Klassifikation mit Clusteranalyse: grundlegende Techniken hierarchischer und K-means-Verfahren

Wiedenbeck, Michael; Züll, Cornelia

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Klassifikation mit Clusteranalyse: Grundlegende Techniken hierarchischer und K-means-Verfahren

Michael Wiedenbeck & Cornelia Züll
Zentrum für Umfragen, Methoden und Analysen, Mannheim

Zusammenfassung

Nach einer Einführung in die Ziele der Clusteranalyse werden die Grundprinzipien der Algorithmen hierarchisch-agglomerativer und K-means-Verfahren dargestellt. Ein Schwerpunkt liegt auf der graphischen Darstellung der Ergebnisse. Außerdem werden einige Verfahren zur Validierung von Clusterlösungen, wie der Vergleich von Lösungen hierarchisch-agglomerativer Verfahren mit K-means-Lösungen sowie Monte-Carlo-Verfahren zur Exploration des Einflusses von Startbedingungen bei K-means-Verfahren, vorgestellt.

Summary

The paper presents a short introduction to the aims of cluster analysis and describes the principles of hierarchical-agglomerative and K-means procedures. Graphical representations play an important role, while validation, for example by comparison of different hierarchical and K-means solutions or by Monte-Carlo simulations, is an important issue.

ZUMA How-to-Reihe, Nr. 10
2001

1 Einleitung

Unter Clusteranalyse versteht man Verfahren zur Einteilung einer Anzahl von Objekten in homogene Gruppen. Die durch eine Anzahl von Variablen beschriebenen Objekte sollen innerhalb einer Gruppe möglichst ähnlich bzgl. der Variablen sein. Objekte aus unterschiedlichen Gruppen sollen möglichst verschieden sein. Die Gruppen nennt man auch Cluster, Klassen oder Typen. Die Gruppeneinteilung wird auch als Klassifikation oder Typologie bezeichnet. [1]

Ein Beispiel aus den Sozialwissenschaften sind die Publikumstypologien in der Medienforschung[2]. Ziel dieser Typologien ist die Unterscheidung von Personengruppen mit unterschiedlichem Verhalten in der Nutzung von Hörfunk und Fernsehen. Grundlage der Typologien sind „einerseits soziodemographische, andererseits aber auch inhaltlich dem Lebensstilansatz verpflichtete Merkmale“ (Hartmann & Neuwöhner 1999, S. 531). Die empirische Basis der Typologien sind Umfragedaten. Eine Typologie wird als gelungen betrachtet, wenn die Gruppen möglichst homogen sind und sich klar voneinander unterscheiden. Dann lassen sie sich durch Repräsentanten beschreiben, die bzgl. der Variablen ein gruppentypisches Profil besitzen. Die in Hartmann & Neuwöhner (1999) dargestellte Typologie umfaßt neun Typen von Mediennutzern, die sich in sehr einprägsamer Weise („Junge Wilde“, „Erlebnisorientierte“ etc.) darstellen lassen.

Die üblichen Konstruktionsverfahren für Typologien oder Cluster teilen sich in zwei Klassen. Einerseits gibt es Verfahren zur Bildung hierarchischer Systeme von Gruppen von Objekten, die von der „feinsten“ Einteilung in einzelne Objekte bis zur „gröbsten“ Gruppierung reicht, bei der alle Objekte in einer einzigen Klasse versammelt sind. Aus diesem System ist dann eine der Gruppierungen zwischen den beiden Extremen als Lösung des Clusterungsproblems auszuwählen. Verfahren einer anderen Klasse, die sogenannten K-means-Verfahren, liefern eine Einteilung in Gruppen, die entsprechend einem globalen Maß optimale Homogenität unter allen Gruppierungen mit einer vorgegebenen Anzahl von Clustern aufweisen.

Die Beispiele der nachfolgenden Abschnitte sind mit zwei Programmen erstellt, die Clusteranalyse anbieten: SPSS (http://www.spss.com/germany) und ClustanGraphics (http://www.gesis.org/software/clustan/ und http://www.clustan.com).

2 Ähnlichkeitsmaße

Vor einer Analyse ist festzulegen, bzgl. welcher Variablen die Objekte miteinander verglichen werden sollen. Dann ist ein Maß zu bestimmen, mit dem die Ähnlichkeit oder Unähnlichkeit zwischen den Objekten numerisch ausgedrückt wird. Da Variablen in der Regel als numerische Codes gespeichert werden, ist jedes Objekt als Punkt in einem endlich-dimensionalen Raum repräsentiert. Seine Dimension stimmt mit der Anzahl der Analysevariablen überein. Als Maße für Unähnlichkeiten werden Metriken in endlichdimensionalen reellen Räumen oder davon abgeleitete Größen wie die Euklidische Metrik oder deren quadrierter Wert verwendet.

Die Lehrbuchliteratur führt neben der Euklidischen und der quadrierten Euklidischen Metrik eine Fülle weiterer Abstandsdefinitionen auf, die zu großen Teilen auch mit den üblichen Softwarepaketen realisiert werden können. Im Fall von binären Variablen werden auch Unähnlichkeitsmaße benutzt, die auf den Anzahlen der Differenzen und der Übereinstimmungen zweier Objekte in diesen Variablen beruhen. Eine detaillierte Diskussion verschiedener Me-

[1] Zur Zielsetzung und Anwendung der Clusteranalyse siehe Bacher (1994), Backhaus et al. (2000), Bailey (1994) und Kaufman & Rosseuuw (1990).
[2] Siehe hierzu Hartmann & Neuwöhner (1999)

triken, Ähnlichkeits- und Unähnlichkeitsmaße für Variablen unterschiedlicher Skalenniveaus und für den Fall, daß diese zusammen in eine Analyse eingehen, findet man z.B. in Kapitel 1 des Buchs von Kaufman und Rousseeuw (1990).

3 Hierarchisch-agglomerative Verfahren

3.1 Das Aggregationsschema

In Abschnitt 1 hatten wir kurz zwei Klassen von Clusterverfahren eingeführt, die hierarchischen und die K-means-Verfahren. Wir stellen im folgenden eine Teilklasse von sehr verbreiteten hierarchischen Verfahren dar, bei denen das hierarchische System durch sukzessives Gruppieren von Objekten und im weiteren Verlauf durch Fusion von Gruppen zu größeren Gruppen konstruiert wird. Man spricht dann auch von hierarchisch-agglomerativen Verfahren.

Die Aggregation beginnt also mit den feinstmöglichen Gruppen, die jeweils aus genau einem Objekt bestehen. Durch Zusammenfassen der zwei - im Sinne des verwendeten Abstandsmaßes - ähnlichsten Objekte wird eine erste zweielementige Gruppe gebildet. Bereits nach der Bildung einer ersten Gruppe muß die ursprüngliche Definition des Abstands zwischen einzelnen Objekten zu einer Definition für Abstände zwischen Gruppen und Objekten bzw. allgemeiner zwischen verschiedenen Gruppen erweitert werden (siehe Abschnitt 3.3). Die Abstände zwischen den Objekten, die im ersten Schritt des Verfahrens nicht aggregiert wurden, bleiben unverändert.

In den nachfolgenden Schritten werden dann Paare von Gruppen und/oder Objekten mit dem jeweils kleinsten Abstand zu neuen Gruppen zusammengefaßt. An jeden Aggregationsschritt schließt sich wieder die Berechnung neuer Abstände analog zum vorangehenden Schritt an. Das Verfahren erzeugt auf diese Weise in jedem Schritt eine neue Gruppe durch Vereinigung zweier bereits konstruierter Gruppen. Dadurch wird laufend eine etwas „gröbere“ Gruppierung generiert. Diese besteht nach dem i-ten Schritt aus n-i Gruppen, wobei n die Anzahl aller Objekte ist. Das Aggregationsschema besteht also aus der Iteration von zwei aufeinanderfolgenden Operationen:

1. Berechnung von Distanzen zwischen den Gruppen der i-ten Stufe;

2. Vereinigung derjenigen Gruppen der i-ten Stufe mit minimaler Distanz, wodurch die Gruppierung der (i+1)-ten Stufe erzeugt wird. Sie enthält eine Gruppe weniger als die Gruppierung der vorhergehenden i-ten Stufe und stimmt in (n-i-2) Gruppen mit denen der i-ten Stufe überein.

Der Prozeß wird bis zur Aggregierung aller n Objekte in einer einzigen Gruppe fortgesetzt. Dies wird nach n-1 Schritten erreicht. Eine charakteristische Statistik des Prozesses ist die Folge der Abstände, die die sukzessiv vereinigten Paare von Gruppen besitzen. Der Verlauf dieser Abstände, die auch als Fusionswerte bezeichnet werden, ist ein Hilfsmittel bei der Bestimmung der Clusterlösung.

