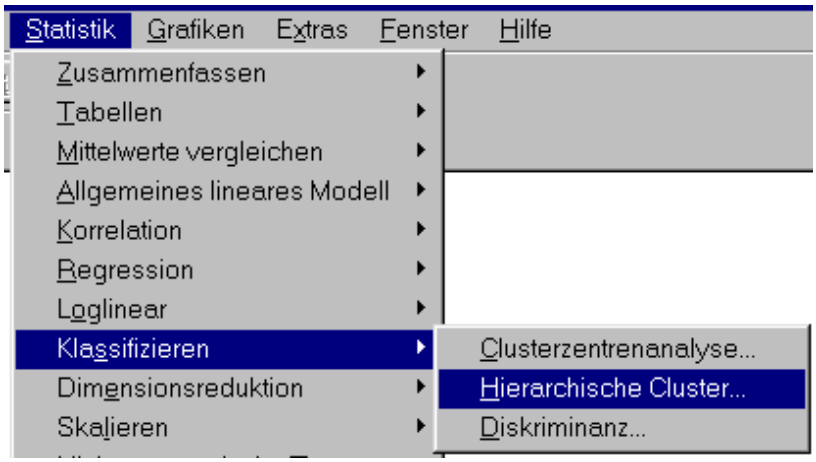
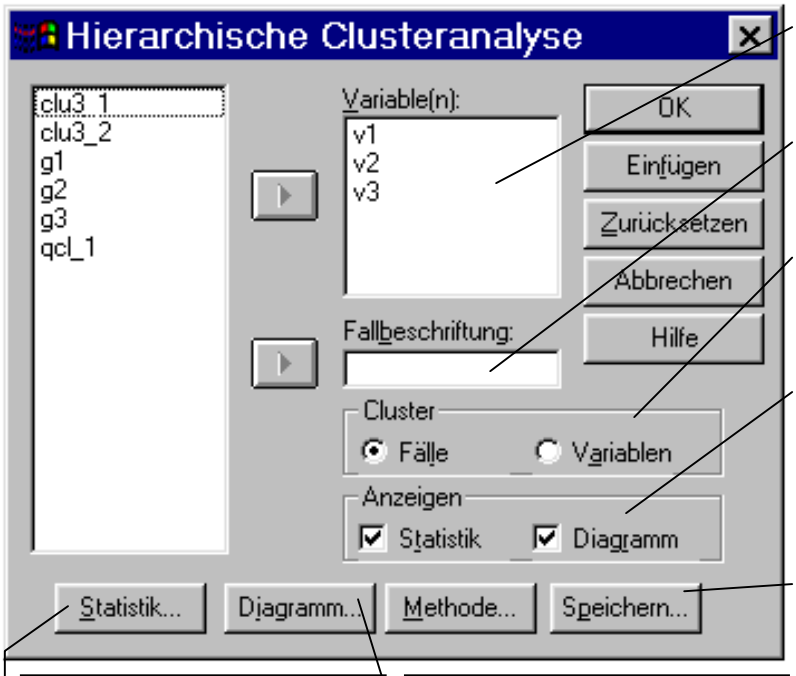
Hierarchische Clusteranalyse



Unter dem Menupunkt „Statistik"-„Klassifizieren" finden sich sowohl agglomerative („hierarchische") als auch partitionierende („Clusterzentren") Clusteranalyseverfahren. Da die hierarchische Analyse das Verfahren der Wahl ist, wenn man noch *keine* Vorstellung von der Zahl der Cluster hat, wird sie hier als erstes vorgestellt.

Zuerst müssen die Variablen ausgewählt werden, welche die Informationen über die Merkmale der zu clusternden Fälle enthalten.



Die Merkmalsvariablen, auf deren Basis geclustert wird.

Variable zur Identifikation der Fälle (z.B. Vpn-Nr.).

Auswahl ob Fälle (Standard, z.B. Personen) oder Variablen (z.B. Items) geclustert werden sollen.

Hier kann die Ausgabe von Statistiken oder Diagrammen unterdrückt werden. Das Ausschalten der „Diagramm"-Option ist, soweit man nicht *wirklich* ein Dendrogramm will, empfehlenswert!

