Clusteranalysen bei Verkehrsuntersuchungen — zum Beispiel bei der Ermittlung von Raumtypen städtischen Unfallgeschehens

VON STEFAN ROMMERSKIRCHEN, BONN

I. Einführung in die Ziele und Anwendungsbereiche der Clusteranalyse

Die Notwendigkeit, eine a priori ungeordnete Vielzahl von Objekten oder Informationen auf einige signifikante Grundtypen zu reduzieren, besteht praktisch in allen Wissenschaftsbereichen. Mit fortschreitender Entwicklung der Datenverarbeitungstechnik beschäftigten sich verschiedene Fachrichtungen wie Wirtschafts- und Sozialwissenschaften, Psychologie, Medizin oder Biologie mit der Entwicklung und Verfeinerung eines Instrumentariums der multivariaten mathematischen Statistik, das man zusammenfassend als Clusteranalyse bezeichnen kann und dessen grundlegende Zielsetzung darin besteht, eine Menge von Elementen unter Verwendung aller relevanten Informationen dergestalt zu Gruppen (Cluster) zusammenzufassen, daß die Elemente einer Gruppe untereinander möglichst ähnlich und die Gruppen untereinander gleichzeitig möglichst verschieden sind.

Dieses Instrumentarium der Clusteranalyse findet auch in den Verkehrswissenschaften vielfache Anwendungsmöglichkeiten. So ermittelten beispielsweise *Krampe, Krieger* und *Tonn* 1975 mittels Clusteranalysen mögliche Einsatzfelder verschiedener neuer Nahverkehrstechnologien[1]. Der Verfasser dieses Beitrags führte 1979 eine Gruppierung verschiedener Städte unter besonderer Berücksichtigung ihrer Nahverkehrsverhältnisse mit der Zielsetzung durch, für weiterführende Analysen eine möglichst repräsentative Auswahl treffen zu können[2]. Unter den oben genannten allgemeinen Zielsetzungen der Clusteranalyse-Technik läßt sich erahnen, daß dieses Instrumentarium in einer quantitativ orientierten Verkehrswissenschaft zahlreiche Einsatzmöglichkeiten besitzt.

Anschrift des Verfassers:
Dr. Stefan Rommerskirchen
Institut für Industrie- und
Verkehrspolitik der Universität Bonn
Adenauerallee 24 – 26
5300 Bonn 1

1) Vgl. *Battelle-Institut e.V., Prognos AG und Studiengesellschaft Nahverkehr mbH,* Gesellschaftliche Beurteilung neuer Nahverkehrssysteme (Vorstudie), Frankfurt, Basel, Hamburg 1975, Teil III.

2) Vgl. *Rommerskirchen, St.,* Sozioökonomische Analyse städtischen Nahverkehrs, Ein empirischer Vergleich des Nahverkehrs im sozioökonomischen Gefüge ausgewählter Städte der Bundesrepublik Deutschland, Berlin 1979.

II. Clusteranalysen bei der Ermittlung von Raumtypen städtischen Unfallgeschehens

1. Die Zielsetzungen der Unfallraumtypenbildung

Zur Veranschaulichung der Beschreibung einiger elementarer Grundzüge der Clusteranalyse wird ein konkretes Beispiel aus dem Bereich des Unfallgeschehens herangezogen: mit Hilfe verschiedener Clusteranalyse-Techniken soll eine leistungsfähige Klassifizierung des städtischen Unfallgeschehens in der Bundesrepublik Deutschland nach „Unfallraumtypen" gefunden werden. Eine derartige Gruppierung kann zum einen einer intensiveren Erforschung des Unfallgeschehens durch die Identifizierung bestimmter Unfallmuster und zum anderen einer Absicherung der Übertragbarkeit von bei Einzelfalluntersuchungen gewonnenen Erkenntnissen dienen. Damit sind zugleich zwei typische Einsatzmöglichkeiten der Clusteranalysen bei Verkehrsuntersuchungen genannt. Als weitere Zielsetzung unseres Beispiels wäre auch eine Schwerpunktsetzung bei der Unfallbekämpfung vorstellbar.

2. Zur Untersuchungsgrundgesamtheit und Datenbasis

Die Grundgesamtheit der Betrachtung des städtischen Unfallgeschehens bilden die 92 kreisfreien Städte der Bundesrepublik Deutschland. Als Basisinformationen stehen folgende Daten zur Verfügung[3]:

a) Straßenverkehrsunfälle insgesamt;
b) darunter Unfälle mit Personenschaden;
c) bei Unfällen verunglückte Personen;
d) darunter Getötete;
e) darunter Schwerverletzte[4].

Da diese absoluten Zahlenangaben wenig geeignet sind, das städtische Unfallgeschehen vergleichbar zu beschreiben, werden zur Ableitung aussagefähigerer Kennziffern (Indikatoren) zwei weitere Merkmale herangezogen:

f) Einwohnerzahl der Stadt;
g) Länge des gesamten Straßennetzes.

3) Die Unfallstatistik stellt auch für kleine Raumeinheiten eine Fülle von Datenmaterial zur Verfügung. Die vorliegende Datenauswahl orientiert sich an der Absicht, das Instrumentarium der Clusteranalyse an einem Beispiel zu veranschaulichen. Das verwendete Datenmaterial entstammt der „Kreisdatenbank Bonn" und gilt im Bereich des Unfallgeschehens für das Jahr 1977, Gebietsstand 1. 5. 1978; vgl. *Rommerskirchen, St.,* Die Kreisdatenbank Bonn, Dokumentation zu einer kleinräumlichen Datensammlung für disaggregierte sozioökonomische Verkehrsuntersuchungen, Bonn 1980.
Zu einer umfassenden Bestandsaufnahme der Determinanten des Unfallgeschehens vgl. *Frerich, J.,* Verkehrssicherheit und Kosten-Nutzen-Analyse, Berlin 1979.

4) Die Daten entstammen der polizeilichen Unfallstatistik und umfassen alle Unfälle mit Personenschaden sowie mit Sachschaden von mehr als 1000 DM bei einem der Beteiligten. Zu den Getöteten zählen auch die innerhalb von 30 Tagen ihren Verletzungen erlegenen Personen. Als Schwerverletzte werden alle Personen betrachtet, bei denen eine stationäre Krankenhausbehandlung erforderlich ist.