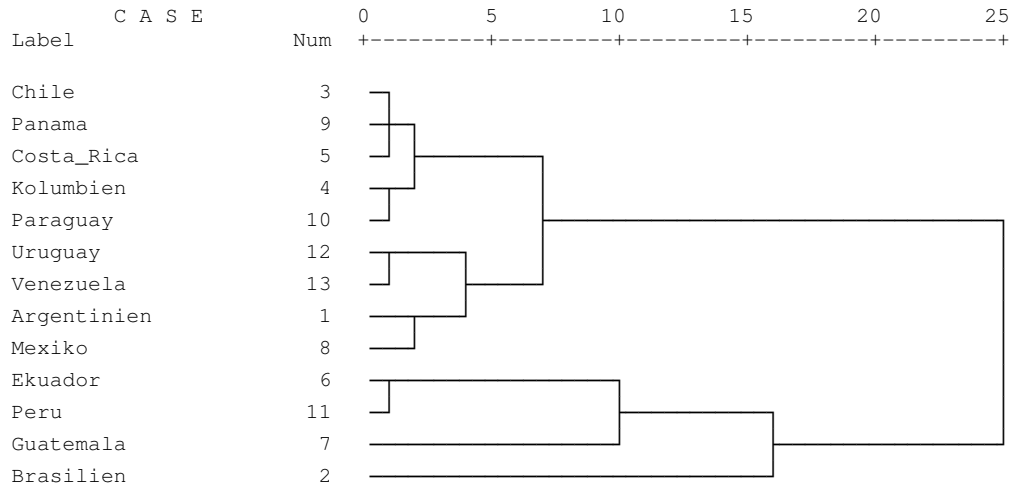
3.2 Darstellungen der Resultate hierarchisch-agglomerativer Verfahren: Dendrogramme

Bevor wir zu weiteren technischen Details der hierarchisch-agglomerativen Verfahren kommen, behandeln wir zunächst ein zentrales Instrument der Darstellung ihrer Resultate. Wie in Abschnitt 3.1 beschrieben wurde, wird durch schrittweise Vereinigung der zwei jeweils ähnlichsten Gruppen ein hierarchisches System erzeugt. Wie findet man aber nun in diesem hierarchischen System eine Lösung des Clusterungsproblems? Aus der etwas komplizierten Struktur der hierarchischen Systeme ließe sich kaum eine Gruppierung auswählen, die als

Clusterlösung interpretierbar wäre, wenn sich das System und der gesamte Aggregationsprozeß nicht in sehr übersichtlicher Weise als graphische Baumstruktur, als sogenanntes Dendrogramm, darstellen ließen. Die Gruppen des hierarchischen Systems bilden die Knoten des Dendrogramms, die Inklusionsrelationen zwischen einer Gruppe der i-ten Stufe und ihren Subgruppen der (i-1)-ten Stufe werden durch die Kanten des Dendrogramms dargestellt. Gruppen sind i.d.R. rechts von ihren Subgruppen positioniert. Üblicherweise werden die Knoten von Dendrogrammen aus Gründen der besseren Lesbarkeit nicht punktförmig, sondern als senkrechte Linien zwischen den waagerecht gezeichneten Vereinigungskanten dargestellt.

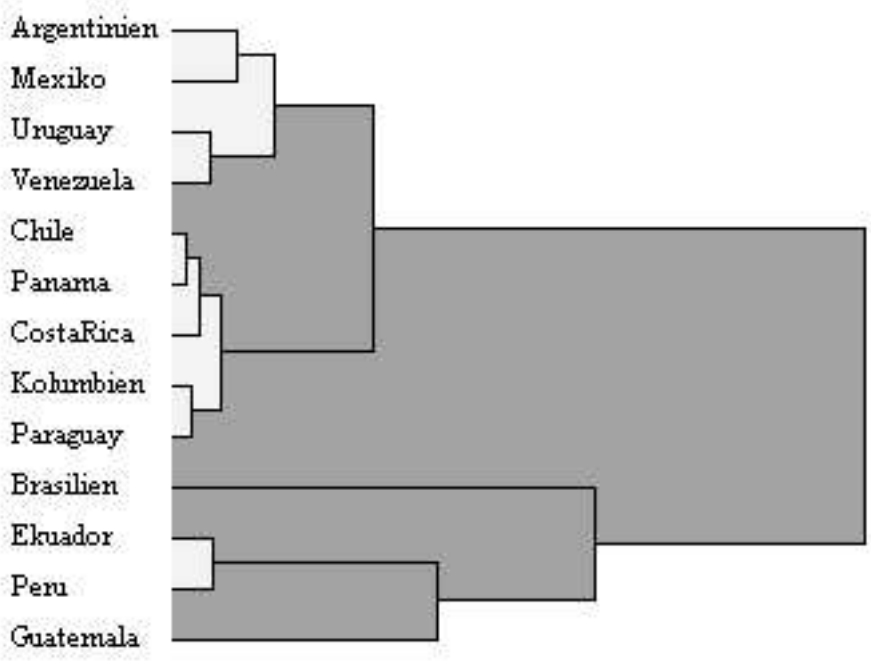
Wir erläutern dies am Beispiel eines sehr einfachen Datensatzes. Er enthält Informationen zu lateinamerikanischen Staaten, die einem UNO Bericht entnommen sind. Zur Clusterbildung wurden die (standardisierten) Variablen „Bevölkerung“, „Fläche“, „Bruttosozialprodukt“, „mittlere Lebenserwartung“, „Kindersterblichkeit“, „Einwohner/Arzt“, „Ernährung (tägliche Kalorienaufnahme)“ und „Alphabetisierungsgrad“ verwendet. Abbildung 1 zeigt das Dendrogramm der Clusteranalyse, wie es von SPSS erstellt wurde.

Abbildung 1: Dendrogramm der Clusteranalyse



Unter „Label“ wird das Land aufgeführt, unter „Num“ erscheint die Fallnummer im Datensatz. Die Skala am oberen Rand der Grafik gibt die skalierten Abstände zwischen den Clustern wieder. Diese Art der Dendrogrammdarstellung in SPSS ist für größere Objektmengen ungeeignet. In diesem Fall sind Hilfsmittel erforderlich, die SPSS nicht bietet, aber in ClustanGraphics verfügbar sind. Daher werden wir in den folgenden Beispielen die Dendrogrammdarstellung aus ClustanGraphics verwenden, obwohl es sich dabei um ein nicht allgemein verbreitetes stand-alone Programm handelt (siehe Abbildung 2).

Abbildung 2: Dendrogramm der Länderdaten (5-er Lösung)



Vor der Aggregation wird jedes der 13 Länder auf der linken Seite des Dendrogramms aufgelistet[3]. Im oberen Teil der Abbildung sehen wir, daß die Länder Argentinien und Mexiko zusammengefaßt werden. Dieses Zusammenführen ist dadurch gekennzeichnet, daß jedes der beiden Länder mit einer waagerechten Kante verbunden ist. Die Vereinigungskanten münden in die senkrechte Linie, die diese Gruppe symbolisiert ("Klammer"). Genauso sind Uruguay und Venezuela zu einer Gruppe verbunden. Auf höheren Stufen werden dann diese beiden Gruppen wiederum zu einer größeren Gruppe aus den Ländern Argentinien, Mexiko, Uruguay und Venezuela zusammengefaßt. Die Gruppierung wird so lange fortgesetzt, bis auf der letzten Stufe die Gruppe aus Argentinien, Mexiko, Uruguay, Venezuela, Chile, Panama, Costa Rica, Kolumbien und Paraguay mit der Gruppe aus Brasilien, Ekuador, Peru und Guatemala zur Gruppe aller Länder zusammengefügt wird.

In einem Dendrogramm sind aber nicht nur das hierarchische System der Gruppen und deren mengentheoretische Inklusionsbeziehungen, sondern auch die Abstandsrelationen zwischen den Gruppen und damit die Reihenfolge der Vereinigungen dargestellt. Im obigen Beispiel sind Chile und Panama das Paar der einander ähnlichsten Länder, die als erste zum kleinsten Fusionswert – ablesbar nach rechts in waagerechter Richtung – gruppiert werden, gefolgt vom Paar aus Kolumbien und Paraguay. Der nächstgrößere Abstand bzw. Fusionswert ist dann bereits ein Abstand zwischen einer Gruppe und einem einzelnen Land: Das Paar Chile und Panama wird mit Costa Rica vereinigt. Analog ist die Reihenfolge der weiteren Gruppenbildung aus dem Dendrogramm ablesbar. Außerdem gibt das Dendrogramm über die Kompaktheit von Clustern Aufschluß. Je „früher“ zwei Cluster vereinigt werden, d.h. je kürzer die Länge der vereinigenden Klammer in Richtung der x-Achse ist, desto kompakter ist das Vereinigungscluster, je weiter die Klammer nach rechts reicht, desto ausgedehnter ist es.