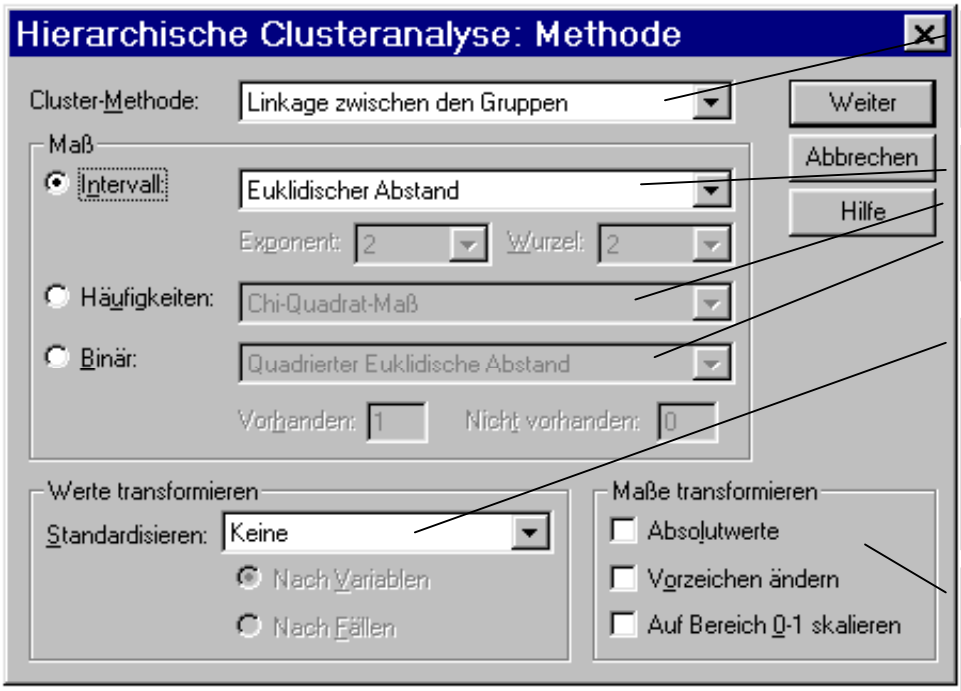
Mit „speichern" können die Clusterzugehörigkeiten auf verschiedenen Stufen des Clusterprozesses (z.B. für eine 3-Cluster-Lösung) als Variablen gespeichert werden.

Hier kann u.a. die Distanzmatrix der Fälle und die Zuordnung zu den Clustern auf verschiedenen Stufen des Clusterprozesses angefordert werden.

Hier können Dendrogramm und Eiszapfendiagramm ein- bzw. ausgeschaltet werden.

Im Anschluß muß unter „Methode" das Skalenniveau, das Distanzmaß und die Cluster-Methode gewählt werden. Da die Standardeinstellungen nur eine Möglichkeit unter einen großen Vielzahl darstellen, sollte dieses Optionsfeld auf keinen Fall ausgelassen werden.

Hierarchische Clusteranalyse: Methode



Hier wird das Kriterium gewählt, nachdem die Cluster fusioniert werden. Zu den Verfahren s.u.

Je nach Skalenniveau stehen unterschiedliche Distanzmaße zur Auswahl. Eine Liste der Möglichen Maße s.u.

Hier können die Datenwerte *vor* der Berechnung der Distanzen sowohl für Fälle als auch für Werte standardisiert werden. Dies ist vor allem bei stark unterschiedlichen Skalen der Merkmalsvariablen sinnvoll.

Über die Gruppe Maße transformieren können die Werte, die durch das Distanzmaß erstellt wurden, transformiert werden. Sie werden *nach* der Berechnung des Distanzmaßes angewendet.

