Bedeutung ausgefallener Werte mindern, so daß diese das Analyseergebnis nur noch marginal oder gar nicht mehr beeinflussen.

Macht die Möglichkeit, robuste Verfahren zu verwenden, es daher unnötig, sich besonders mit Ausreißern zu beschäftigen? Ich meine nein! Zum einen bedeuten robuste Verfahren stets einen gewissen Informationsverlust. Zum anderen kann die Beschäftigung mit Ausreißern zu wichtigen - neuen - Erkenntnissen über den Prozeß der Datenermittlung oder über die Datenstruktur führen. Beides unterbleibt jedoch, wenn die verwendeten robusten Verfahren es überflüssig erscheinen lassen, sich besonders mit Ausreißern auseinanderzusetzen. Das ist jedoch nur möglich, wenn man vorab Ausreißer identifiziert. Ein einfaches Verfahren hierzu soll im folgenden demonstriert werden. Dabei sollen die Vorteile der rechnerischen Exaktheit mit denen der Anschaulichkeit einer graphischen Darstellung, zeitgemäß im Dialog am PC, kombiniert werden.

2. Begriff

Rönz und Strohe (1994, S. 28) bezeichnen als Ausreißer einen extremen Beobachtungswert, der ein qualitatives, von der Gesamtheit abweichendes Element signalisiert. Weiter heißt es, „Ausreißer können durch Meß-, Übertragungs-, Berichts- oder Rechenfehler verursacht werden. Möglicherweise hat aber auch die Grundgesamtheit eine andere als die angenommene Verteilung, z. B. eine Mischverteilung. Ausreißer können folglich entweder falsche Werte sein oder korrekte Werte, die ihre Ausreißerrolle erst durch eine falsche Modellwahl erhalten. Auf diese zweite, engere Definition beschränkt sich auch Hawkins in der Encyclopedia of Statistical Sciences (1985/6, S. 539).

Barnett und Lewis versuchen, den Ausreißerbegriff anders einzugrenzen (1994, S. 8 ff.). Sie unterscheiden zwischen Extremwerten, Ausreißern und Irrläufern (contaminants).

Extremwerte sind die beiden Randwerte der geordneten Merkmalswerte, also der größte und der kleinste Wert. Ausreißer sind Beobachtungen, die mit dem Rest der Werte unvereinbar erscheinen. Irrläufer stammen aus anderen Verteilungen.

Jede Verteilung hat zwangsläufig Extremwerte. Das können, müssen aber keinesfalls Ausreißer sein. Da es unter Umständen auch mehrere Ausreißer geben kann, die inkompatibel mit der Masse der übrigen Werte sind, gibt es folglich auch Ausreißer, die keine Extremwerte sind. Das ist immer dann der Fall, wenn es eben mindestens einen noch größeren oder kleineren Wert gibt.

Irrläufer schließlich können, müssen aber ebenfalls keine Ausreißer sein. Eine Chance, entdeckt zu werden, haben sie aber nur dann, wenn sie Ausreißer sind. Verstecken sie sich dagegen unter den normalen Werten, bleiben sie unerkannt.

Präzise, weil operational, ist die Definition, die als Ausreißer einen Wert bezeichnet, der mit einer bestimmten Verteilungsannahme inkompatibel ist, wobei die Entscheidung darüber in aller Regel durch einen geeigneten Test erfolgt. Allerdings ist diese Definition zu eng, da sie nichts darüber sagt, was mit einem identifizierten Ausreißer zu geschehen hat. Soll man zu robusteren Modellen übergehen, so lange, bis der Ausreißer modellkompatibel ist? Außerdem

bleibt bei dieser Definition außer acht, daß auch in Gesamtheiten Ausreißer auftreten können.

Zweckmäßig erscheint mir daher die Definition von Barnett und Lewis, als Ausreißer die Werte zu bezeichnen, die mit der Masse der übrigen Werte unvereinbar erscheinen. Mit anderen Worten, es gibt kein objektives Kriterium, anhand dessen eindeutig entschieden wird, ob ein Wert ein Ausreißer ist oder nicht. Es ist vielmehr Ermessenssache. Im Grunde ist ja auch das Verteilungskriterium Ergebnis einer subjektiven Entscheidung, nur daß die Entscheidung nicht den Wert direkt, sondern die Wahl des Verteilungsmodells betrifft.

Wird ein Wert aufgrund subjektiver Entscheidung als Ausreißer identifiziert, stellt sich die Frage, was mit ihm geschehen soll. Die Antwort hängt von einer der folgenden Situationen ab:

1) Die zu untersuchende Gesamtheit ist extrem schief verteilt. Es kommen also auch besonders große - oder kleine - Werte vor.

   a) Die Gesamtheit ist vollständig erfaßt. Die Ausreißer sind Teil der Gesamtheit und müssen folglich unverändert berücksichtigt werden. Als Beispiel mag die Einwohnerzahl deutscher Städte und Gemeinden dienen. Berlin ist mit 3,5 Millionen Einwohnern zweifellos ein Ausreißer. Andererseits ist es aber auch Teil der Gesamtheit.

   b) Wird die Einwohnerzahl über eine Stichprobe erfaßt, so wäre eine Stichprobe, die Berlin enthält, extrem ungünstig, um daraus etwa die durchschnittliche Einwohnerzahl deutscher Städte und Gemeinden zu schätzen. Hier wäre die Anwendung eines robusten Schätzverfahrens am Platze.

2) Versehentlich wird die Einwohnerzahl Nürnbergs mit 4.985.000 registriert, also um den Faktor 10 zu hoch. Hier wäre eine Korrektur des Erfassungs- bzw. Übermittlungsfehlers angebracht. Ähnlich ist die Situation, wenn aus einer Paralleluntersuchung die Stadtregion Paris mit 9 Millionen Einwohnern in die deutsche Analyse geraten wäre. Hier ist eine Elimination des Irrläufers der einzige Weg.

   Wird man in beiden Fällen robuste Verfahren anwenden, wäre der numerische Effekt der beiden fehlerhaften Werte möglicherweise gering und insofern zu vernachlässigen. Man vergibt damit aber die Chance, die Datenqualität zu überprüfen. Es ist ja durchaus möglich, daß derartige Ausreißer nicht lediglich isolierte Einzelfälle sind, sondern sichtbares Resultat einer allgemeinen Nachlässigkeit bei der Behandlung der Daten, die, weil die

Fehler nicht sofort ins Auge fallen, auch nicht erkannt wird. Die Auswirkungen auf die Ergebnisse können nichtsdestoweniger erheblich sein.