Der Bezug des Unfalldatenmaterials auf diese Größen stellt sicherlich nur eine Näherungslösung dar, die in Ermangelung leistungsfähigerer Bezugsgrößen, insbesondere der Verkehrsleistung, erforderlich ist.

Als Ausgangsinformation zur Beschreibung und Typisierung des städtischen Unfallgeschehens werden die folgenden Indikatoren gebildet:
(1) Unfälle je 1000 Einwohner;
(2) Unfälle je Straßenkilometer;
(3) Anteil der Unfälle mit Personenschaden;
(4) Verunglückte je 1000 Einwohner;
(5) Anteil der Getöteten an allen Verunglückten;
(6) Anteil der Schwerverletzten an allen Verunglückten;
(7) Getötete je 1000 Unfälle mit Personenschaden;
(8) Verletzte je 1000 Unfälle mit Personenschaden.

Die Variablen (1) und (2) sind Indikatoren für die relative Unfallhäufigkeit, Variable (3) ist eine Unfallstruktur- bzw. Schadenstrukturkennziffer. Die Variable (4) relativiert die absolute Zahl der Verunglückten. Die Variablen (5) und (6) beschreiben die Struktur der Personenschäden, die Indikatoren (7) und (8) schließlich messen die Verletzungsintensität je Unfall. Eine Niveauvariable – z. B. die absolute Anzahl der Unfälle – soll hier bewußt nicht herangezogen werden, weil sie für die Klassifizierung keinen eigenständigen Informationsgehalt besitzt. Die Berücksichtigung solcher Merkmale ist nur sinnvoll, wenn die Variation der absoluten Merkmalsreihe für das Gruppierungsergebnis von Bedeutung ist. So kann man beispielsweise die absolute Einwohnerzahl einer Stadt bei der Typisierung der Einsatzfelder neuer Nahverkehrstechnologien dazu heranziehen, Größenklassen zu bilden, die die erforderliche Mindestnachfrage nach einem bestimmten Verkehrsmitteltyp gewährleisten. Eine Verwendung absoluter Angaben im vorliegenden Beispiel würde jedoch wegen der unterschiedlichen Größe der betrachteten Räume vermutlich die typischen Strukturen des Unfallgeschehens nur verwischen. Somit bildet eine Datenmatrix mit 736 Elementen (92 Städte, 8 Kennziffern) die Grundlage der Ermittlung von Raumtypen des städtischen Unfallgeschehens mittels Clusteranalysen.

3. *Die Arbeitsschritte der Clusteranalyse*

Bei der Anwendung der Clusteranalyse sind folgende Schritte durchzuführen[5]:
1. Datenaufbereitung (Variablen- bzw. Indikatorenauswahl, Quantifizierung der Merkmale, Bereinigung und Transformation der Daten);
2. Identifizierung und Eliminierung von multivariaten „Ausreißer-Objekten";
3. Clusteranalysen i. e. S. (Auswahl eines Verfahrenstyps und geeigneter Verfahren, Ermittlung der optimalen Gruppenzahl, Festlegung einer Startpartition, Berechnungen);
4. Diskussion der Ergebnisse (Darstellung, Interpretation, Vergleich).

5) Vgl. hierzu und zu einer formal ausführlicheren Beschreibung der Clusteranalyse-Technik *Rommerskirchen, St.*, Die Clusteranalyse als Instrument der Verkehrswissenschaft, in: *Neumann, R., Zachcial, M.* (Hrsg.), Verkehrssysteme im Wandel, Festschrift zum 70. Geburtstag von Fritz Voigt, Berlin 1980, S. 27 ff. Ebenda findet sich auch eine ausführliche Dokumentation weiterführender Literatur.

3.1 *Die Datenaufbereitung*

Der erste Arbeitsschritt wurde mit der Darstellung der Datenbasis der Clusteranalyse bereits angesprochen. Die Quantifizierung der Merkmale ist jedoch nur notwendige, nicht aber hinreichende Voraussetzung für die eigentlichen Berechnungen. Darüber hinaus ist noch zu prüfen, ob der zur Klassifizierung verwendete Merkmalssatz nicht-redundant ist. D. h. es ist zu prüfen, ob die Merkmalsreihen voneinander unabhängige Informationen wiedergeben oder miteinander korreliert sind. Eine Ausschaltung redundanter Informationen ist deswegen erforderlich, weil die Bestimmung der Ähnlichkeiten zwischen den Objekten auf der Grundlage der Distanzen erfolgt, die zwischen den im Merkmalsraum verteilten Objekten bestimmt werden können. Das am häufigsten verwendete Distanzmaß, die Euklidische Distanz, setzt jedoch einen orthogonalen (rechtwinkligen) Variablenraum voraus, der nur bei unkorrelierten Variablen gegeben ist, so daß eine Verwendung korrelierter Merkmale einer Gewichtung in Richtung auf diese Variablen gleichkäme[6].

Die Ausschaltung störender Abhängigkeiten zwischen den Variablen erfolgt auf der Basis der partiellen Korrelationskoeffizienten, die zusammen mit ihren Irrtumswahrscheinlichkeiten (oberhalb der Diagonalen die Korrelationskoeffizienten, unterhalb der Diagonalen die Irrtumswahrscheinlichkeiten für einen zweiseitigen Signifikanztest) in Tabelle 1 wiedergegeben sind.