[3] Die Anordnung der Länder im ClustanGraphics-Dendrogramm unterscheidet sich von der im SPSS-Dendrogramm. Da in keinem der beiden Programme diese Reihenfolgen durch den Nutzer beeinflußt werden kann, lassen sich keine direkt vergleichbaren Dendrogramme erzeugen.

Für die meisten der üblichen Verfahren gilt eine Monotonieeigenschaft (siehe Abschnitt 3.3): Die Abstände oder Fusionswerte wachsen im Verlauf der Aggregation. Es werden zwar in einem Schritt unter allen Paaren von Objekten und/oder Gruppen diejenigen mit dem kleinsten Abstand vereinigt, doch ist dieser Abstand mindestens so groß wie die Abstände aller Paare, die in den vorangehenden Schritten zusammengeführt wurden. Betrachtet man den Abstand eines Vereinigungspaares als ein Maß ihrer Heterogenität, so kann man sagen, daß im Verlauf der Aggregation die Gruppen immer heterogener werden.

Da nun der Abstand einer Gruppe vom linken Rand des Dendrogramms dem Abstand zwischen dem Paar entspricht, aus dem es gebildet wird, so wird wegen der o.g. Monotonieeigenschaft eine Gruppe immer rechts von ihren Subgruppen positioniert sein.

Wenn die Folge der Abstände für die ersten Stufen der Aggregation nur kleine Zuwächse zeigt, dann liegt auch nur eine geringe Zunahme der Heterogenität der Gruppen vor. Werden aber die Gruppen nachfolgender Stufen mit einem sprunghaft angewachsenen Fusionswert erreicht, so nimmt auch die Heterogenität sprunghaft zu. Da die Gruppen möglichst homogen sein sollen, wenn sie als Cluster oder Typus gelten sollen, wird man als Lösung des Clusterproblems die Gruppen auf der Stufe unmittelbar vor dem „großen Sprung“ ansehen.

Dazu läßt sich die Folge der Fusionswerte, wie sie im Dendrogramm ablesbar sind, in ClustanGraphics durch eine explizite Option und in SPSS durch einen kleinen Trick[4] als Lineplot darstellen. Aus dessen Verlauf ist sichtbar, nach wie vielen Schritten bzw. bei welcher Anzahl von Gruppen sich die Zunahme der Heterogenität deutlich erhöht.

Das Kriterium des Wachstums der Fusionswerte verdeutlichen wir im folgenden noch mit den Daten des ALLBUS 98 (Allgemeine Bevölkerungsumfrage 1998, http://www.gesis.org/Dauerbeobachtung/Allbus/). Der Datensatz enthält 473 Variablen und 3234 Fälle. Für die folgende Clusteranalyse wurden drei Variablen ausgewählt, die zur Gruppierung der Fälle hinzugezogen werden sollen:

V51: durchschnittliche Fernsehgesamtdauer pro Tag in Minuten;

V64: durchschnittliche Radiohör-Dauer pro Tag in Minuten;

V66: Häufigkeit des Zeitungslesens pro Woche.

Fehlende Werte wurden fallweise eliminiert, so daß die Analysedatei 2957 Fälle enthält.

Abbildung 3 zeigt ein auf die letzten Fusionen verkürztes Dendrogramm der Clusteranalyse mit den Medienvariablen des ALLBUS: Durch eine senkrechte Trennlinie ist ein Schwellenwert markiert, oberhalb dessen die Zuwächse der Fusionswerte augenscheinlich stark zunehmen. Der Teil des Dendrogramms vom linken Rand bis zu den vier Gruppen, deren Fusionswerte unterhalb der Markierung liegen, ist eingefärbt und hebt dadurch diese Clusterlösung hervor. Alternativ dazu ist in Abbildung 4 ein kleinerer Schwellenwert mit einer Lösung von zehn Clustern markiert; auch hier ist oberhalb des Schwellenwerts ein deutlicher Anstieg der Fusionswerte zu verzeichnen.

Bei Umfragedaten mit hohen Stichprobenumfängen sind die Dendrogramme nur darstellbar, wenn sie verkürzt werden, d.h. wenn nur das Teildendrogramm der letzten Fusionen ab einem bestimmten Schwellenwert bzw. unterhalb einer bestimmten Anzahl von Gruppen gezeigt wird. Dann werden am linken Rand des Dendrogramms keine einzelnen Objekte, sondern

[4] SPSS gibt die Folge der Fusionswerte als Tabelle aus. Diese Tabelle kann durch Doppelklick zur Bearbeitung aktiviert werden. Die gesamte Spalte der Koeffizienten oder nur die letzten m Fusionswerte können nun markiert werden. Durch Klicken mit der rechten Maustaste wird ein Menü aktiviert, aus dem die Option „Create graph“ und „line“ ausgewählt werden kann. Es wird ein Lineplot der Fusionswerte angezeigt. Diesem Verfahren sind allerdings durch Restriktionen des SPSS Output Viewers Grenzen in Bezug auf die Zahl der Agglomerationsschritte in der Tabelle, die noch bearbeitbar ist, gesetzt.

Gruppen dargestellt. Diese werden in dem obigen Beispiel jeweils durch das Objekt bezeichnet, das in der Gruppe den kleinsten Fusionswert besitzt und in der Reihenfolge der Daten als erstes im Datensatz erscheint. In den eckigen Klammern stehen die zugehörigen Gruppenumfänge. Diese Möglichkeit der Verkürzung des Dendrogramms ist in ClustanGraphics, aber nicht in SPSS realisiert (siehe Abschnitt 4).

Abbildung 3: Dendrogramm der Allbus-Daten (4-er Lösung)

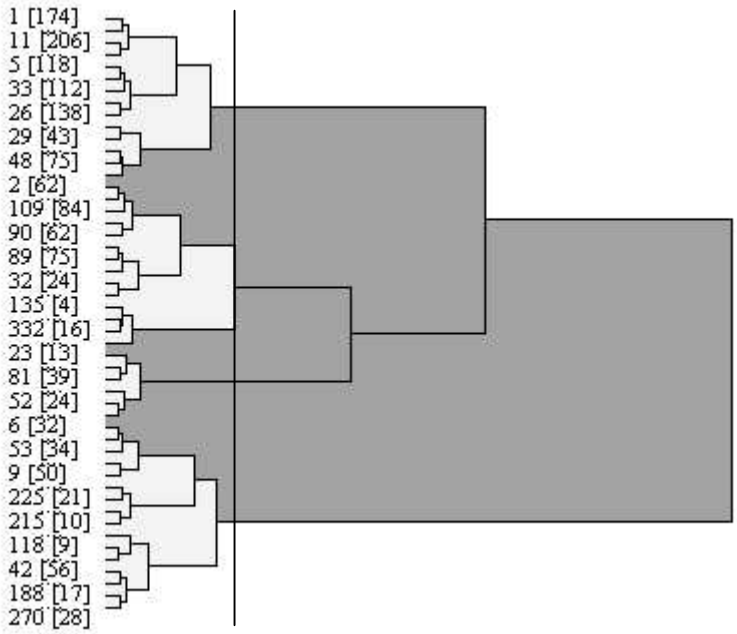
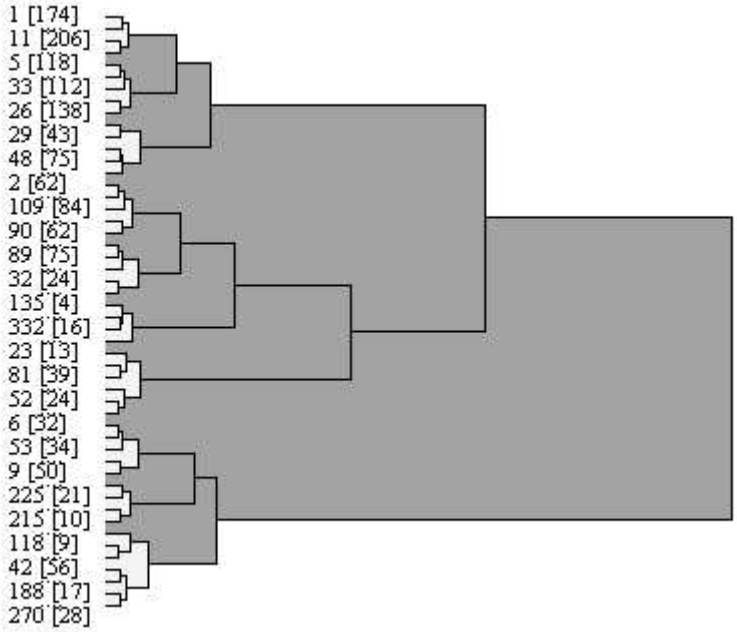


Abbildung 4: Dendrogramm der Allbus-Daten (10-er Lösung)



In den Dendrogrammen der Abbildungen 3 und 4 lassen sich die 4-Cluster- und die 10-Clusterlösung erkennen. Zur Interpretation der Ergebnisse sind allerdings weitere Analysen erforderlich (siehe dazu Abschnitt 5).