3) Auch bei der Analyse von Zufallsvariablen, für die keine konkreten Gesamtheiten existieren, können fehlerhafte Daten auftreten. Ausreißer bei einer Verteilung bedeuten daher nicht automatisch, daß die Verwendung dieser Verteilung falsch ist, also eine besser passende Verteilung zu wählen ist.

Grundsätzlich sollte also gelten, daß vor der Analyse empirischer Daten eine Kontrolle auf Ausreißer erfolgen sollte. Offenbar fehlerhafte Werte sind zu korrigieren oder durch plausible Werte zu ersetzen. Korrekte Ausreißer in Gesamtheiten bleiben unverändert. Sie geben lediglich Anlaß zu der Frage, ob die gewählte Maßzahl wirklich geeignet ist, die Gesamtheit angemessen zu beschreiben. Korrekte Ausreißer in Stichproben legen die Verwendung robusterer Verfahren oder Verteilungen nahe.

3. Wirkung von Ausreißern

Da in der Statistik Einzelwerte stets nur Mittel zum Zweck sind, Aussagen über Gesamtheiten zu erhalten, sind Ausreißer auch nicht per se interessant, weil sie von der Masse der übrigen Werte deutlich abweichen, sondern wegen ihrer Wirkung auf das Analyseergebnis. Sie „verzerren“ die statistischen Resultate, indem sie sie, bei korrekten Ausreißern, als nicht mehr charakteristisch für die Gesamtheit erscheinen lassen, oder schlimmer, bei fehlerhaften Werten, indem sie das Ergebnis verfälschen.

Einen stärkeren Einfluß auf das Ergebnis können jedoch nur Werte metrischer Merkmale haben. Nur dort läßt auch das Skalenniveau das Entstehen von Ausreißern zu. Nur die Abstandsinformation metrischer Merkmale kann dazu führen, daß einzelne Werte sich deutlich von der Masse der Werte unterscheiden. Bei ordinalen Merkmalen können - im Sinne der Einteilung von Barnett und Lewis - zwar Extremwerte auftreten, aber keine Ausreißer. Bei nominalen Merkmalen gibt es nicht einmal Extremwerte, obwohl auch dort, wegen der häufig unterschiedlichen Affinität der Ausprägungen, aus dem Rahmen fallende Daten auftauchen können. Wenn man z. B. die Berufe von Akademikern untersucht, so ist das Auftreten eines Taxifahrers fast schon ein Ausreißer im Vergleich zu den üblichen akademischen Berufen.

Wenn man den Ausreißerbegriff auf metrische Merkmale beschränkt, bedeutet das, daß bei der Verwendung skalenadäquater Methoden die Ergebnisse mehr oder minder stark von Ausreißern beeinflußt werden. Das Ausmaß der Beeinflussung hängt dabei, neben dem Abstand des Ausreißers von den übrigen Werten, auch ab von deren Anzahl. Da Stichproben üblicherweise kleiner sind als Gesamtheiten, spielen Ausreißer folglich in Stichproben auch die größere Rolle.

Bedeutsam ist aber auch die Art und Weise, wie die Abstände in den jeweiligen Methoden gewichtet werden. Werden, wie in vielen Verfahren, etwa bei der Varianz, die Abstände quadriert, wirken sich Ausreißer überproportional aus. Auch die Methode der kleinsten Quadrate in der Regressionsrechnung reagiert besonders empfindlich auf Ausreißer. Das Ausmaß des Einflusses zeigt sich, wenn man die jeweiligen Maßzahlen einmal mit und einmal ohne den Ausreißer berechnet.

Allerdings ist nur bei univariaten Verfahren der Ausreißereffekt auf diese Weise sofort zu erkennen. In multivariaten Verfahren sind Ausreißer so nicht in jedem Falle zu erkennen. Es wird daher zwischen einflußreichen und einflußlosen Ausreißern unterschieden. Während die einflußreichen Ausreißer einen deutlichen Effekt auf das Modell haben, verändern die einflußlosen das Modell und seine Parameter kaum. Ihr Einfluß zeigt sich meist nur mittelbar, man braucht dafür weitere Kenngrößen. Bei einem Regressionsmodell kann dies beispielsweise das Bestimmtheitsmaß sein. Wird z. B. aus einem Modell mit hoher Anpassungsgüte durch Ausschluß eines Wertes ein Modell mit deutlicher niedrigerer Güte, obwohl sich die Parameter kaum ändern, liegt offensichtlich ein einflußloser Ausreißer vor.

Genau genommen sind also auch die einflußlosen Ausreißer eigentlich gar nicht ohne Einfluß, da sie uns einen unter Umständen falschen Eindruck von der Modellqualität verschaffen. Auch scheinbar einflußlose Ausreißer sollten daher gesondert unter die Lupe genommen werden.

4. Identifikation von Ausreißern

Üblicherweise unterscheidet man zwischen graphischen und rechnerischen Methoden. Im mehrdimensionalen Fall ist jedoch eine Trennung nicht mehr exakt möglich, da die einzelnen Merkmale vor einer graphischen Präsentation in aller Regel erst einmal vereinheitlicht, d. h. in bezug auf die Streuung vergleichbar gemacht werden müssen.

Hier soll daher die andere Einteilungsmöglichkeit gewählt werden, nämlich die nach ein- und nach mehrdimensionalen Analysen.

4.1 Ausreißer im eindimensionalen Fall

Graphische Darstellungen sind das einfachste und anschaulichste Instrument für die Identifikation von Ausreißern. Das gilt allerdings nur für einzelne oder - analog - auch für zwei Merkmale gemeinsam.

In erster Linie werden dafür Häufigkeitsverteilungen verwendet, wobei allerdings die Ränder der Verteilung nicht klassiert sein dürfen, um die extremen Werte explizit auszuweisen. Bei diesem Vorgehen ist es Ermessenssache zu entscheiden, ob einer oder mehrere der Randwerte als Ausreißer bezeichnet werden sollen oder nicht.

Bei Vorliegen mehrerer Merkmale kann man selbstverständlich jeweils einfache Häufigkeitsverteilungen zeichnen. Die Beziehung zwischen den Merkmalen geht dabei jedoch verloren. Die Frage, ob einzelne Objekte mehrdimensionale Ausreißer sind, d. h. in bezug auf mehrere Merkmale ausgefallene Werte aufweisen, läßt sich dabei nur indirekt ermitteln, indem man die Randwerte wieder den Objekten zuordnet.