Tabelle 1: *Korrelationskoeffizienten (r) und ihre Irrtumswahrscheinlichkeiten (α) im Merkmalssatz*

Var. Nr.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	–	−0,025	−0,772	0,730	−0,239	−0,256	−0,236	−0,234
(2)	0,815	–	0,201	0,346	0,292	0,345	0,336	0,441
(3)	0,000	0,055	–	−0,203	0,260	0,246	0,249	0,211
(4)	0,000	0,001	0,052	–	−0,046	−0,078	−0,023	−0,023
(5)	0,022	0,005	0,012	0,667	–	0,139	0,994	0,175
(6)	0,014	0,001	0,018	0,461	0,187	–	0,151	0,978
(7)	0,024	0,001	0,017	0,824	0,000	0,152	–	0,206
(8)	0,025	0,000	0,044	0,826	0,096	0,000	0,049	–

Die Irrtumswahrscheinlichkeit für den Korrelationskoeffizienten gibt an, mit welcher Sicherheit dieser gegen die Behauptung, daß der aus dem Datenmaterial ermittelte Zusammenhang nicht zutrifft (Nullhypothese), abgesichert ist. Da bei der Anwendung der Euklidischen Distanz (statistisch) gesicherte Korrelationen zwischen den Merkmalen

6) Die Möglichkeit, andere Distanzmaße wie die *Mahalanobis*-Distanz oder anstelle korrelierter Variablen deren Hauptkomponenten zur Klassifizierung heranzuziehen, soll an dieser Stelle nicht diskutiert werden, da dieses zu rechentechnischen bzw. Interpretationsproblemen führt, die es im allgemeinen angebrachter erscheinen lassen, die Anwendungsvoraussetzungen für die Euklidische Distanz wenigstens annähernd herzustellen.

vermieden werden müssen, ist eine Grenze festzulegen, oberhalb derer Korrelationen nicht zugelassen werden sollen[7]. Dazu bietet sich die Tabelle der Zufallshöchstwerte des Korrelationskoeffizienten an, die in Abhängigkeit von der Anzahl der Beobachtungen bzw. der daraus resultierenden Freiheitsgrade für bestimmte Irrtumswahrscheinlichkeiten den Schwellenwert für statistisch gesicherte Korrelationen angibt[8]. Will man in unserem Beispiel Abhängigkeiten, die mit 0,1 Prozent gegen die Nullhypothese gesichert sind, ausschließen, so müssen Korrelationskoeffizienten mit ($|r| > 0{,}338$) vermieden werden, für das 1-Prozentniveau mit ($|r| > 0{,}267$).

Da mit abnehmender Anzahl von Beobachtungen der Zufallshöchstwert des Korrelationskoeffizienten ansteigt, kann man zum Ausschluß korrelierter Variablen zusätzlich oder alternativ einen anderen, ebenfalls aus der Korrelationsanalyse abgeleiteten Schwellenwert verwenden: man legt unter sachlogischen Erwägungen einen maximalen Prozentsatz fest, zu dem sich zwei Variable gegenseitig erklären dürfen. Rechnerisch ergibt sich dieser Prozentsatz aus dem Bestimmtheitsmaß, dem Quadrat des Korrelationskoeffizienten. Will man beispielsweise eine gegenseitige Erklärung zweier Variablen von maximal 10 Prozent zulassen, so müssen alle Korrelationen mit ($|r| > 0{,}316$) vermieden werden. Dieser Schwellenwert ist von der Anzahl der Beobachtungen unabhängig und sollte vor allem bei kleineren Grundgesamtheiten (von weniger als 40 Beobachtungen)[9] Verwendung finden.

In unserem Beispiel sollen keine unter dem Ein-Prozent-Niveau abgesicherten Korrelationen zugelassen werden; d. h. bei Auftreten von Korrelationen mit ($|r| > 0{,}267$) ist eine der beiden Variablen auszuschließen. Um möglichst viele Informationen zu bewahren, werden zunächst die Variablen mit den meisten signifikanten Korrelationen ausgeschlossen, in unserem Beispiel zuerst Variable (2) und dann Variable (1). Von den verbleibenden sechs Merkmalen sind noch die Variablen (5) und (7) sowie die Variablen (6) und (8) miteinander korreliert, und zwar als Folge eines straffen Zusammenhangs zwischen den Bezugsgrößen sogar sehr eng, so daß man praktisch von einer definitorischen Verknüpfung der Variablen sprechen kann. Wegen der leichteren Assoziationen bei den Dimensionen der Variablen (5) und (6) (Prozentsätze) wird diesen Variablen der Vorzug gegeben, so daß die Clusteranalysen mit den Variablen (3) bis (6) durchgeführt werden.

Als weiteres Problem des ersten Arbeitsschrittes ist noch zu beachten, daß die Euklidische Distanz – im Gegensatz zum Korrelationskoeffizienten – auch gegenüber linearen Skalentransformationen nicht invariant ist, so daß bei Indikatoren, deren Dimensionsunterschiede für den Gruppierungsprozeß nicht von inhaltlicher Bedeutung sind, eine Normierung vorzunehmen ist. Beispielsweise würde die Verwendung des Indikators „Verunglückte je Stadtbewohner" anstelle von „Verunglückte je 1000 Stadtbewohner" unter Beibehaltung der Dimensionen der übrigen Variablen zu anderen Gruppierungsergebnissen

7) Das Postulat einer völligen Unkorreliertheit der Variablen (mit $r = 0$) läßt sich in der Praxis nicht realisieren. In diesem Zusammenhang ist auch anzumerken, daß die Bedingung ($r = 0$) nicht hinreichend für die Unabhängigkeit zweier Variablen ist. Deshalb sollte man sich auf das Verfahren beschränken, die auf einem bestimmten Signifikanzniveau gesicherten Korrelationen auszuschließen.

8) Vgl. beispielsweise *Förster, E., Rönz, B.*, Methoden der Korrelations- und Regressionsanalyse, Berlin 1979, S. 306.

9) Bei 40 Beobachtungen liegt der Zufallshöchstwert für den Korrelationskoeffizienten mit einer Irrtumswahrscheinlichkeit von 5 Prozent etwa bei $r = \pm 0{,}31$.

führen. Um diesen Einfluß auszuschalten, empfiehlt es sich, die Daten einheitlich auf einen standardisierten Wert mit dem Mittelwert 0 und der Varianz 1 (Z-Transformation) zu normieren, womit zugleich der erste Arbeitsschritt erledigt ist.