Für die Beurteilung der Clusterstruktur sollte man immer das gesamte Dendrogramm bzw. das gekürzte Dendrogramm der letzten Aggregationsstufen heranziehen. Wie man am Beispiel der Allbus-Daten sieht, weisen einzelne Gruppen, die als Cluster interpretierbar sind, selbst wieder eine Clusterstruktur auf. In einer derartigen Situation würde man nur eine unvollständige Analyse der Daten liefern, wenn man sich für eine einzige Clusterstruktur oberhalb eines bestimmten Heterogenitätslevels als Clusterlösung entscheiden würde.

3.3 Abstandsmaße zwischen Gruppen

In Abschnitt 3.1 wurde angemerkt, daß die sukzessive Aggregation eine Erweiterung der Definition von Metriken auf Abstände zwischen Gruppen von Objekten erfordert. Es gibt dafür verschiedene Möglichkeiten, jede von ihnen impliziert ein spezielles Fusionskriterium für den Aggregationsprozeß.

Einige Definitionen verwenden den Abstand zwischen zwei Objekten aus je einer Gruppe. Beim *single linkage*-Verfahren (SPSS: „nearest neighbor“) wird der minimale Abstand verwendet: Der Abstand zwischen zwei Gruppen ist gleich dem minimalen Abstand zweier Objekte aus je einer der Gruppen. Im Aggregationsprozeß begünstigt dieses Maß die Vereinigung von Gruppen, die aneinander grenzen oder die durch dazwischenliegende Objekte „verbunden“ sind. Dadurch kann es zur Ausbildung „fadenförmiger“ Cluster kommen, wenn sich Gruppen entlang Pfaden zu den nächstgelegenen einzelnen Objekten in ihrer Umgebung fortsetzen.

Beim Verfahren des *complete linkage* (SPSS: „furthest neighbor“) liegt der genau umgekehrte Fall vor: Hier ist der Abstand zweier Gruppen gleich dem maximalen Abstand zweier Objekte aus je einer Gruppe. Ein möglicher Nachteil besteht darin, daß der „Zwischenraum“ zwischen einem Vereinigungspaar u. U. größer ist als bei einigen alternativen Paaren. Man wird das Minimax-Kriterium von complete linkage dann wählen, wenn Cluster mit möglichst geringem Durchmesser gesucht werden.

Ein Mittelweg zwischen single und complete linkage ist *average linkage* („between-groups linkage“): Hier wird der Abstand zweier Gruppen als Mittelwert der Abstände zwischen allen Paaren von Objekten definiert.

Die bisher genannten Verfahren besitzen die o.g. Monotonieeigenschaft: Die Folge der Abstände zwischen den Gruppen, die schrittweise zusammengeführt werden, ist monoton steigend, d.h. die Heterogenität der Gruppen wächst im Verlauf des Verfahrens. Dies ist sowohl eine plausible Eigenschaft des Aggregationsprozesses als auch eine wesentliche Voraussetzung für die Dendrogramm-Darstellung mit überschneidungsfreien Kanten (siehe Abschnitt 3.2).

Die Wahl des Abstands zwischen den Schwerpunkten zweier Gruppen („centroid clustering“) ist eine weitere Möglichkeit der Definition. Dies ist ein anderer Mittelweg zwischen single linkage und complete linkage, für den aber die o.g. Monotonieeigenschaft nicht allgemein garantiert ist: Der gemeinsame Schwerpunkt zweier vereinigter Gruppen liegt u.U. näher an den Schwerpunkten anderer Gruppen, d.h. der Abstand zu diesen Gruppen verringert sich nach der Fusion. In einem Dendrogramm zeigt sich diese Situation daran, daß die Position einer Gruppe nicht wie in den obigen Beispielen rechts von den beiden konstituierenden Subgruppen in Richtung größerer Fusionswerte, sondern links davon liegt. In der praktischen Anwendung werden drastische „Umkehreffekte“ aber kaum registriert, da aus geometrischen Gründen der numerische Spielraum für solche Effekte nicht sehr groß ist.

Ein weiteres sehr häufig verwendetes Verfahren legt als Fusionskriterium fest, daß die Binnenvarianz der Gruppierungen minimal wächst. Es wird auch als *Ward*'sches Verfahren („Ward's method") bezeichnet. Wegen seiner Berücksichtigung der Heterogenität in den Gruppen gemessen durch die Binnenvarianz, zu der alle Objekte einer Gruppe gemäß ihrem Abstand zum Gruppenschwerpunkt gleichmäßig beitragen, wird es häufig anderen Verfahren vorgezogen. Es bildet konvexe Gruppen und begünstigt eine gleichmäßige Besetzung von Gruppen. Vermutet man die Existenz konvexer, balanciert besetzter Cluster, so können diese durch das Ward'sche Verfahren identifiziert werden.

Diese und weitere Abstandsdefinitionen zusammen mit den zugehörigen Agglomerationsverfahren und ihren Eigenschaften sind ausführlich in Kaufman & Rousseeuw (1990) diskutiert. Insbesondere behandeln die Autoren topologische Eigenschaften der Aggregation, also der räumlichen Ausdehnung der Gruppen verglichen mit den gegenseitigen Abständen der Objekte in den Gruppen. Die oben angesprochene Tendenz von single linkage, entfernte Objekte über eine „Brücke" von benachbarten Objekten miteinander in einer Gruppe zu vereinigen, kann geometrisch anschaulich als raumzusammenziehend (space contracting) interpretiert werden. Der umgekehrte Effekt bei complete linkage wird als raumstreckend oder raumdehnend (space dilating) bezeichnet: Zwei Gruppen werden nicht zusammengefaßt, auch wenn sie nahe aneinandergrenzen, wenn die Vereinigung eine der Gruppen mit einer dritten einen kleineren Durchmesser besitzt. Als raumerhaltend (space conserving) werden schließlich Verfahren bezeichnet, bei denen die Zwischenräume zwischen den Gruppen und ihre Ausdehnung im Verlauf der Aggregation – weitgehend – der geometrischen Konfiguration ihrer einzelnen Objekte entsprechen. Hierzu zählen das Ward'sche Verfahren und average linkage.

Die unterschiedlichen topologischen Eigenschaften der verschiedenen Abstandsmaße bzw. Fusionskriterien können zur Konstruktion unterschiedlicher hierarchischer Strukturen und unterschiedlichen Lösungen des Klassifikationsproblems führen. Das hängt in starkem Maß auch von den Daten ab. Diese unterschiedlichen Strukturen sind mitunter auch unterschiedlich gut geeignet für eine Auswahl und Interpretation von Clusterlösungen. Je kompakter Cluster eines Datensatzes sind, und je klarer diese voneinander getrennt sind, desto stabiler ist auch die Rekonstruktion dieser Cluster durch hierarchische Verfahren mit unterschiedlichen Fusionskriterien.