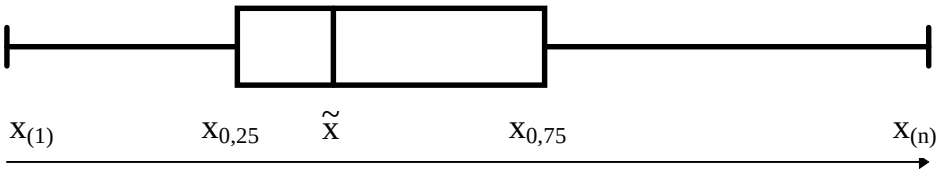
Für zwei Merkmale läßt sich auch eine zweidimensionale Häufigkeitsverteilung erstellen. Bei unterschiedlichen Maßeinheiten bzw. absoluten Streuungen sind die Merkmale jedoch - wie erwähnt - vorab zu normieren. Dies geschieht meist durch Standardisierung, auf die später noch eingegangen wird.

Eine andere eindimensionale Darstellungsmethode sind die sog. Box-Plots, die auch bei uns immer mehr Verbreitung finden (Schlittgen 1993, S. 38 ff.). In

ihnen wird versucht, sowohl die Homogenität der Merkmalswerte als auch deren Spannweite in einem Diagramm zu veranschaulichen.

Zur Charakterisierung der Homogenität können z. B. das 1. und das 3. Quartil verwendet werden. Die 50 % der Merkmalswerte zwischen dem 1. und dem 3. Quartil werden durch einen Kasten (Box) repräsentiert. Die Spannweite wird zusätzlich durch eine rechts und links sich anschließende Gerade verdeutlicht.

Abbildung 1: Box-Plot mit Quartilsabstand und Spannweite



Es bedeuten
$x_{(1)}$ kleinster Wert
$x_{(n)}$ größter Wert
$x_{0,25}$, $x_{0,75}$ 0,25 und 0,75-Quantil
$\tilde{x}$ Median.

Für eine Ausreißerkennzeichnung muß die übliche Darstellung noch etwas modifiziert werden, da zwar die Lage der beiden extremen Werte, des kleinsten und des größten, klar ist, nicht jedoch die Position der benachbarten Werte. Von Ausreißern wird man dann wohl kaum sprechen, wenn sich die Werte unterhalb und oberhalb der beiden angegebenen Quartile ziemlich gleichmäßig über den jeweiligen Wertebereich verteilen.

Es ist daher erforderlich, an den Rändern der Verteilung jeweils mehrere Einzelwerte auszuweisen. Falls die Zahl der Werte zu groß ist, um alle größeren und kleineren Werte einzeln aufzuzeichnen, kann man die Box, je nach Anzahl der Werte, auf alle Werte zwischen dem 1. und 9. Dezil oder dem 1. und 99. Perzentil ausdehnen.

Zur rechnerischen Identifikation von Ausreißern werden meist Testverfahren verwendet. Das setzt allerdings grundsätzlich Zufallscharakter der Daten voraus.

Das fundamentale Problem der Testverfahren ist, daß die zugrundezulegende Verteilung selten bekannt ist, so daß die Ausreißereigenschaft nicht unerheblich davon abhängt, wie gut die tatsächliche - aber unbekannte - Verteilung mit der hypothetischen übereinstimmt. Je kompakter eine Verteilung ist, desto eher werden die Randwerte als Ausreißer identifiziert. Während also die Normalverteilung sehr rasch zu Ausreißern führt, gibt es bei der Cauchy-Verteilung eigentlich gar keine Ausreißer. Auch bei der Pareto-Verteilung kommen häufig extreme Werte vor.

Werden einzelne ausgefallene Werte auf ihre Ausreißereigenschaft getestet, kann sich das Problem der Maskierung (masking) ergeben (Barnett/Lewis S. 109 ff.). Weitere ausgefallene Werte führen dazu, daß der getestete Wert nicht mehr als Ausreißer erkannt wird. Testet man deshalb mehrere Werte gleichzeitig auf Ausreißereigenschaft, kann es passieren, daß normale Werte durch Ausreißer überlagert werden (swamping) und so fälschlich zu Ausreißern deklariert werden.

4.2 Ausreißer im mehrdimensionalen Fall

Bei mehrdimensionalen Analysen wird das Problem der Ausreißeridentifikation schwieriger. Wie Barnett und Lewis (1994, S. 269) betonen, gibt es für multivariate Ausreißer keine so eindeutig klare Fixierung in einer Häufigkeitsverteilung, da mehrdimensionale Verteilungen keine einfachen Ränder besitzen.

Um dennoch eine eindeutige Ermittlung der Randwerte zu erhalten, schlagen sie eine reduzierte Subordnung $R(\underline{x})$ vor, durch die die Merkmalsvektoren $\underline{x}_i$, $i = 1,...,n$ in Skalare umgewandelt werden. Üblicherweise wird dafür die Distanzmessung zum Schwerpunkt $\underline{\bar{x}}$ oder zum Nullpunkt $\underline{0}$ verwendet. Die Addition der Abstände erfordert aber zusätzlich eine Normierung der Merkmalswerte, um vergleichbare Streuungen zu erhalten.

Die wichtigsten Möglichkeiten sind

$R(\underline{x},\underline{W}) = (\underline{x} - \underline{\bar{x}})'\underline{W}^{-1}(\underline{x} - \underline{\bar{x}})$

mit
$\underline{W}$ = Diagonalmatrix mit
(1) s_j Standardabweichung oder
(2) R_j = Spannweite für Merkmal j, j = 1,..., m in der Hauptdiagonalen.

Barnett und Lewis berücksichtigen nicht, daß es inzwischen eine Fülle von graphischen Möglichkeiten gibt, mehrdimensionale Objekte zweidimensional darzustellen, und zwar ohne Informationsverlust.

Der einfache Trick besteht darin, daß jede Merkmalsdimension in der Ebene durch unterschiedliche Zeichen wiedergegeben wird. Da hierfür viele Möglichkeiten denkbar sind, gibt es entsprechend viele potentiellen Darstellungsformen.

Hartung und Elpelt (1989, S. 610 ff.) geben einen Überblick über die wichtigsten Formen. Da gibt es Sterne, Sonnen, Profile, Glyphs, um nur einige zu nennen. Am amüsantesten sind die Flury-Riedwyl-Gesichter, bei denen die Gesichtsteile, Augen, Augenbrauen, Haare usw. die einzelnen Merkmale repräsentieren.

So anschaulich derartige Diagramme - nach einiger Übung - sind, so haben sie den Nachteil, daß sie sich eigentlich nur für kleinere Datenmengen eignen. Es dürfte einem rasch vor den Augen flimmern, wenn man zwecks Ausreißerdiagnose größere Objektzahlen überprüfen sollte. Abhilfe würde hier nur ein Algorithmus schaffen, der vorab potentielle Ausreißer aussortiert. Das wiederum würde aber auf eine reduzierte Subordnung hinauslaufen.