Hierarchische Clusteranalyse: Cluster-Methoden

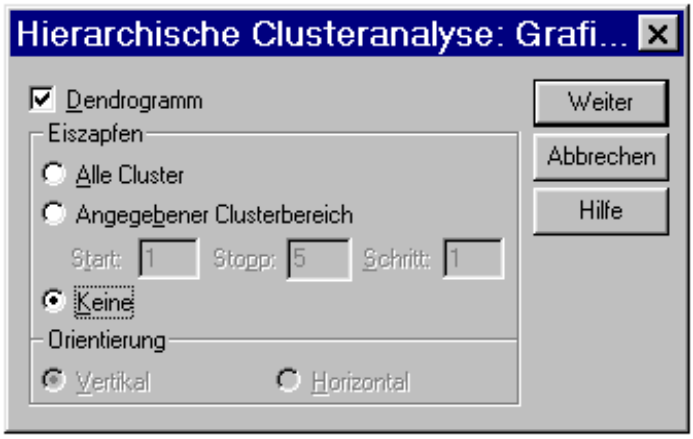
Hat man keine Gründe für eine bestimmte Wahl unter den möglichen Clusterverfahren, ist die Standardeinstellung („Linkage zwischen den Gruppen“) oder das Ward-Verfahren ein guter Ausgangspunkt (Bei Ward sollte als Distanzmaß die quadrierte euklidische Distanz gewählt werden). Eine wiederholte Analyse mit unterschiedlichen Verfahren kann auch einer exploratorischen Analyse der Stabilität der Ergebnisse dienen.

Listen mit den SPSS-Hilfe-Beschreibungen der Clustermethoden und Distanzmaße finden sich auf den beiden letzten Seiten.

Hierarchische Clusteranalyse: Ausgabe

Die Standardausgabe von „Hierarchische Clusteranalyse“ besteht aus folgenden Punkten im Ausgabe-Navigator:

1) Titel: „Distanzen“
2) Anmerkungen (werden nur bei Doppelklicken auf den entsprechenden Punkt im Navigator angezeigt; enthalten u.a. die SPSS-Syntax der durchgeführten Analyse)
3) Verarbeitete Fälle: Zahl der verarbeiteten und fehlenden Fälle bei der Erstellung der Proximitäts- bzw. Distanzmatrix.
4) Eigentliche Analyse; Titel je nach Clustermethode, z.B. „Ward-Linkage“)
5) Zuordnungsübersicht: Hier wird für jeden Fusionsschritt aufgelistet, welche Fälle oder Cluster zusammengeführt wurden. Unter „Koeffizient“ wird die Distanzen zwischen den Fällen oder Clustern, die kombiniert wurden, aufgeführt (Diese Größe ist abhängig von Distanzmaß und Clustermethode). Sprünge im Koeffizientenverlauf können bei der Suche nach einer sinnvollen Clusterzahl hilfreich sein.
6) Vertikales Eiszapfendiagramm: Außer bei kleiner Fallzahl ist diese Ausgabe, die eine grafische Veranschaulichung der Fusionsschritte darstellen soll, unübersichtlich und wenig hilfreich. Für eine grafische Veranschaulichung sollte unter im Optionsfeld „Diagramm“ „Dendrogramm“ gewählt und bei „Eiszapfen“ „keine“ gewählt sein (s. Abb.). Ansonsten empfiehlt sich, im Startfeld der Analyse (s.o.) die „Diagramm“-Option auszuschalten.



Einstellungen im Dialogfeld „Diagramm“ zum Wählen des Dendrogramms und Ausschalten der Eiszapfen.

Hierarchische Clusteranalyse: Cluster-Methoden

Auszug aus der SPSS-Online-Hilfe:

Linkage zwischen den Gruppen: Kombiniert Cluster, um die Durchschnittsdistanz zwischen allen Itempaaren zu verkleinern, in denen ein Teil des Paares aus jeweils einem Cluster stammt. Diese Methode verwendet Informationen über alle Distanzpaare, nicht nur das nächstgelegene oder das entfernteste.

Linkage innerhalb der Gruppen: Kombiniert Cluster auf die Art, daß die Durchschnittsdistanz zwischen allen Items innerhalb des entstandenen Clusters so klein wie möglich ist. Die Distanz zwischen zwei Clustern wird dann als Durchschnittswert aller Distanzen zwischen allen möglichen Fallpaaren des Clusters genommen, der entstehen würde, wenn sie kombiniert wären.

Nächstgelegener Nachbar: Diese Methode kombiniert zunächst die beiden Items mit der kleinsten Distanz oder mit der größten Ähnlichkeit. Die Distanz zwischen dem neuen Cluster und einzelnen Fällen wird sodann als die Mindestdistanz zwischen einem einzelnen Fall und einem Fall im Cluster berechnet. Die Distanzen zwischen Items, die nicht verbunden wurden, bleiben unverändert. Bei jedem Schritt wird die Distanz zwischen zwei Clustern als Distanz zwischen den beiden Punkten genommen, die am engsten beieinander liegen.