3.2 Die Behandlung multivariater Ausreißerobjekte

Der zweite Arbeitsschritt besteht in der Identifizierung multivariater Ausreißerelemente. In unserem Beispiel sind das solche Städte, die unter gleichzeitiger Berücksichtigung aller Gruppierungsmerkmale gegenüber der Grundgesamtheit eine derart atypische Struktur aufweisen, daß sie praktisch mit keiner anderen Stadt vergleichbar und daher auch nicht sinnvoll einem Unfallraumtyp zuzuordnen sind. Solche Ausreißerelemente können den Klassifizierungsprozeß unter Umständen erheblich stören bzw. destabilisieren und müssen daher aus der Grundgesamtheit eliminiert werden.

Da die Ermittlung multivariater Ausreißer auch mit den Methoden der Clusteranalyse erfolgt, ist es an dieser Stelle angebracht, ganz kurz die wesentlichen Verfahrenstypen dieses Instrumentariums zu skizzieren. Generell lassen sich hierarchische und nicht-hierarchische Verfahren unterscheiden. Bei den (agglomerativ-) hierarchischen Verfahren[10] wird zunächst jedes Element (jede Stadt) als eine Gruppe betrachtet. Diese Gruppen werden dann – nach unterschiedlichen Methoden und Vorschriften für die Distanz- bzw. Ähnlichkeitsmessung – schrittweise zusammengefaßt (fusioniert), bis alle Elemente eine einzige Gruppe bilden. Charakteristisch für die hierarchischen Verfahren ist, daß die auf einer bestimmten Stufe gefundenen Gruppierungen im nachfolgenden Schritt nicht wieder aufgelöst werden können. Der Gruppierungsprozeß kann mittels eines „Dendrogramms" graphisch veranschaulicht werden (vgl. Abb. 1 – 3), das eine nützliche Hilfe zur Beurteilung der Gruppierungen ist und einen der wesentlichen Gründe darstellt, zu Beginn der Clusteranalysen hierarchische Verfahren einzusetzen.

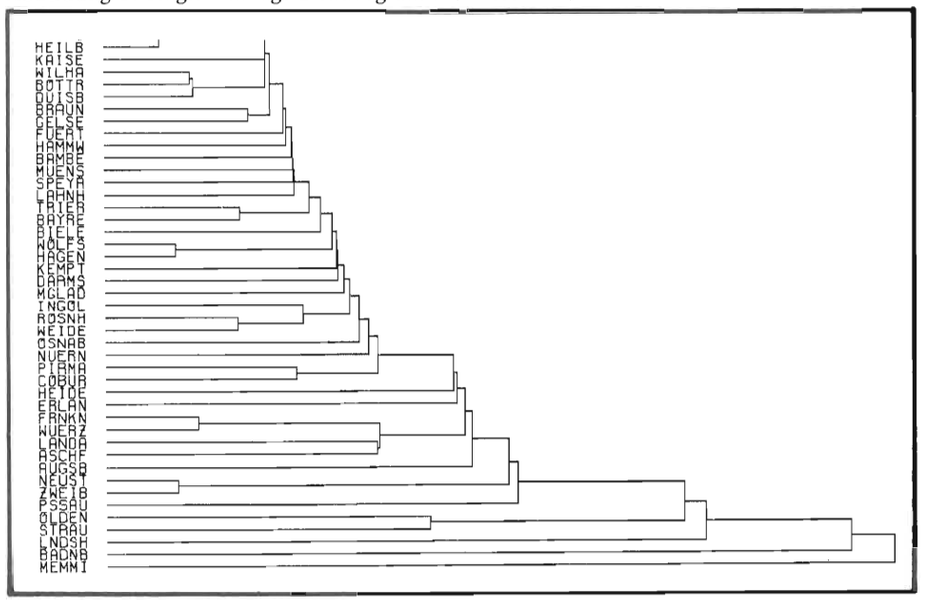
Die nicht-hierarchischen Verfahren zielen darauf ab, eine beliebige Anfangsgruppierung (Startpartition) solange iterativ zu verbessern, bis ein vorgegebenes Gütekriterium für das Typisierungsergebnis nicht mehr verbessert werden kann. Diese iterativen Algorithmen sind vor allem deswegen schwieriger anzuwenden, weil sie – sofern man die Berechnungen nicht für alle möglichen Gruppenzahlen durchführen möchte – voraussetzen, daß man eine Teilmenge von Gruppenzahlen fixiert, für die jeweils die Berechnungen vorgenommen werden, und außerdem und insbesondere deshalb, weil man eine Erstaufteilung der Elemente auf die jeweils betrachtete Gruppenzahl finden muß, mit der der iterative Optimierungsprozeß gestartet werden kann. Da die Anzahl der möglichen Startpartitionen schon bei kleinen Gruppenzahlen und Objektmengen sehr groß ist, sind die gefundenen optimalen Typisierungsergebnisse oftmals nur lokal stabil; d. h. unter Verwendung anderer Anfangsaufteilungen lassen sich möglicherweise andere optimale Gruppierungen finden.

10) Die divisiven hierarchischen Verfahren gehen genau umgekehrt vor wie die agglomerativen: sie betrachten zunächst die Grundgesamtheit der Elemente als eine Gruppe und spalten diese solange auf, bis jedes Objekt ein eigenes Cluster darstellt. Die Algorithmen zur Lösung dieses Problems sind in der Regel rechenaufwendiger als agglomerative, so daß sie praktisch kaum zur Anwendung gelangen.

Die Ausreißeranalyse erfolgt mittels hierarchischer Clusteranalysealgorithmen. Dabei ist es zweckmäßig, sowohl das „Single-Linkage-" als auch das „Group-Average-Verfahren" anzuwenden[11], die schwer einzuordnende Elemente erst am Ende des Fusionsprozesses eingruppieren, so daß man Ausreißer relativ leicht mittels der Dendrogramme herausfinden kann. Die Abbildungen 1 und 2 zeigen die unteren Hälften des Single-Linkage- und Group-Average-Dendrogramms bei Gruppierung aller 92 Städte auf der Basis der vier durch Z-Transformation standardisierten Unfallmerkmale. Mit Single-Linkage werden als letztes die Städte Memmingen, Baden-Baden und Landshut eingruppiert, bei Group-Average findet sich am unteren Ende des Dendrogramms ein fünfelementiges Cluster mit den Städten Memmingen, Landshut, Passau, Coburg und Pirmasens, außerdem noch Baden-Baden als schwer einzuordnendes Einzelelement.