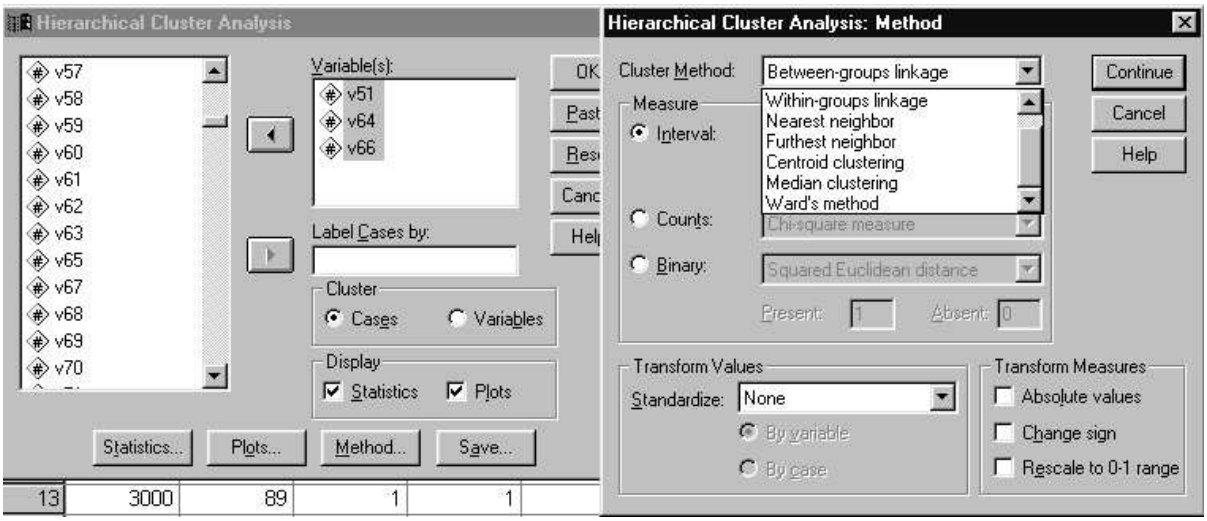
4 Durchführung einer Clusteranalyse

In den folgenden Abschnitten verwenden wir wieder den Allbus 98-Datensatz, den wir bereits in Abschnitt 3.2 für Dendrogramme herangezogen haben.

SPSS ist ein Statistikprogramm, das neben vielen verschiedenen Statistik-Analysemöglichkeiten hierarchische Clusteranalyse und K-means-Verfahren anbietet. Ähnliche Möglichkeiten werden auch in anderen Statistik-Programmen, z.B. in Statistica, angeboten, auf die wir uns hier aus Platzgründen nicht beziehen können. ClustanGraphics (Wishart 1999) ist ein reines Clusteranalyse-Programm, das speziell auf die Probleme der Clusteranalyse zugeschnitten ist und im Vergleich zu SPSS erweiterte Analyse- und graphische Darstellungsmöglichkeiten erlaubt (siehe 3.2).

SPSS bietet unter dem Menü-Punkt „Classify" u.a. hierarchische Clusteranalyse. In Abbildung 4 (links) ist das Hauptfenster der Clusteranalyse gezeigt, in dem die Variablen, die zur Clusterung verwendet werden sollen, ausgewählt werden können.

Abbildung 5: SPSS Menü zur hierarchischen Clusteranalyse



Unter „Statistics“ können verschiedene Dokumentationen des Agglomerationsprozesses angefordert werden:

- eine Agglomerationstabelle, die alle Schritte der Aggregation dokumentiert;
- die Ähnlichkeitsmatrix;
- eine Tabelle der Clusterzugehörigkeit für jeden Fall auf einer (oder mehreren) festzulegenden Stufen.

Unter „Plots“ können Diagramme zur Verdeutlichung des Agglomerationsprozesses angefordert werden: Ein Icicle-Plot, der die Agglomerationstabelle bildlich darstellt, und das Dendrogramm. Beide Ausgaben sind bei Umfragedaten mit großen Fallzahlen nur sehr schwer zu lesen. SPSS empfiehlt diese Darstellungen nur für kleine Fallzahlen (SPSS Base 8.0, S. 296). Beim Icicle-Plot besteht die Möglichkeit, die Ausgabe des Diagramms auf die letzten Fusionsschritte zu begrenzen. Dies ist beim Dendrogramm nicht möglich. Eine Alternative dazu bietet das Programm ClustanGraphics, das dem Forscher die Möglichkeit gibt, das Dendrogramm erst ab einer bestimmten Stufe anzuzeigen.

Es sei auch noch auf die nützliche Option in ClustanGraphics hingewiesen, nach der in einem Dendrogramm durch Mausklick auf der Höhe eines bestimmten Fusionswerts die Gruppen hell eingefärbt werden, deren Fusionswerte unterhalb dieses Schwellenwerts liegen. Dadurch treten die Clusterstrukturen besser hervor (siehe Abschnitt 3.2).

Die Abstände im Dendrogramm werden in der Software immer so skaliert, daß es in der gesamten Breite am Bildschirm bzw. auf dem Papier zu sehen ist. In SPSS bedeutet das z.B., daß alle Abstände so umgerechnet werden, daß sie in eine Skala zwischen 0 und 25 passen. In ClustanGraphics werden die Abstände kontinuierlich skaliert. Es kann in beiden Fällen vorkommen, daß Unterschiede bei sehr kleinen Abständen nicht mehr sichtbar sind, und daß es vor allem auf den unteren Ebenen oft so erscheint, als ob mehr als zwei Gruppen im gleichen Schritt vereinigt würden (siehe auch Abbildung 1).

Unter „Save“ kann ausgewählt werden, ob die Clusterzugehörigkeit auf einer oder mehreren Aggregationsstufen den Objekten, d.h. Fällen des Datensatzes, zugespielt werden soll. Als Resultat enthält die SPSS-Datei danach eine oder mehrere Variablen, die die Clusterzugehörigkeit eines Objekts beschreiben.

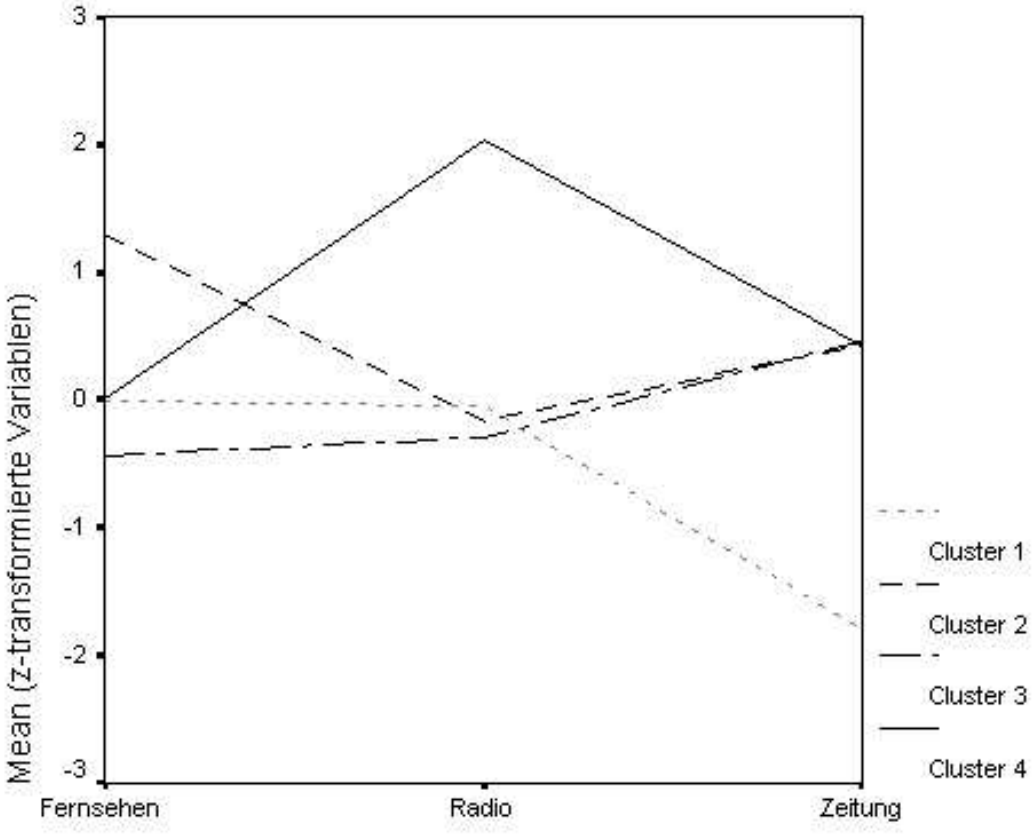
Während unter den bisher beschriebenen Optionen die Form der Ausgabe festgelegt wird, werden unter „Method“ (siehe Abbildung 5, rechts) Vorgaben für die Clusteranalysen gemacht. Zunächst wird das Distanzmaß für das Zusammenfügen von Gruppen bestimmt (Cluster Method). Als Voreinstellung bietet SPSS „between-groups linkage“ (average linkage between groups). Ein bei Umfragedaten sehr häufig verwendetes Maß ist dagegen „Ward’s method“ (siehe 3.3). Zusätzlich kann das Ähnlichkeitsmaß für das Zusammenfügen der Objekte festgelegt werden. Hier wird im Menü von intervallskalierten Variablen ausgegangen und die quadrierte Euklidsche Metrik als Maß angeboten. Im Menü kann unter einigen weiteren Maßen gewählt werden. In der Syntaxsprache von SPSS (über „Paste“) werden zusätzliche Optionen angeboten. Es kann auch angegeben werden, ob die Variablen standardisiert werden sollen (z.B. durch Bilden von z-Scores). Dies bietet sich an, wenn die Wertebereiche der Variablen sehr unterschiedliche Ausdehnung aufweisen: Bei der Berechnung der Ähnlichkeitsmaße kann dies zu sehr ungleichgewichtigen Einflüssen der Variablen auf das Distanzmaß führen.