Im folgenden soll eine Möglichkeit vorgestellt werden, mehrdimensionale Objekte sowohl quantitativ als auch graphisch in übersichtlicher Form darzustellen.

4.3 Graphische und zahlenmäßige Darstellung mehrdimensionaler Objekte

Da eine eindeutige Reihung mehrdimensionaler Objekte ohne eine Subordnung nicht möglich ist, wird vorgeschlagen , dies über die paarweisen Abstände der Objekte zu realisieren.

Die Subordnung wird erreicht, indem die Abstände eines jeden Objektes zu allen übrigen aufsummiert werden. Potentielle Ausreißer sind die Objekte mit den größten Abstandssummen. Da auch hier eine Vorab-Normierung der einzelnen Merkmale erforderlich ist, bieten sich die weiter vorn erwähnte Standardisierung oder die Vereinheitlichung der Spannweite an.

Eine andere Möglichkeit der Normierung besteht darin, die paarweisen Abstände eines jeden Merkmals über alle Paarvergleiche aufzusummieren und mit diesem Gesamtabstand die Einzelabstände zu normieren. (Buttler u. Fickel 1995).

Von den verschiedenen Möglichkeiten einer Abstandsmessung soll hier nur die formal einfachste, die quadrierte enklidische Distanz, verwendet werden.

Es ergibt sich für die Gesamtdistanz des Merkmals j

$$D_j = \sum_{i=1}^{n}\sum_{k=1}^{n}(x_{ij} - x_{kj})^2, j = 1,\ldots,m$$

Die Doppelzählung der Abstände ist notwendig, wenn man den Distanzbeitrag eines jeden Objektes für die Subordnung angeben will.

Der normierte Abstand der Objekte i und k bei Merkmal j ergibt sich folglich nach

$$g_{ik,j} = \frac{1}{D_j}(x_{ij} - x_{kj})^2$$

Summiert über alle Paarvergleiche folgt für die relativierte Gesamtdistanz von Merkmal j

$$\sum_{i=1}^{n}\sum_{k=1}^{n} g_{ik,j} = \frac{1}{D_j}\sum_{i=1}^{n}\sum_{k=1}^{n}(x_{ij} - x_{ik})^2 = 1$$

Da dies für alle Merkmale j = 1,...,m gilt, hat die relativierte Gesamtdistanz über alle Paarvergleiche und Merkmale den Wert m.

Für die Subordnung der Objekte wird noch der Distanzanteil der einzelnen Objekte, aufsummiert über die Merkmale, benötigt. Es ist dies

$$g_i = \sum_{j=1}^{m} \frac{1}{D_j} \sum_{k=1}^{n} \left(x_{ij} - x_{kj}\right)^2$$

Die Objekte werden nach der Größe ihrer Distanzanteile geordnet. Potentielle Ausreißer sind solche Objekte, deren Distanzanteil erheblich über den der anderen Objekte hinausgeht.

Es dürfte im übrigen interessant sein, die Verteilung der Gesamtdistanzen zu analysieren. Da die Masse der Objekte normalerweise nahe beieinander liegt, ergibt sich regelmäßig eine rechts-schiefe Verteilung. Das gilt auch ohne Ausreißer. Möglicherweise läßt sich sogar ein Verteilungsmodell finden, das verwendet werden kann, um zu testen, ob einzelne extreme Werte noch zur Verteilung gehören oder nicht, folglich bei Zugrundelegung dieser Verteilung als Ausreißer auszusortieren sind.

Zusätzliche Informationen liefert eine Matrix der paarweisen Distanzen

$$\underline{G} = \begin{bmatrix} 0 & g_{12} & g_{13} & \cdots & g_{1n} \\ g_{21} & 0 & & & \\ \vdots & & \ddots & & \\ g_{n1} & & & & 0 \end{bmatrix}$$

Während die Aufstellung der Subordnung verhindert, daß Überlagerungseffekte auftreten - jedes Objekt wird für sich betrachtet - werden Maskierungseffekte anhand der Matrix $\underline{G}$ sichtbar. Wenn nämlich zwei Objekte i und k deutlich von den übrigen getrennt sind, dies zeigt sich an ihren Distanzwerten g_i und g_k, ist zunächst nicht ersichtlich, ob sie nahe beisammen liegen oder aber ob es sich um isolierte Ausreißer handelt. Wie groß der Abstand zwischen ihnen beiden ist, wird deutlich anhand ihres paarweisen Abstandes g_{ik}.
Da der Abstand der Objekte zu den übrigen sich additiv aus den Distanzen der einzelnen Merkmale ergibt, läßt sich die Gesamtdistanz der einzelnen Objekte auch in die Distanzanteile der einzelnen Merkmale aufteilen. Dies hat den Vorteil, daß bei den potentiellen Ausreißern nicht nur die Gesamtdistanz bekannt ist, sondern auch noch, welche Merkmale besonders dazu beigetragen haben.

Um die Übersichtlichkeit zu verbessern, ist es zweckmäßig, für das Merkmal mit dem größten Gesamtdistanzwert die einzelnen Merkmale nach der Größe ihres Beitrags zu ordnen.

Es ergibt sich dann folgendes Tableau:

Tableau der Gesamtdistanz und der merkmalsspezifischen Distanzen

Objekt	Gesamtdistanz	Merkmalsspezifische Distanzen		
(i)	$g_{(i)}$	$g_{(i),1}$	. . .	$g_{(1),m}$
(1)	$g_{(1)}$	$g_{(1),1}$		
(2)				
(3)				
.	.	.		.
.	.	.		.
.	.	.		.
(n)	$g_{(n)}$	$g_{(n),1}$		$g_{(n),m}$

In der ersten Zeile steht das Objekt mit der größten Gesamtdistanz, dessen Wert in der zweiten Spalte aufgeführt ist.Wegen der größeren Anschaulichkeit sind die Gesamtdistanzen der einzelnen Objekte so normiert, daß ihre Summe nicht m, sondern 1 bzw. 100 % beträgt. Die dritte Spalte besetzt das Merkmal, das bei diesem Objekt die größte Einzeldistanz beigetragen hat. Es folgt das Merkmal mit dem zweitgrößten Beitrag usw.

In der zweiten Zeile ist das Objekt mit der zweitgrößten Gesamtdistanz aufgeführt sowie die Beiträge der einzelnen Merkmale dazu, die jetzt allerdings meist nicht mehr in jedem Fall monoton fallen.
Potentielle Ausreißer sind, wie bereits erwähnt, nur Objekte an der Spitze der Tabelle. Durch welche Merkmale die Ausreißereigenschaft bewirkt wird, wird deutlich anhand der Beiträge der einzelnen Merkmale.