Um die Entscheidung über die Auswahl der Ausreißerobjekte abzusichern, wurde für die komplette Grundgesamtheit auch eine Gruppierung nach dem WARD-Verfahren, dem wohl leistungsfähigsten hierarchischen Klassifikationsverfahren, durchgeführt, das aufgrund seiner Vorgehensweise allerdings zur Identifizierung multivariater Ausreißer nicht geeignet ist. Dabei ergab sich, daß schon bei der Unterscheidung von nur drei Gruppen ein sechselementiges Cluster separiert wird, das neben den Städten Memmingen, Landshut, Passau, Coburg und Pirmasens noch Augsburg enthält. Somit stehen insgesamt sieben Städte als mögliche Ausreißer (im Hinblick auf die Unfallraumtypisierung) zur Disposition.

Abbildung 1: *Single-Linkage-Dendrogramm (Ausschnitt)*



11) Vgl. hierzu beispielsweise *Sitterberg, G.*, Multivariate Analyse der Struktur und Entwicklung von Städten, Münster 1977, S. 120 ff.

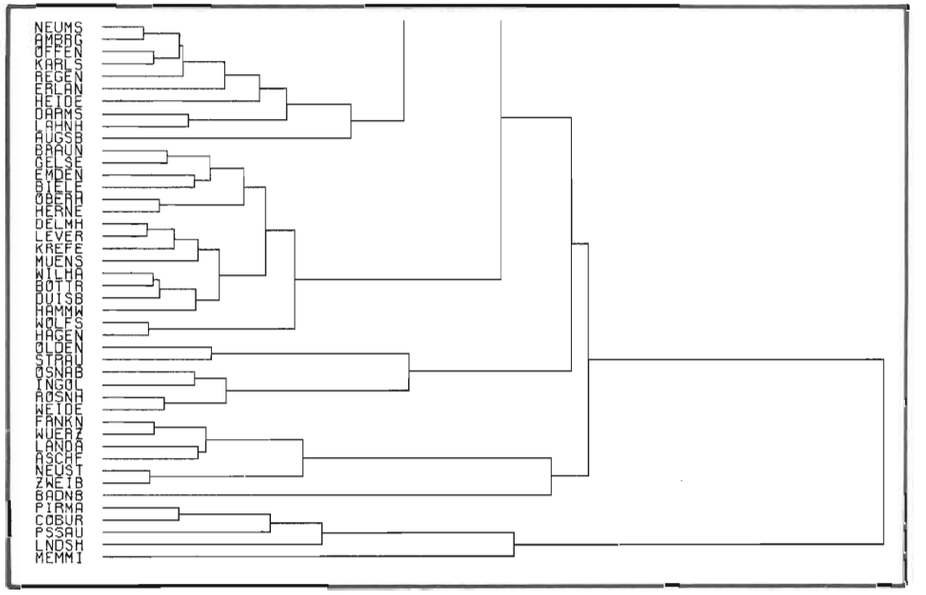
Um Klarheit darüber zu erhalten, ob nur drei Städte (Memmingen, Baden-Baden, Landshut) oder alle sieben bei der eigentlichen Typisierung ausgeschlossen werden müssen, wurde die Grundgesamtheit zunächst nur um die drei genannten Städte verringert und nochmals eine Ausreißeranalyse durchgeführt. Nach Single-Linkage zeigten sich wiederum Passau und Augsburg als Ausreißer. Bei Group-Average bildeten diese Städte zusammen mit Coburg und Pirmasens eine (Ausreißer-)Gruppe. Daher wurden schließlich alle sieben Städte aus der eigentlichen Gruppierung ausgeschlossen, so daß 85 Städte die Grundgesamtheit der weiteren Berechnung bilden.

3.3 Die Durchführung der Klassifizierung

Der dritte Arbeitsschritt, der in der Suche nach der optimalen Zuordnung der verbleibenden Städte zu signifikant unterschiedlichen Unfallraumtypen besteht, wird üblicherweise mit einer hierarchischen Gruppierung nach dem WARD-Verfahren begonnen. Dieses bereits kurz erwähnte Verfahren zielt darauf ab, die bei der Fusionierung zweier Cluster auftretenden Varianzzuwächse innerhalb der neuen Cluster zu minimieren. Als Fusionskriterium dient also im Gegensatz zu den meisten anderen hierarchischen Verfahren nicht ein minimaler Distanzzuwachs, sondern ein möglichst geringer Homogenitätsverlust, gemessen durch ein Heterogenitätsmaß (Fehlerquadratzuwächse).

Das Klassifizierungsergebnis nach dem WARD-Algorithmus zeigt Abbildung 3. Es verbleibt noch die Aufgabe, die Gruppenanzahl zu ermitteln, die im Hinblick auf die angestrebte Differenzierung der Unfallraumtypen optimal erscheint. Als Entscheidungskrite-

Abbildung 2: *Group-Average-Dendrogramm (Ausschnitt)*



rium kann man die relativen Distanz- bzw. Heterogenitätszuwächse heranziehen, die bei den einzelnen Fusionsschritten auftreten. Man wählt dann diejenige Gruppenzahl, nach der durch den Fusionsprozeß besonders hohe Zuwachsraten zu verzeichnen sind. Demnach ist es im vorliegenden Fall z. B. sinnvoll, entweder eine vier oder eine acht Klassen umfassende Lösung zu betrachten, wobei die 4-Cluster-Lösung im Sinne des genannten Kriteriums überlegen ist. Dieses Kriterium stellt allerdings lediglich eine Entscheidungshilfe dar. Es ist durchaus möglich und zulässig, andere Gruppenzahlen zu betrachten, wenn gewisse Vorstellungen über die in etwa zu betrachtende Gruppenzahl vorliegen, zumal die Funktion der Distanz- oder Homogenitätszuwächse in Abhängigkeit von der Gruppenzahl zumeist mehrere relative Minima aufweist. Exogene Vorgaben bestehen beispielsweise häufig bei der Auswahl von repräsentativen Untersuchungsobjekten, wenn durch den Zeit- und/oder Finanzrahmen eine ungefähre Zahl zu betrachtender Objekte fixiert worden ist.