5 Inhaltliche Interpretationen: Lineplots

Wenn mit Hilfe des Dendrogramms eine Clusterstruktur ausgewählt wurde, dann ist man natürlich an der inhaltlichen Bedeutung der Cluster interessiert. Im Idealfall ist jedes Cluster durch die Ausprägungskombinationen der Variablen beschreibbar, die für die Bestimmung der Distanzen bzw. Ähnlichkeiten herangezogen wurden. Es ist also plausibel, das mittlere Profil dieser Variablen in einem Cluster, d.h. den (multivariaten) Mittelwert der Variablen, als ein das Cluster charakterisierendes Profil anzusehen. Wenn sich unter den Variablen nominalskalierte oder ordinale Variable befinden, dann lassen sich die Mittelwerte gar nicht oder nur eingeschränkt interpretieren. Im Falle nominalskalierter Variabler ist es daher besser, sie als ein System von dichotomen 0-1- codierten Variablen darzustellen, da sich hier der Mittelwert als relativer Anteil der Objekte mit der Ausprägung „1“ interpretieren läßt. Aus den Mittelwerten aller zu einer nominalen Variablen gehörigen dichotomen Variablen lassen sich dann die relativen Anteile der Kategorien der Variablen ablesen.

Für die Interpretation der Cluster ist es wichtig, die Kombination der Ausprägungen aller Variablen simultan darzustellen bzw. dies für deren mittleres Profil zu tun. Bei mehr als drei Komponenten - und darum handelt es sich in der Regel - ist eine Darstellung als Streudiagramm ausgeschlossen. Man muß hier zu anderen Formen der Darstellung multivariater Größen greifen. Von den verschiedenen Möglichkeiten wollen wir hier kurz auf die Darstellungsweise durch Lineplots eingehen, da diese auch mit einfachen Mitteln in SPSS realisiert werden können. In ClustanGraphics ist die Präsentation von Mittelwertprofilen eine eigene direkt aufrufbare Option. Zur Berechnung der Profile in SPSS muß durch Wahl der entsprechenden Optionen des „Save“-Menüpunkts in der Clusteranalyse-Prozedur die Zugehörigkeit der Objekte zu den Clustern in Form einer eigenen Variablen gespeichert werden. Dann kann - nach vorheriger Sortierung nach Clusterzugehörigkeit - durch die SPSS-Prozedur „Aggregate“, bei der die Clusterzugehörigkeit als „Break“-Variable fungiert, eine Datei erzeugt werden, deren Fälle genau den Clustern entsprechen und bei der die Clustervariablen zu Mittelwerten innerhalb der Cluster aggregiert sind. Werden die Zeilen und Spalten dieses Files durch die Prozedur „Transpose“ vertauscht, so erscheinen die aggregierten Variablen im transponierten File als Fälle und jedes der Cluster als separate Variable, deren Ausprägungen die Mittelwerte der Clustervariablen sind. Mit der Option (multiple) „Lineplot“ unter dem Menüpunkt „Graph“ lassen sich die Mittelwerte der Variablen in den Clustern als Linienzüge darstellen. Jeder Linienzug repräsentiert genau ein Cluster und zeigt das charakteristische Profil der Cluster bzgl. der Clustervariablen.

Abbildung 6: Lineplot der Mittelwerte der 4-er Lösung der Allbus-Daten



Die Ergebnisse der Allbus-Daten führen anhand der Mittelwerte der einzelnen Variablen zu folgender Interpretation. Es können vier verschiedene Typen von Mediennutzung unterschieden werden:

1. durchschnittliche Radio und Fernsehnutzung, es wird aber kaum Zeitung gelesen
2. häufige Fernsehnutzung (> 4,5 Stunden)
3. eher geringe Mediennutzung insgesamt
4. häufige Radionutzung (> 9 Stunden)

Ein Schwachpunkt dieser Darstellungsform ist allerdings, daß die Heterogenität innerhalb der Cluster nicht darstellbar ist. Natürlich lassen sich die Varianzen der Clustervariablen innerhalb der Cluster berechnen bzw. läßt sich durch Boxplots die Separierung der Cluster darstellen. Dies sind jedoch Hilfsmittel, die in ClustanGraphics überhaupt nicht und in SPSS nur sehr umständlich einsetzbar sind.

6 Kapazitätsprobleme bei hierarchisch-agglomerativen Verfahren

Kapazitätsprobleme bei Software und Hardware bei größeren Fallzahlen haben sich seit einiger Zeit deutlich vermindert. Sie können aber, abhängig von Software und Fallzahl, vor allem bei den hierarchisch-agglomerativen Verfahren immer noch auftreten. Die Kapazitätsgrenzen unterscheiden sich für die gebräuchlichen Softwareprogramme stärker als früher: Analysen, die mit einer Software problemlos möglich sind, können durchaus in anderen Programmen scheitern. Da die Kapazitätsgrenzen auch von der verfügbaren Hardware (Kernspeicher und Plattenplatz) abhängen, lassen sich genaue Grenzen nicht angeben.

Der Agglomerationsprozeß läßt sich anhand des Dendrogramms sehr gut nachvollziehen - sofern die Fallzahl nicht allzu groß ist. Analoges gilt für alle Teile der Druckausgabe, mit denen der gesamte Aggregationsvorgang dokumentiert wird. Eine Grenze für übersichtliche Darstellungen dürfte etwa bei 250 Objekten liegen. In der Umfrageforschung findet man allerdings sehr oft erheblich größere Zahlen. Die Nutzung von Dendrogrammen ist dann nicht mehr für einzelne Objekte möglich, kann aber bei der Darstellung des Agglomerationsprozesses beispielsweise der letzten 250 Schritte eingesetzt werden.

Diese Verkürzung des Dendrogramms ist aber nur dann bequem durchführbar, wenn sie explizit zu den Optionen der Software gehört, wie beispielsweise bei ClustanGraphics. Anderenfalls muß für die Darstellung - z.B. der letzten 250 Schritte - die Distanzmatrix der Gruppen der (n-250)-ten Stufe berechnet und als Eingabe verwendet werden. Der Zugriff auf solche Distanzmatrizen im Verlauf der Aggregation und eine Dendrogrammverkürzung sind mit SPSS nicht direkt umsetzbar, die Berechnung der Distanzmatrix der Gruppierung der (n–250)-ten Stufe kann je nach Definition der Distanz auch nur sehr aufwendig berechnet werden.

7 K-means

7.1 Grundschema der Algorithmen

Die Resultate hierarchisch-agglomerativer Verfahren hängen neben den Daten von ihren topologischen Eigenschaften ab, die wiederum von den verschiedenen Definitionen von Abständen abhängen. Die K-means-Verfahren konstruieren dagegen Gruppen ohne einen Aggregationsprozeß, in dessen Verlauf topologische Eigenschaften wirksam werden könnten. Ihre Resultate sind deswegen raumerhaltend.

Grundlage ist wieder die Repräsentierung der Objekte in einem reellen Variablenraum mit einer Metrik, auch hier i.d.R. die Euklidische Metrik. Mit Hilfe dieser Metrik wird ein globales Maß für die Binnenheterogenität der Gruppen definiert. Dann wird nach einer Gruppierung gesucht, die dieses Maß minimiert, wobei die Anzahl der Gruppen vorgegeben ist. Das Verfahren ermittelt also keine Gruppierungen, in denen einzelne Cluster besonders kompakt auf Kosten hoher Heterogenität anderer Gruppen sind, sondern Gruppen mit einer „mittleren“ Homogenität.

Ein solches globales Heterogenitätsmaß ist z.B. die Summe der quadrierten Abstände der Objekte einer Gruppe von den Gruppenschwerpunkten. Dieses Maß liegt der folgenden Darstellung zugrunde. Das Kriterium kann als Maß der Binnenvarianz in den Gruppen interpretiert werden. Es ist in SPSS und ClustanGraphics implementiert.