Insbesondere bei einer größeren Zahl von Merkmalen ist es durchaus möglich, daß auch bei Objekten, die in den Gesamtabständen nicht hervorstechen, bei

einzelnen Merkmalen Ausreißer auftreten können. Auch das läßt sich grundsätzlich aus der Tabelle erkennen.

Welche Wirkung die Eliminierung eines Ausreißers auf die verbleibenden Abstände hat, wird deutlich, wenn man das betreffende Objekt streicht und die verbleibenden Distanzen neu berechnet.

Um das Distanztableau nicht zu groß und unübersichtlich werden zu lassen, empfiehlt sich eine Beschränkung auf die Objekte , deren Distanzbeitrag überdurchschnittlich ist. Schwellenwert ist folglich

$$s = \frac{1}{n}$$

Bei dieser Beschreibung ist nicht nur gewährleistet, daß sämtliche potentiellen Ausreißer ausgewiesen werden, auch ausgefallene Werte bei einzelnen Merkmalen dürften in aller Regel sichtbar werden. Da die Verteilung der Distanzbeiträge rechts-schief ist, für die Masse der Objekte ergeben sich nur kleine Werte, wird die Zahl der ausgewiesenen Objekte deutlich kleiner sein als n/2. Selbstverständlich läßt sich auch ein anderer Schwellenwert fixieren, der eine noch stärkere Verringerung der Zahl ausgewiesener Objekte bewirkt.

Das skizzierte Tableau der Objektdistanzen eignet sich bestens für eine graphische Präsentation. Zu diesem Zweck bietet sich ein dreidimensionales Diagramm an, bei dem die Objekte und die Merkmale auf den beiden Basisachsen abgetragen werden, die zugehörigen Distanzbeiträge dagegen auf der Senkrechten.

Das Diagramm verdeutlicht noch besser als das Distanztableau, welche Objekte potentielle Ausreißer sind und welche Merkmale dies bewirkt haben. Aber auch bei den Objekten, die in ihrem Gesamtbild nicht besonders auffallen, fallen einzelne außergewöhnliche Merkmalsbeiträge sofort auf.

4.4 Verfahrensdemonstration an einem Beispiel

Das Verfahren zur quantitativen und graphischen Identifikation von Ausreißern soll an einem kleinen Zahlenbeispiel demonstriert werden. Zu diesem Zweck wurden aus dem Statistischen Jahrbuch deutscher Gemeinden für die

Städte (über 10.000 Einwohner) in Mittelfranken folgende Merkmale entnommen:

1. Fläche (qkm)
2. Bevölkerung
3. Ausländeranteil (%)
4. Ausgaben (1.000 DM)
5. Steuereinnahmen (1.000 DM)
6. Schulden pro Kopf (DM)

Da es sich dabei um einige größenabhängige Merkmale handelt, steht zu vermuten, daß Nürnberg als die mit Abstand größte Stadt ein potentieller Ausreißer ist, auch wenn es sich hier um durchaus korrekte Angaben handelt.

Für die Berechnung der Distanzen und ihre graphische Darstellung hat Jürgen Kroha ein kleines Programm in Turbo-Pascal geschrieben, das auf Diskette beigefügt ist.

Das Programm erlaubt es nicht nur, das Distanztableau sowie das Distanzdiagramm (Auswertung # 1) am Monitor zu betrachten, es ist auch möglich, einzelne Objekte zu eliminieren um zu sehen, welchen Effekt das auf die Abstände der übrigen Objekte hat.

Für die insgesamt 20 Städte sind folgende Daten berücksichtigt:

Tabelle 1: Städte über 10.000 Einwohner in Mittelfranken nach ausgewählten Merkmalen

Nr.	Stadt	Fläche qkm	Bevölkerung Pers.	Ausländeranteil %	Ausgaben 1000 DM	Steuereinnahmen 1000 DM	Schulden pro Kopf DM
1	Altdorf	48,61	13585	6,6	39460	13049	995
2	Ansbach	99,94	37893	5,7	168670	60098	1844
3	Bad Windsheim	78,35	11734	8,5	33626	13632	982
4	Dinkelsbühl	75,12	11226	3,0	37764	11350	1571
5	Erlangen	76,97	102440	12,2	485000	197408	2979
6	Feuchtwangen	139,41	11096	5,3	37616	12429	306
7	Fürth	63,34	103362	13,7	481590	170845	2202
8	Gunzenhausen	82,48	16028	5,1	55353	8540	669
9	Hersbruck	22,92	11961	6,2	43597	15797	494
10	Herzogenaurach	47,73	20464	7,2	55297	26341	1535
11	Hilpoltstein	90,00	10781	4,2	33521	9859	731
12	Höchstadt/Aisch	72,00	11756	7,7	50827	10500	861
13	Lauf	99,83	23390	9,8	80990	34383	1121
14	Neustadt/Aisch	61,22	11484	3,6	34785	15716	1660
15	Nürnberg	185,78	493692	14,2	2773776	1032000	2439
16	Oberasbach	12,21	15871	4,7	37600	18312	1072
17	Röthenbach	13,13	12646	15,3	37458	19028	530
18	Roth	96,25	21737	6,1	61300	23498	239
19	Schwabach	40,71	35514	7,4	132704	48940	1319
20	Weißenburg	97,56	17933	8,0	65118	24974	528