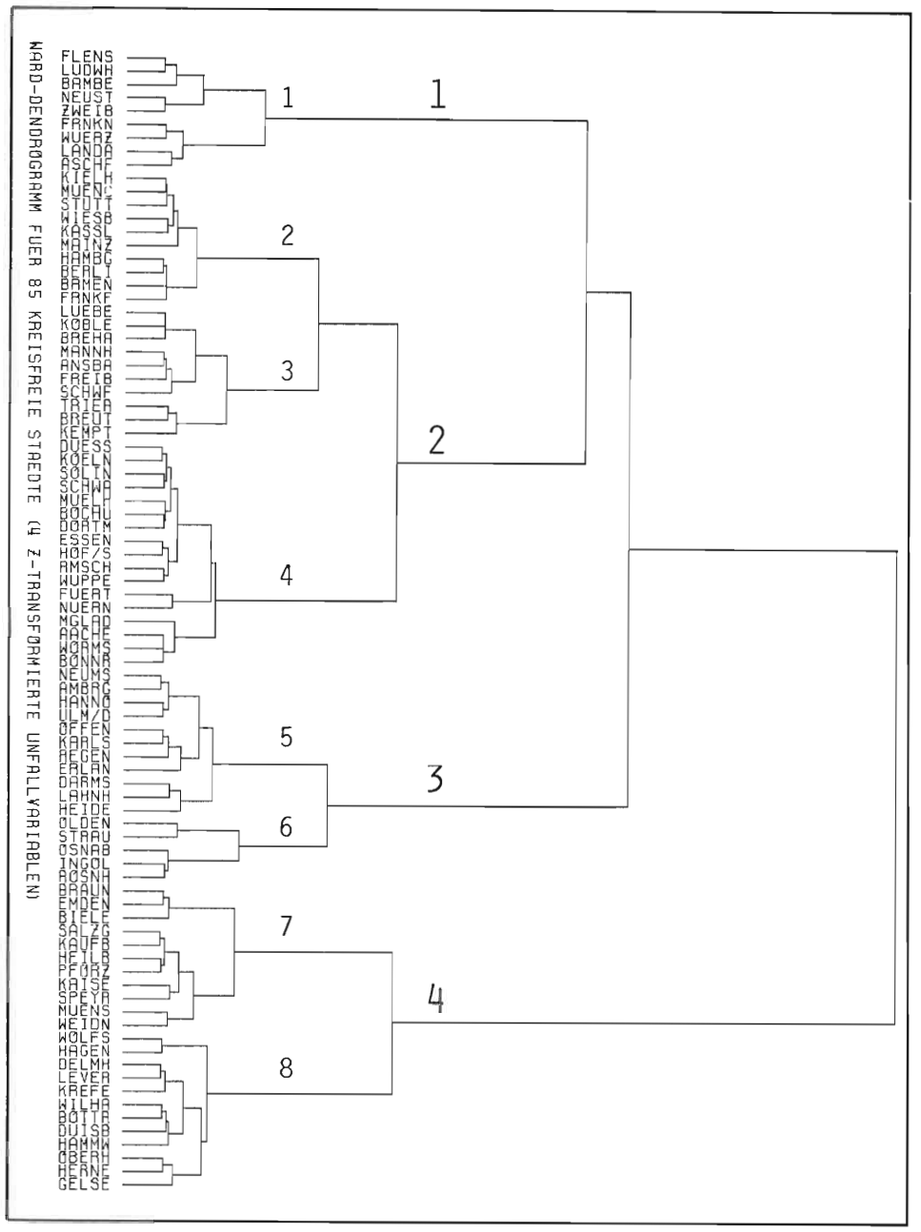
Bei der Anwendung iterativer Clusteranalysealgorithmen sollte man die Informationen, die bei der hierarchischen Klassifizierung gewonnen werden konnten, möglichst weiter nutzen. Das gilt sowohl im Hinblick auf die Fixierung der Gruppenzahl als auch hinsichtlich der Festlegung der Startpartition. Im vorliegenden Beispiel wurde daher versucht, die mit dem WARD-Verfahren gefundenen 4- und 8-Cluster-Lösungen mit dem am häufigsten eingesetzten iterativen Algorithmus KMEANS zu verbessern. Der KMEANS-Algorithmus zielt darauf ab, eine Startpartition durch Verschieben der Objekte zwischen den Gruppen solange zu verändern, bis die Abstandsquadratsummen innerhalb aller Gruppen minimal sind. Verfügt man nicht über eine Anfangsaufteilung der Elemente, so startet man KMEANS üblicherweise mit der „Standardanfangspartition", bei der die einzelnen Objekte der Reihe nach auf die fortlaufenden Gruppen verteilt werden.

Die Verbesserung der Standardanfangspartition mit KMEANS führte bei unserem Beispiel im Sinne des verwendeten Zielkriteriums sowohl bei vier als auch bei acht Gruppen zu schlechteren Ergebnissen als die mit den WARD-Gruppierungen gestarteten Iterationen. Verbessert man die beiden WARD-Lösungen mit KMEANS, so kommt es in beiden Fällen zu 12 Verschiebungen. Für die 4-Cluster-Lösung – das mag an dieser Stelle genügen – ergeben sich folgende Verschiebungen (die Cluster-Nummern sind dem Dendrogramm (Abb. 3) zu entnehmen): Flensburg und Ludwigshafen wechseln von Cluster 1 zu Cluster 2, Freiburg von Cluster 2 zu Cluster 3; Trier, Kempten (Allgäu), Mülheim und Essen werden von Cluster 2, Straubing von Cluster 3 nach Cluster 4 verschoben, und aus Cluster 4 wechseln Heilbronn, Kaiserslautern, Speyer und Weiden zu Cluster 3. Diese 4-Cluster-Lösung nach KMEANS soll im letzten Arbeitsschritt noch etwas näher beleuchtet werden.