Sein prinzipieller Nachteil ist, daß die Anzahl der zu prüfenden Gruppierungen auch bei Stichproben kleineren Umfangs schon enorm groß ist, sodaß Kapazitätsgrenzen auch sehr leistungsfähiger Rechner schnell erreicht werden. Daher wird das Kriterium der minimalen globalen Binnenvarianz durch ein relatives Kriterium ersetzt: Es werden Gruppierungen gesucht, bei denen jedes Objekt zum Schwerpunkt seiner Gruppe einen kleineren Abstand besitzt als zu den anderen Gruppenschwerpunkten. Es leuchtet intuitiv ein, daß für die unter dem globalen Kriterium optimale Gruppierung diese sogenannte Minimal-Distanz-Eigenschaft zutrifft. Allerdings legt diese Eigenschaft Gruppierungen nicht eindeutig fest: Es kann für einen Datensatz tatsächlich mehrere Gruppierungen mit der Minimal-Distanz-Eigenschaft geben. Unterschiedliche Lösungen können von den Gruppierungen abhängen, mit denen der Lösungsalgorithmus startet, aber auch von der Reihenfolge, mit der die Objekte abgearbeitet werden. Eine ausführliche Darstellung hierzu findet sich in (Kaufmann & Pape 1984).

Der Algorithmus für Minimal-Distanz-Lösungen ist ein Austauschalgorithmus (Kaufmann & Pape 1984, S. 409ff), der nach dem folgenden Schema funktioniert: Zu einer beliebigen ersten Gruppierung ermittele man die Schwerpunkte ihrer Gruppen. Anschließend bilde man eine neue Gruppierung, indem jedes Objekt demjenigen Gruppenschwerpunkt zugeordnet wird, dem es am nächsten ist. Für die neuen Gruppen berechne man wieder die Schwerpunkte und bilde mit diesen analog wieder eine neue Gruppierung. Das Verfahren bricht ab, wenn mit einem neuen Schritt keine neue Gruppierung erzeugt wird. Die resultierende Gruppierung hat die Minimal-Distanz-Eigenschaft. Bei jedem Schritt sinkt auch die globale Binnenvarianz. Dabei wird jedoch nicht notwendigerweise ein globales Minimum erreicht, sondern mitunter nur ein lokales Minimum.

Startgruppierungen lassen sich durch Vorgabe von Punkten erzeugen, denen die Objekte dann nach Maßgabe des kürzesten Abstands zugeordnet werden. Diese Punkte werden auch als Startzentren oder „cluster centers“ (SPSS) bezeichnet.

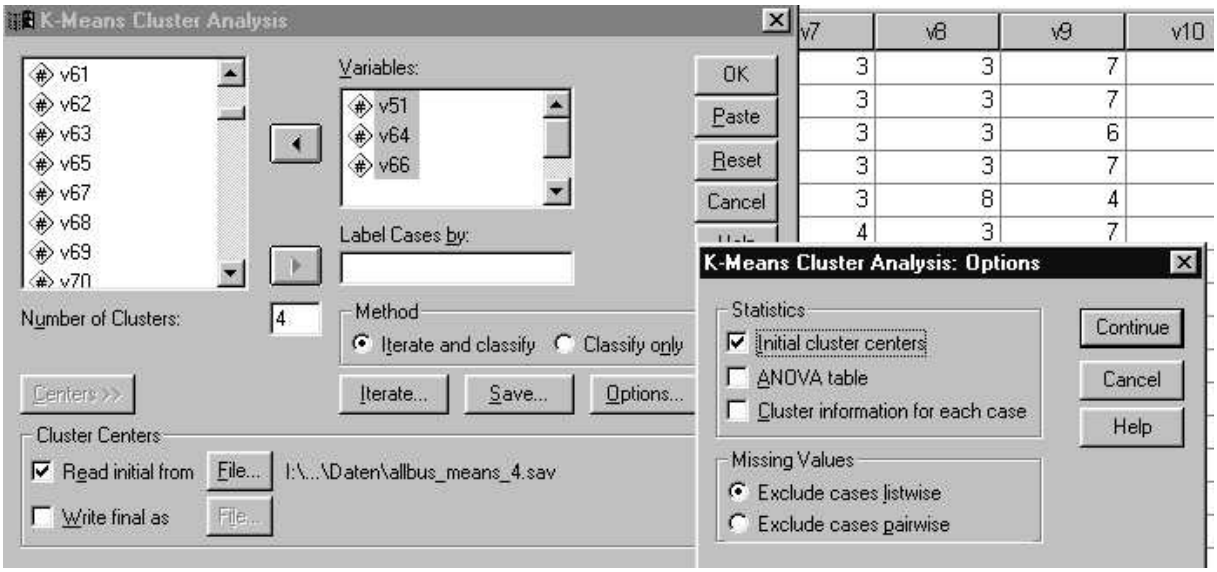
Nur in Situationen außerordentlich klarer Clusterstrukturen in den Daten wird die Lösung eines K-means-Verfahrens invariant gegenüber den Startgruppierungen sein. Die Eindeutigkeit einer Lösung bei einer Vielzahl von zufällig gewählten Startgruppierungen ist umgekehrt ein Hinweis auf die Ausgeprägtheit der Clusterstruktur (vgl. Monte-Carlo-Verfahren in Abschnitt 8). Eine Variante des Algorithmus besteht darin, daß die Schwerpunkte als neue Zuordnungszentren nicht erst nach der vollständigen Konstruktion der neuen Gruppierung, sondern bereits nach jeder Zuordnung eines einzelnen Objekts berechnet werden („running means“). Diese Option kann zwar die Konvergenz des Verfahrens beschleunigen, aber bei gleichen Startgruppierungen zu wieder anderen Minimal-Distanz-Lösungen führen.

Ein Problem des K-means-Verfahrens ist die Vorgabe der Anzahl der Cluster. Hierzu sind Vorinformationen über die Clusterstruktur nötig. Diese können aus der Analyse anderer Daten stammen, aber auch aus hierarchisch-agglomerativen Analysen oder aus einer Theorie abgeleitet sein. In jedem Fall wird man sich in der Anwendung nicht auf eine einzige Vorgabe beschränken, sondern Lösungen zu einem Intervall von möglichen Clusteranzahlen ermitteln. Aus der Abhängigkeit der minimalen Binnenvarianz von der Anzahl der Cluster erhält man Hinweise auf plausible Clusterstrukturen. Je höher die vorgegebene Clusteranzahl ist, desto kleiner ist auch die Binnenvarianz. Steigt diese sprunghaft an, wenn die Clusterzahl unter einen bestimmten Wert sinkt, so wird man als Anzahl der Cluster diesen Schwellenwert wählen. Man sollte bei diesem Vorgehen aber sicher sein, daß man zu einer vorgegebenen Clusterzahl auch tatsächlich das globale und nicht nur ein lokales Minimum ermittelt hat. Dies erhält man durch eine hinreichende Anzahl von Monte-Carlo-Studien mit zufällig variierenden Startgruppierungen (siehe dazu auch Abschnitt 8).

7.2 K-means in SPSS

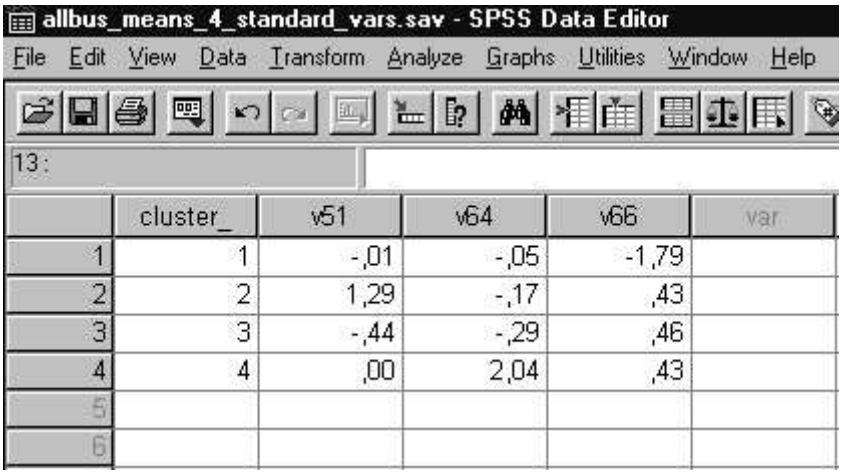
Unter dem Menüpunkt „Classify“ wird in SPSS u.a. K-means-Clusterung angeboten (siehe Abbildung 7). Im Hauptfenster zu diesem Menü können die Variablen und die Zahl der Cluster ausgewählt werden. Daneben kann die Zahl der Iterationen bestimmt und festgelegt werden, ob die Clusterzugehörigkeit zu den Fällen der Datei zugespielt werden soll.