Quelle: Stat. Jahrbuch deutscher Gemeinden 1994

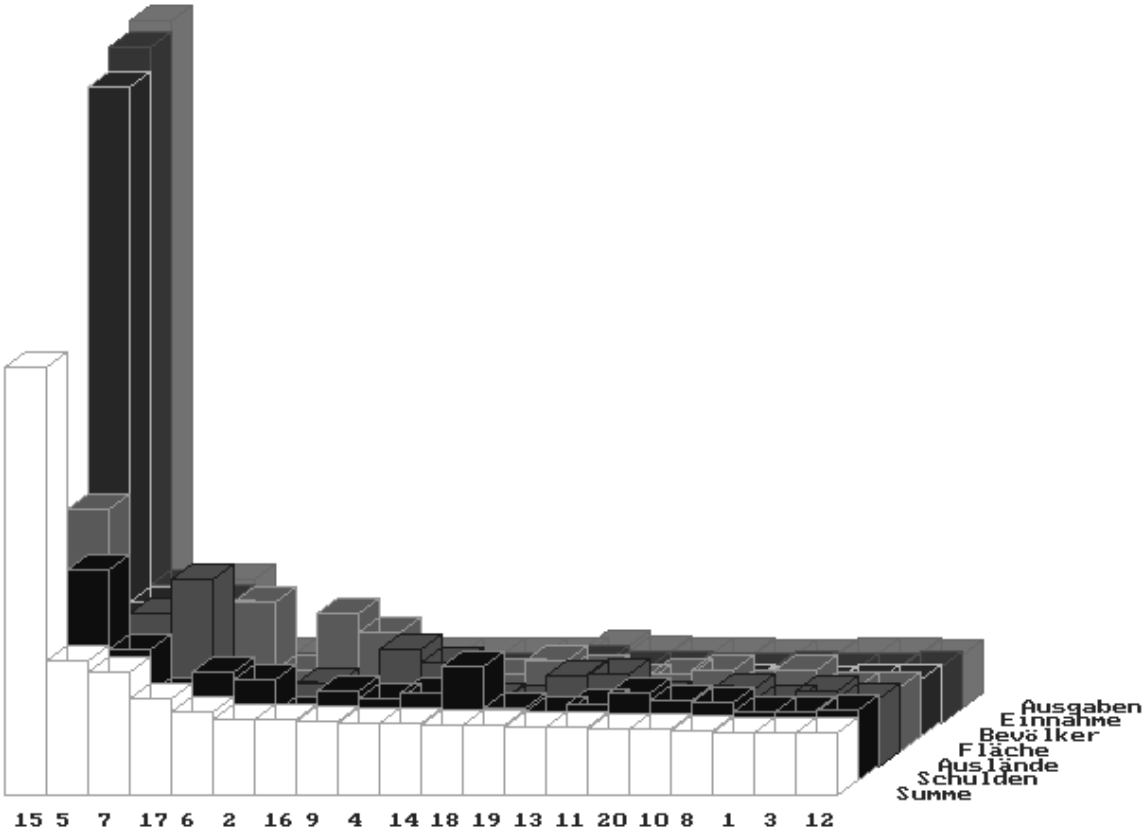
Daraus ergibt sich in einem ersten Analyseschritt folgendes Distanztableau, in dem überdurchschnittliche Werte durch Fettdruck hervorgehoben sind.:

Tabelle 2: Gesamte und merkmalsspezifische Distanzen der mittelfränkischen Städte

Merkmal Objekt	Ausgaben TDM	Einnahmen TDM	Bevölkerung	Fläche 1000qm	Ausländer %	Schulden DM/EW	Summe %
15	**36,28**	**35,71**	**34,31**	**12,77**	**8,63**	**8,06**	**22,63**
5	**6,82**	**7,29**	**7,10**	3,40	**6,73**	**11,09**	**7,07**
7	**6,78**	**6,48**	**7,16**	3,64	**8,11**	**6,87**	**6,51**
17	2,87	2,85	2,93	**7,17**	**9,93**	4,56	**5,05**
6	2,87	2,91	2,99	**7,95**	3,95	**5,65**	4,39
2	3,64	3,55	3,61	4,30	3,74	**5,31**	4,03
16	2,87	2,85	2,90	**7,27**	4,40	3,56	3,98
9	2,85	2,86	2,95	**6,27**	3,56	4,71	3,87
4	2,87	2,93	2,98	3,40	**6,20**	4,34	3,79
14	2,90	2,86	2,97	3,71	**5,49**	4,62	3,76
18	2,86	2,87	2,98	4,06	3,59	**6,03**	3,73
19	3,34	3,29	3,50	4,83	3,53	3,84	3,72
13	2,97	3,02	3,03	4,30	4,84	3,60	3,63
11	2,91	2,97	3,01	3,77	4,86	3,97	3,58
20	2,87	2,88	2,91	4,14	3,73	4,56	3,52
10	2,84	2,90	2,95	4,35	3,51	4,25	3,47
8	2,84	3,02	2,90	3,51	4,08	4,12	3,41
1	2,86	2,90	2,92	4,29	3,51	3,56	3,34
3	2,91	2,89	2,95	3,42	3,99	3,57	3,29
12	2,84	2,96	2,95	3,44	3,61	3,72	3,25

Eindeutiger Spitzenreiter ist Nürnberg (# 15). Bei insgesamt 20 Städten entfällt auf Nürnberg allein ein Distanzanteil von 22,6 %. Die nächste Stadt Erlangen (# 5) bringt es demgegenüber nur noch auf 7 %. Im Distanzdiagramm kommt die Sonderstellung von Nürnberg noch deutlicher zum Ausdruck.

Abbildung 2: Distanzdiagramm der mittelfränkischen Städte



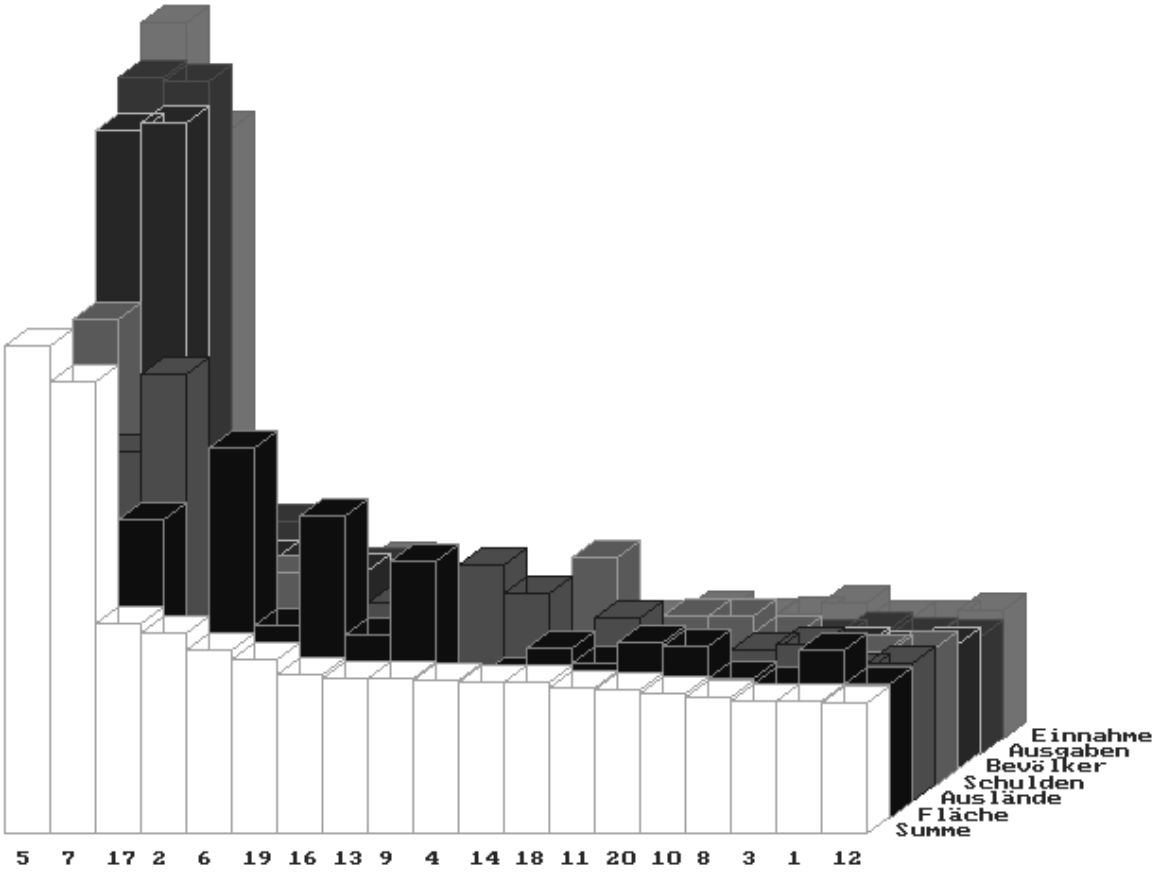
Vom Gesamtbetrag der Distanzen entfällt auf den „Ausreißer“ Nürnberg ein sehr großer Anteil, der allerdings überwiegend durch die größenabhängigen Merkmale, insbesondere Ausgaben, Einnahmen und Bevölkerungszahl hervorgerufen wird. Entsprechend entfallen auf die übrigen Städte nur vergleichsweise kleine Beiträge. Sie erscheinen dadurch recht homogen.