3.4 Zur Interpretation der Ergebnisse

Den vierten und letzten Arbeitsschritt bilden die Erläuterungen zu den Ergebnissen der Clusteranalysen. Dazu zählt nicht nur die Beschreibung der einzelnen Cluster im Hinblick auf die Typisierungszielsetzung, sondern auch eine Interpretation der Gruppierungsunterschiede bei Anwendung verschiedener Verfahren bzw. Verfahrenstypen sowie eine Diskussion der Aussagegrenzen der verwendeten Informationen (Daten) und der eingesetzten Methoden. Insofern kommt diesem Arbeitsschritt der Ergebnisdiskussion eine

Abbildung 3: *WARD-Dendrogramm zur Klassifizierung von Unfallraumtypen*



sehr große Bedeutung zu, wie ja ganz generell der Wert statistischer Resultate davon abhängt, in welchem Maße ihr Zustandekommen nachprüfbar dokumentiert und die Gültigkeitsbereiche kritisch offengelegt werden. Dies gilt um so mehr, wenn dem Anwender – wie beim Instrumentarium der Clusteranalyse – in jeder Arbeitsphase Ermessensspielräume verbleiben, sei es bei der Auswahl der Variablen, bei der Transformation der Daten, bei der Fixierung der Gruppenzahl oder bei der Entscheidung über sonstige offene Fragen. Die Verpflichtung zu einer sorgfältigen Dokumentation und Begründung der angewendeten Verfahren und ausgewählten Ergebnisse ist bei Clusteranalysen zudem auch dadurch begründet, daß in der Regel – wie bereits erwähnt – keine eindeutigen und global stabilen Ergebnisse erzielt werden können.

Zur Beschreibung der Clusteranalyseergebnisse stehen verschiedene Hilfsgrößen zur Verfügung. Zur Erklärung der Besonderheit eines Clusters gegenüber der Grundgesamtheit in Bezug auf die einzelnen Klassifikationsvariablen bildet man die Differenz zwischen den Mittelwerten einer Variablen im Cluster und in der Grundgesamtheit und normiert diese Differenz auf die Standardabweichung der Variablen in der Grundgesamtheit. Vergleichsweise hohe Werte für diese Prüfgröße weisen auf diejenigen Variablen hin, hinsichtlich derer sich das Cluster besonders deutlich von der Grundgesamtheit unterscheidet. In Tabelle 2 sind Mittelwert[12] und Standardabweichung für die vier Klassifikationsvariablen in der Grundgesamtheit und in den vier Gruppen, die bei der iterativen Verbesserung des WARD-Ergebnisses durch KMEANS entstanden, aufgeführt.

Tabelle 2: *Statistische Angaben zur Beurteilung der 4-Cluster-Lösung nach KMEANS*

Statistische Kennziffer	Variable	Grundgesamtheit	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Mittelwert	(3)	52,2793	45,3112	47,7191	53,2551	59,9597
	(4)	8,4852	10,3876	7,8918	9,9090	7,5848
	(5)	1,7961	0,9508	1,6147	2,0631	2,0770
	(6)	24,5883	28,7430	22,6090	21,1016	29,0863
Standardabweichung	(3)	8,7710	3,7219	7,1197	7,7472	6,8648
	(4)	1,4197	1,1693	0,8493	0,9830	0,9687
	(5)	0,4913	0,3884	0,3266	0,3885	0,3922
	(6)	4,8177	3,3418	2,9277	4,1481	3,5493

In Cluster 1 ist die deutlichste „Unterscheidungsvariable" die Variable (5), in Cluster 2 die Variable (3), in Cluster 3 die Variable (4) und in Cluster 4 die Variable (6). Es sei aber nochmals in Erinnerung gerufen, daß für das Zustandekommen der einzelnen Cluster alle Variablen verantwortlich sind. Die genannten Einzelangaben dienen nur der Interpretation der Ergebnisse.

12) Die angegebenen Mittelwerte sind die arithmetischen Mittel der Kennziffern und nicht die „wahren" Durchschnittswerte für die Grundgesamtheit bzw. die einzelnen Cluster.

Zur Beurteilung der Frage, welche Variablen innerhalb des Clusters besonders wenig variieren und damit einen hohen Erklärungsgehalt für die Zusammengehörigkeit der Objekte im Cluster aufweisen, berechnet man das Verhältnis der Variablenvarianzen[13] innerhalb der Cluster zu denen in der Grundgesamtheit. Niedrige Werte für diesen Quotienten deuten auf vergleichsweise geringe Variablenvariationen innerhalb der Cluster hin und sind daher zu deren Charakterisierung gut geeignet. Cluster 1 erreicht beispielsweise hinsichtlich der Variablen (3) und (6) die geringsten Werte und könnte somit als das Cluster mit einer unterdurchschnittlichen Personenschadensquote bei gleichzeitig hoher Verletzungsintensität bezeichnet werden. Die übrigen Cluster kann man nach dem gleichen Muster charakterisieren.

Will man die Homogenität eines Clusters unter Berücksichtigung aller Variablen ermitteln, so bildet man den Durchschnitt der Distanzen aller Clusterelemente vom Clustermittelpunkt (Centroid). Ist dieser „Radius"[14] im Verhältnis zu den Radien anderer Cluster klein, so kann man das Cluster als homogen bezeichnen. Für alle genannten Hilfsgrößen zur Interpretation der Clusteranalyseergebnisse gilt, daß ihre absoluten Werte keinen eigenständigen Informationsgehalt besitzen, sondern daß sie erst durch den Vergleich mit den Werten für die anderen Variablen bzw. Cluster Bedeutung erlangen.

III. Schlußbemerkungen

Mit diesen Ausführungen soll das Beispiel zur Demonstration des Einsatzes der Clusteranalyse-Technik bei Verkehrsuntersuchungen abgeschlossen sein. Eine weiterführende Beschreibung der einzelnen Unfallraumtypen – beispielsweise unter sozioökonomischen Aspekten – würde die Zielsetzungen des vorliegenden Beitrags übersteigen. Das Beispiel sollte insbesondere verdeutlichen, welche Probleme bei der Anwendung von Clusteranalysen auftreten und welche Lösungsansätze zu ihrer Bewältigung bestehen. Zugleich sollten Einsatzmöglichkeiten und -grenzen dieses Instrumentariums aufgezeigt werden. Wie bei anderen Verfahren der multivariaten Statistik verbleibt auch dem Anwender der Clusteranalyse ein gewisser Freiraum für Subjektivität und Intuition. Dennoch erscheint das dargestellte Verfahren der multidimensionalen Klassifikation rein gefühlsmäßigen und/oder monothetischen (d. h. auf einem einzigen Kriterium beruhenden) Verfahren eindeutig überlegen. Insofern stellt es eine wertvolle Bereicherung einer quantitativ orientierten Verkehrswissenschaft dar.