Abbildung 7: Menü zu K-means in SPSS



Als Startzentren (cluster centers) werden per Voreinstellung von SPSS die Werte der ersten k Fälle (bei k vorgegebenen Clustern) verwendet. Der Forscher hat aber auch die Möglichkeit, die Startzentren als Datei einzulesen. Diese müssen in einer eigenen SPSS-Datei gespeichert werden, die ein fest vorgegebenes Format haben muß. Die erste Variable der Datei muß die Variable „cluster_“ sein, die die Nummer des jeweiligen Clusters enthält. Dann folgen die Variablen, die die Mittelwerte enthalten. Für jede zur Clusterbildung verwendete Variable muß es eine entsprechende Variable geben, die zu jedem Cluster den Mittelwert dieser Variablen enthält. Die Variablennamen und die Reihenfolge der Variablen müssen genau denen der Originalvariablen in der SPSS-Datei entsprechen. Für unser Beispiel (Allbus-4-Clusterlösung) sieht die Mittelwertdatei wie in Abbildung 8 angegeben aus.

Abbildung 8: Mittelwertdaten in SPSS



	cluster_	v51	v64	v66	var
1	1	-,01	-,05	-1,79	
2	2	1,29	-,17	,43	
3	3	-,44	-,29	,46	
4	4	,00	2,04	,43	
5					
6					

Unter „Options“ können bei K-means verschiedene Ausgaben gewählt werden:

- die Startzentren der Cluster (die Schwerpunkte der Cluster einer Lösung werden immer ausgegeben);

- eine ANOVA Tabelle, in der neben der Varianz innerhalb eines Clusters und der Varianz zwischen den Clustern der F-Wert und die Signifikanz angegeben werden. Als Anhaltspunkt kann dienen: Je kleiner die Signifikanz und je größer der F-Wert, um so größer die Distanz zwischen den Clustern;

- Cluster-Informationen zu jedem Fall (Clusterzugehörigkeit, Abstand vom Clusterschwerpunkt).

Zusätzlich wird immer eine Tabelle der Abstände der Clusterschwerpunkte voneinander ausgegeben.

Weiter kann unter „Options“ festgelegt werden, wie Fälle mit fehlenden Werten zu handhaben sind (fallweiser oder paarweiser Ausschluß der Werte).

8 Validierung von Lösungen durch Vergleich von Verfahrensalternativen, Kreuzvalidierung, Diskriminanzanalysen und Monte-Carlo-Studien

Die Abhängigkeit der Clusteranalysen von Verfahrensparametern erfordert, wie wir an verschiedenen Stellen betont haben, zusätzliche Entscheidungen bei der Durchführung der Analysen. Dabei kann man auch den Vergleich zwischen unterschiedlichen Resultaten bei unterschiedlichen Verfahren der Clusteranalyse nutzen. Ein Beispiel dafür ist das folgende allgemein empfohlene Vorgehen zur Kontrolle der Ergebnisse von hierarchisch-agglomerativen Verfahren.

Man finde zunächst eine Clusterlösung mit einem hierarchisch-agglomerativen Verfahren. Dann ordne man den (multivariaten) Mittelwerten, die die Clusterschwerpunkte bzgl. der Clustervariablen charakterisieren, die Objekte nach Maßgabe des kleinsten Abstands neu zu. Bei diesem Verfahren kann es neben Neuzuordnungen von Objekten, die zwischen den Clustern liegen, zu erheblichen Verschiebungen kommen, die auf die topologischen Eigenschaften des hierarchisch-agglomerativen Verfahrens zurückgehen. Diese Verschiebungen kann man dadurch dokumentieren, daß man sowohl für die ausgewählte hierarchische Lösung als auch für die durch Neuzuordnung modifizierten Cluster die Clusterzugehörigkeit speichert und deren gemeinsame Häufigkeitsverteilung in einer Kreuztabelle betrachtet (Abbildung 9). Wenn die Cluster der hierarchischen Lösung die Zeilen, und die Cluster nach der Neuzuordnung die Spalten bilden, dann läßt sich diese Tabelle wie eine Überströmmatrix lesen: In den Feldern neben der Hauptdiagonale befinden sich die neuzugeordneten Objekte. Man findet dann nicht selten stark besetzte Zellen neben der Diagonalen vor. Dies ist die Folge der raumverzerrenden topologischen Eigenschaften des hierarchisch-agglomerativen Verfahrens. Ist man aber an „raumerhaltend“ gebildeten Clustern interessiert, so sollte man prüfen, ob eine Lösung mit mehr, aber dafür weniger stark aggregierten Clustern zu einer geringeren Zahl von Verschiebungen führt.

Abbildung 9: Kreuzvalidierung der Allbus-4-Clusterlösung

		K-means				Total
		1	2	3	4	
Ward	1	1556	105		39	1700
	2	26	404		4	434
	3		1	208	3	212
	4			20	591	611
Total		1582	510	228	637	2957

Stabilität der Lösungen bei variierenden Verfahren kann man als ein Validitätskriterium für Resultate ansehen. Stabilität kann aber auch von den Daten selbst abhängen. Es wird daher mitunter empfohlen (Gordon 1999), an zufällig gezogenen Substichproben gewonnene Lösungen mit dem vollständigen Datensatz zu bestätigen.

Eine weitere Möglichkeit der Validitätsprüfung durch Prüfung der Stabilität sind Monte-Carlo-Studien. ClustanGraphics bietet mit dem Modul „FocalPoint“ Monte-Carlo-Verfahren an, bei denen durch Variation der Startgruppierungen und/oder durch zufällige Reihenfolgen der Objekte Lösungen bestimmt und mit den Häufigkeiten ihres Auftretens in einer Reihe von Durchläufen aufgelistet werden. Die Anzahl der Lösungen und die Häufigkeiten, mit der sie in einer Monte-Carlo-Studie reproduziert werden, können wichtige Hinweise auf die Eindeutigkeit der Clusterstruktur in den Daten geben. Insbesondere kann so die - unbekannte - Anzahl sämtlicher Minimal-Distanz-Lösungen in einer hinreichend langen Reihe von Durchläufen bestimmt werden.

ClustanGraphics bietet für Startgruppierungen neben rein zufälligen oder fest vorgegebenen Startzentren, wie speziell ausgewählte Objekte, auch Lösungen aus hierarchisch-agglomerativen Verfahren an, sowie die Definition von Startgruppierungen mit Hilfe besonders kompakter Cluster, sogenannter Cliquen. Schließlich können Startgruppierungen als „Zellen“ von Kontingenztabellen festgelegt werden, die eine erste inhaltliche Klassifikation darstellen.

Die fehlende konfirmatorische Komponente der Clusteranalyse wird häufig dadurch ersetzt, daß die Gruppierung, die man als Clusterstruktur ausgewählt hat, in einer Diskriminanzanalyse „getestet“ wird. Dies bedeutet, daß die bei der Clusteranalyse verwendeten Variablen daraufhin überprüft werden, ob sich mit ihnen im Sinne der Diskriminanzanalyse die Zugehörigkeit zu den Clustern vorhersagen läßt bzw. welche Trefferhäufigkeit dabei erreicht wird. Der Nutzen einer solchen Bestätigung erscheint uns allerdings begrenzt, da die gleichen Variablen verwendet werden wie bei der Clusteranalyse selbst. Dadurch ist eine „Bestätigung“ durch eine Diskriminanzanalyse etwas tautologisch.

Literatur

Bacher, J. (1994). *Clusteranalyse*. Oldenbourg, München.

Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2000). *Multivariate Analysemethoden*. Berlin: Springer.

Bailey, K. D. (1994). *Typologies and Taxonomies: An Introduction to Classification Techniques.* SAGE, Thousand Oaks.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler*. Springer, Berlin.

Feller, W. (1968). An Indroduction to Probability Theory and 1st Applications. Vol. I (3rd ed.), 160-161. New York: Wiley.

Gordon, A. D. (1999): *Classification*. Boca Raton, Chapman & Hall/CRC.

Hartmann P., Neugewöhner, U. (1999): *Lebensstilforschung und Publikumssegmentierung*. Media-Perspektiven 10, S. 531-539.

Kaufman, L., Rousseeuw, P. J. (1990): *Finding Groups in Data*. New York, Wiley.

Kaufmann, H.; Pape, H. (1984): *Clusteranalyse*. In Fahrmeir, L.; Hamerle, A. (Hrsg.): Multivariate statistische Verfahren. Berlin: de Gruyter.

SPSS Base 8.0, Applications Guide. Chicago, IL: SPSS.

Wishart, D. (1999): *ClustanGraphics Primer*. A Guide to Cluster Analysis. Edinburgh: Clustan Limited.