Welchen Effekt die Elimination von Nürnberg hat, wird in der Tabelle 3 sowie in der Abb. 3 deutlich, die Nürnberg nicht enthält.

Tabelle 3: Distanzen mittelfränkischer Städte ohne Nürnberg

Merkmal Objekt	Ausgaben TDM	Einnahmen TDM	Bevölkerung	Fläche 1000qm	Ausländer %	Schulden DM/EW	Summe %
5	**19,97**	**18,87**	**17,81**	**13,02**	**7,98**	3,73	**13,56**
7	**16,98**	**18,72**	**18,00**	**8,10**	**9,76**	3,94	**12,58**
17	3,28	3,34	3,42	4,72	**11,92**	**8,29**	**5,83**
2	**5,98**	**6,49**	**5,90**	**6,10**	3,85	5,12	**5,57**
6	3,43	3,33	3,58	**5,94**	4,07	**10,32**	5,11
19	5,02	**5,27**	**5,52**	4,16	3,73	**5,36**	4,84
16	3,27	3,33	3,35	3,74	4,56	**8,42**	4,45
13	3,96	3,79	3,87	3,79	**5,51**	5,10	4,34
9	3,29	3,28	3,46	4,89	3,67	**7,16**	4,29
4	3,50	3,33	3,55	4,85	**6,60**	3,71	4,26
14	3,29	3,43	3,51	5,21	**5,79**	4,02	4,21
18	3,37	3,35	3,69	**6,37**	3,69	4,76	4,21
11	3,64	3,48	3,64	4,09	5,08	4,33	4,04
20	3,42	3,42	3,43	4,73	4,02	4,87	3,98
10	3,48	3,27	3,58	4,73	3,68	4,76	3,92
8	3,79	3,27	3,36	4,26	4,22	3,92	3,80
3	3,36	3,47	3,48	3,72	4,37	3,76	3,69
1	3,39	3,31	3,38	3,71	3,64	4,70	3,69
12	3,58	3,26	3,48	3,85	3,85	3,74	3,62

Abbildung 3: Distanzdiagramm der mittelfränkischen Städte ohne Nürnberg



Da hierbei die Gesamtdistanz wiederum auf den Wert m normiert wurde, kommen die Unterschiede zwischen den verbleibenden Städten jetzt viel deutlicher zum Ausdruck.

In Auswertung # 2 bringt das Demonstrationsprogramm zusätzlich die Matrix D der paarweisen Distanzen. Auch hier zeigt sich der große Abstand von Nürnberg zu den übrigen Städten, wobei der Abstand zu den beiden anderen Großstädten Erlangen (# 5) und Fürth (# 7) vergleichsweise am geringsten ist.

Nach Eliminierung von Nürnberg erhalten Erlangen und Fürth den größten Abstand zu den übrigen Städten, wobei beide, wie die Graphik zeigt, sehr nahe beieinander liegen.

Die Frage, was mit dem „Ausreißer“ Nürnberg geschehen soll, wurde bewußt nicht beantwortet. Eine automatische Eliminierung kommt nicht in Frage, da es sich ganz offensichtlich um korrekte Werte handelt. Andererseits kann die Berücksichtigung, wie eingangs gesagt wurde, zu „schlechten“ Resultaten führen. Es kommt vielmehr auf den Analysezweck und die vorgesehenen Methoden an, was mit solch einem korrekten Ausreißer gemacht bzw. wie auf seine Existenz reagiert werden soll.

Literatur

Barnett, V., Lewis, T., Outliers in Statistical Data, 3. Aufl., Chichester u. a. 1994.

Buttler, Günter, Fickel, Norman, Clusteranalyse mit gemischtskalierten Merkmalen, Lehrstuhl für Statistik und Ökonometrie, Lehrstuhl für Statistik und empirische Wirtschaftsforschung, Diskussionspapier 2/1995.

Deutscher Städtetag (Hg.), Statistisches Jahrbuch Deutscher Gemeinden, 81. Jg. 1994, Köln.

Hartung, Joachim, Elpelt, Bärbel, Multivariate Statistik, 3. Aufl., München, Wien 1989.

Hawkins, Douglas M., Outliers, in: Kotz, S., Johnson, N. L. (Hg.), Encyclopedia of Statistical Sciences, Bd. 6, New York u. a. 1985, S. 539 - 543.

Rönz, Bernd, Strohe, Hans Gerhard, Lexikon Statistik, Wiesbaden 1994.

Schlittgen, Rainer, Einführung in die Statistik, 4. Aufl., München, Wien 1993.

Programmbeschreibung

Das Programm hat den Namen „STAT.EXE“.
Nach Aufruf erscheint das Hauptmenü mit den Unterpunkten

- Eingabe
- Ändern
- Datei
- Ausgabe
- Fin

Die Menüpunkte können entweder durch die Pfeiltasten ausgewählt und mit der Return-Taste aktiviert oder aber durch die jeweiligen Anfangsbuchstaben direkt aktiviert werden.

Der Menüpunkt **Eingabe** hat als Unterpunkte die Eingabe von Werten, die Anzeige von Eingabewerten und die Freigabe des Speichers.