13) Die Varianz errechnet sich als Quadrat der Standardabweichung.

14) Die Durchschnittsdistanzen innerhalb der Cluster werden bei dem vom Verfasser für die Beispielsberechnungen eingesetzten und sehr empfehlenswerten Software-Programmpaket für Clusteranalysen „CLUSTAN 1.C" (Release 2) nach *David Wishart* leider nicht angegeben. FORTRAN-Programme zur Interpretation von Clusteranalyseergebnissen finden sich beispielsweise bei *Anderberg, M. R.*, Cluster analysis for applications, New York, London 1973, S. 326 ff.

Summary

In traffic science, as in many other scientific fields, we often face the problem of classifying a great number of a *priori* unorganized elements with due and simultaneous regard to several aspects. The cluster analysis has been found to be an appropriate instrument to solve this problem in mathematical and statistical terms. In this paper, the basic characteristics of this technique and the problems of application involved are shown by means of determining and classifying specific types of urban accidents. A classification of this nature may aim at determining high-priority countermeasures, examining the applicability of results from individual studies or finding representatives for further studies. The paper demonstrates that the cluster analysis, as an instrument of multi-dimensional classification, is of great value to a quantitatively oriented traffic science.

Résumé

Comme dans d'autres spécialités, on trouve dans la science traitant le trafic souvent le problème de devoir grouper un nombre d'éléments a priori sans ordre en tenamt compte de plusieurs aspects simultanément. Un instrument approprié à résoudre ce problème à l'aide de méthodes mathématiques et statistiques est l'analyse Cluster. Le présent exposé montre les caractéristiques de cette méthode et les problèmes d'application de la technique de l'analyse Cluster à l'aide de la détermination de types d'accidents urbaines. Le but d'une telle classification est de par exemple déterminer les priorités dans la lutte contre les accidents, de vérifier l'applicabilité des résultats de recherches individuelles ou de trouver des représentants pour d'autres recherches. Cette étude montre que l'analyse Cluster est, comme intrument de classification multidimensionnelle, un enrichissement précieux pour la science traitant le trafic qui se base sur la quantité.

b.va.

Die Leistung des Verkehrsbetriebes — Bemerkungen zu einer Untersuchung von Thies Claussen [1]

VON SÖNKE PETERS, BERLIN

1. Vorbemerkungen

In seiner Arbeit unternimmt *Claussen* den Versuch, die „Produkte", d. h. die Leistungen des Verkehrsbetriebes einer Analyse zu unterziehen, weil er in ihnen zutreffend den Ausgangspunkt einer betriebswirtschaftlichen Theorie des Verkehrsbetriebes erblickt. Dabei bezeichnet er die Leistung eines Verkehrsbetriebes als eine komplexe Erscheinung, die unter mehreren Aspekten untersucht werden muß, um sie gedanklich erfassen und darstellen zu können.[2] Bei dieser Untersuchung geht es *Claussen* um die Systematisierung der verkehrsbetrieblichen Produktionsfaktoren, die Behandlung der Teilfunktionen und -prozesse der Verkehrsleistung sowie die Analyse der Verkehrsleistung als Leistungsergebnis und Leistungsprozeß, als Marktleistung und Betriebsleistung und als Gelegenheits- und Linienverkehrsleistung.[3]

Dieser Betrachtungsweise und ihren Ergebnissen kann nicht gefolgt werden, allerdings nur wegen der Unzweckmäßigkeit des Ansatzes, nicht jedoch, weil dieser für falsch gehalten wird, da Definitionen keiner Richtig-Falsch-Entscheidung unterliegen können. Die Haupteinwände gegen *Claussens* Untersuchung sind darin zu sehen, daß er nicht versucht, die Verkehrsleistung und darauf aufbauend die Verkehrsbetriebslehre im Rahmen der Allgemeinen Betriebswirtschaftslehre zu betrachten, sondern daß er vielmehr Gefahr läuft, Besonderheiten der Verkehrsleistung und damit auch des Verkehrsbetriebes herauszuarbeiten, die letztlich keine sind[4], anstatt auf Gemeinsamkeiten mit anderen Betrieben oder Gruppen von Betrieben abzustellen. Hierzu kann aber durchaus die Tatsache geführt haben, daß Aussagen der Allgemeinen Betriebswirtschaftslehre herangezogen wurden, die tatsächlich nicht so allgemein sind, daß sie für alle (denkbaren oder real existierenden) Betriebe Gültigkeit besitzen, da sich die Allgemeine Betriebswirtschaftslehre in ihrer historischen Entwicklung mehr oder weniger stark an den Gegebenheiten in für den anonymen Markt produzierenden Industriebetrieben orientiert hat.[5]

Anschrift des Verfassers:
Prof. Dr. Sönke Peters
Technische Universität Berlin
Institut für Betriebswirtschaftslehre
Fachbereich Wirtschaftswissenschaften (FB 18)
Uhlandstraße 4 – 5
1000 Berlin 12

1) *Claussen, T.*, Zur Diskussion des „Verkehrsleistungs"-Begriffs, in: Zeitschrift für Verkehrswissenschaft, 50. Jg. (1979), S. 245 ff.
2) Vgl. *Claussen, T.*, a.a.O., S. 245.
3) Vgl. ebenda.
4) Vgl. *Diederich, H.*, Die allgemeine Betriebswirtschaftslehre als Grundlage der Betriebswirtschaftslehre des Verkehrs, in: Gegenwartsfragen der Verkehrsbetriebslehre, Köln 1975, S. 160 f.
5) Vgl. zu dieser Problematik ebenda sowie *Kirsch, W./Bamberger, I./Gabele, E./Klein, H. K.*, Betriebswirtschaftliche Logistik, Wiesbaden 1973, S. 11 f.