Eingabe:

Wenn sich beim Start der Eingaberoutine noch Daten im Speicher befinden, wird dies gemeldet und der Benutzer hat die Möglichkeit, diese zu speichern oder aus dem Speicher zu löschen.
Bevor die einzelnen Werte eingegeben werden können, muß die Anzahl der Merkmale und Objekte eingegeben werden. Für jedes einzelne Merkmal wird vor der eigentlichen Werteingabe Name und Einheit abgefragt. Die Werteingabe erfolgt merkmalsweise, die Eingabe eines Wertes ist mit Return abzuschließen, es wird eine Eingabeüberprüfung durchgeführt. Nach jedem Merkmal erfolgt eine Abfrage, ob die Eingabeprozedur abgebrochen werden soll. Alle Eingaben außer ´n´ und ´N´ ermöglichen die Eingabe der weiteren Merkmale. Falls die Eingabe abgebrochen wurde, besteht die Möglichkeit, eine Auswertung mit den bestehenden Daten durchzuführen und diese in eine Datei zu speichern.

Anzeige der Werte:

Alle eingegebenen Ursprungswerte können angezeigt werden. Falls die Wertematrix zu groß für den Bildschirm ist, kann sie mit den Pfeiltasten verschoben werden. Durch die Taste F6 können Sie den Wert, der momentan schwarz hinterlegt ist, ändern. In dem hierfür angezeigten Fenster wird neben dem Namen des Merkmals und der Nummer des ausgewählten Objektes der momentane Wert, der Wertebereich der Merkmalsausprägungen sowie der Mittelwert der Merkmalsausprägungen mit und oh-

ne den ausgewählten Wert angezeigt. Wenn Sie eine Wertänderung durchführen, ist zu beachten, daß bei der darauf folgenden Darstellung der Originalwerte die Reihenfolge der Objekte und Merkmale schon anhand der neuen Daten sortiert sind. Durch das Drücken der ESC-Taste beenden Sie den Anzeigemodus und gelangen wieder in das Untermenü. Falls keine anzuzeigende Werte vorhanden sind, erscheint lediglich eine entsprechende Meldung, die Sie mit Return bestätigen müssen.

Speicherfreigabe:

Hier können Sie die eingegebenen Werte löschen und somit den Speicher wieder freigeben.

Der Menüpunkt **Ändern** enthält die Befehle für das Einfügen bzw. Löschen von Merkmalen/Objekten. Diese Befehle können Sie entweder durch die Pfeiltasten auswählen oder mit den Tasten F3 bis F6 anwählen.

Löschen:

Beim Löschen wird das betreffende Objekt/Merkmal mit den Pfeiltasten ausgewählt, das Löschen wird durch die Taste F10 vorgenommen. Wieder ist zu beachten, daß die verbleibenden Werte bei der neuen Anzeige bereits neu sortiert sind.

Einfügen:

Wenn Sie ein neues Merkmal einfügen wollen, werden zuerst der Merkmalsname und die Einheit abgefragt, die Eingabe der Werte erfolgt wie beim Eingeben einer komplett neuen Matrix. Die Objektreihenfolge ist jedoch von der Sortierung der Werte abhängig und erfolgt nicht automatisch von Objektnummer 1 ... n.

In dem Menüpunkt **Datei** sind die Unterpunkte Speichern, Laden, Dir, Ren und lw-wechseln enthalten.

Speichern:

Sie haben die Möglichkeit, Ihre eingegebenen Werte in einer Datei zu speichern. Falls die Datei, die Sie angegeben haben, bereits existiert, müssen Sie das Überschreiben der bestehenden Datei bestätigen.

Laden:

Sie haben natürlich auch die Möglichkeit, Ihre abgespeicherten Werte wieder zu laden, hierfür müssen Sie den Dateinamen angeben.

Dir:

Mit diesem Befehl haben Sie die Möglichkeit, sich alle Dateien anzeigen zu lassen, die im aktuellen Verzeichnis existieren. Im unteren Bereich des Bildschirmes werden zu der jeweils ausgewählten Datei ergänzende Informationen angezeigt.

Es ist möglich, das Verzeichnis innerhalb eines Laufwerkes mit dieser Option zu wechseln. Hierfür positionieren Sie den Cursor auf das jeweilige Verzeichnis und drücken die Return-Taste. Wenn Sie mit dieser Methode eine Datei auswählen, wird diese beim Laden einer neuen Datei als Vorschlag angegeben, sie ersparen sich damit die Eingabe des Dateinamens.

Ren:

Sie können eine Datei, die im aktuellen Laufwerk ist, umbenennen. Hierfür geben Sie bitte den alten und den neuen Dateinamen an. Mit der ESC-Taste kommen Sie wieder in das Untermenü.

lw-wechseln:

Mit diesem Befehl können Sie das aktuelle Laufwerk wechseln.

In dem Menüpunkt Auswertung können Sie sich die beiden besprochenen Auswertungen anzeigen lassen.

Auswertung 1:

In der Auswertung 1 wird für jede Merkmalsausprägung eines Objektes die Abstandssumme zu den Merkmalsausprägungen der anderen Objekte gebildet und mit der Gesamtabstandssumme ins Verhältnis gesetzt. Die Spalte Summe gibt den Abstandsanteil des jeweiligen Objektes über alle Merkmale an. Die Werte sind so sortiert, daß das Objekt mit der größten Gesamtabstandssumme ganz oben steht. Das Merkmal, daß für dieses Objekt die höchsten Abstandswerte erreicht, steht ganz links. Falls die Werte nur teilweise angezeigt werden, können Sie mit den Pfeiltasten den Ausschnitt verschieben. Durch die ESC-Taste kommen Sie wieder in das Submenü. Mit der F9-Taste können Sie auch die graphische Darstellung der Werte aktivieren, mit der Return-Taste schalten Sie wieder auf die konventionelle Darstellung.

Auswertung 2:

In dieser Auswertung werden die Abstandssummen der einzelnen Objekte zueinander ermittelt. Diese Auswertung hat auch eine Graphikoption, die analog zur Auswertung 1 zu aktivieren ist.

Mit dem Menüpunkt Fin könne Sie das Programm beenden.

Demonstrationsbeispiel

Das Programm „STAT.EXE“ enthält das weiter vorn vorgestellte Demonstrationsbeispiel. Es hat den Namen „DATEN.SAV“ und kann über den Menuepunkt „Datei“ geladen werden.

An dem Demonstrationsbeispiel kann unter anderem gezeigt werden, wie sich die Eliminierung einzelner Objekte oder Merkmale auf die Ergebnisse auswirkt. Die Eliminierung kann über den Menuepunkt „Ändern“ durchgeführt werden.