Entferntester Nachbar: Die Distanz zwischen zwei Clustern wird als die Distanz zwischen den zwei Punkten berechnet, die am weitesten auseinanderliegen.

Zentroid-Clustering: Berechnet die Distanz zwischen zwei Clustern als Distanz zwischen den Mittelwerten für alle Items. Die Distanz, in der Cluster kombiniert werden, kann von einem zum nächsten Schritt abnehmen.

Median-Clustering: Die beiden kombinierten Cluster werden bei der Berechnung des Zentroidwerts gleich gewichtet; dabei spielt es keine Rolle, wieviele Fälle jeder enthält. Auf diese Weise können kleine Gruppen bei der Charakterisierung größerer Cluster, in die sie integriert werden, gleich große Effekte haben.

Wards Methode: Mit dieser Methode werden zuerst die Mittelwerte für jede Variable innerhalb der einzelnen Cluster berechnet. Anschließend wird für jeden Fall die Quadrierte Euklidische Distanz zu den Cluster-Mittelwerten berechnet. Diese Distanzen werden für alle Fälle summiert. Bei jedem Schritt sind die beiden zusammengeführten Cluster diejenigen, die die geringste Zunahme in der Gesamtsumme der quadrierten Distanzen innerhalb der Gruppen ergeben.

Hierarchische Clusteranalyse: Distanzmaße

Auszug aus der SPSS-Online-Hilfe

Maße für Intervalldaten

Euklidische Distanz. Die Quadratwurzel der Summe der quadrierten Differenzen zwischen den Werten der Einträge. Dies ist die Voreinstellung für Intervalldaten.
Quadrierte euklidische Distanz. Die Summe der quadrierten Differenzen zwischen den Werten der Einträge.
Pearson-Korrelation. Die Produktmomentkorrelation zwischen zwei Vektoren von Werten.
Kosinus. Der Kosinus des Winkels zwischen zwei Wertevektoren.
Tschebyscheff. Die größte absolute Differenz zwischen den Werten der Einträge. („Supremum-Metrik", entspricht Minkowski mit $r = \infty$)
Block. Die Summe der absoluten Differenzen zwischen den Werten der Einträge. Auch als Manhattan-Distanz bekannt.
Minkowski. Die p-te Wurzel der Summe der absoluten Differenzen zwischen den Werten der Einträge zur p-ten Potenz.
Benutzerdefiniert. Die r-te Wurzel der Summe der absoluten Differenzen zwischen den Werten zur p-ten Potenz.

Unähnlichkeitsmaße für binäre Daten (Auszug):

Euklidische Distanz. Wird als SQRT(b+c) aus einer Vier-Weg-Tabelle berechnet, wobei b und c diagonale Zellen für Fälle darstellen, die in einem Objekt vorhanden sind und im anderen fehlen.
Quadrierte Euklidische Distanz. Wird aus der Anzahl der nicht miteinander harmonierenden Fälle berechnet. Ihr Minimalwert ist 0, und es ist keine Obergrenze vorhanden.
Größendifferenz. Ein Index der Asymmetrie. Der Bereich liegt zwischen 0 und 1.
Musterdifferenz. Unähnlichkeitsmaß für binäre Daten in einem Bereich von 0 bis 1. Wird als bc/(N**2) aus einer Vier-Weg-Tabelle berechnet, wobei b und c diagonale Zellen für Fälle darstellen, die in einem Objekt vorhanden sind und im anderen fehlen, und n der Gesamtzahl der Beobachtungen entspricht.
Varianz. Wird als (b+c)/4n aus einer Vier-Weg-Tabelle berechnet, wobei b und c diagonale Zellen für Fälle darstellen, die in einem Objekt vorhanden sind und im anderen fehlen, und n der Gesamtzahl der Beobachtungen entspricht. Der Bereich liegt zwischen 0 und 1.
Streuung. Dieser Ähnlichkeitsindex weist einen Bereich von -1 bis 1 auf.
Form. Dieses Distanzmaß liegt im Bereich von 0 bis 1. Asymmetrische Nicht-Übereinstimmungen werden bestraft.
Einfache Übereinstimmung. Dies ist das Verhältnis von Übereinstimmungen zur Gesamtzahl der Werte. Übereinstimmungen und Nicht-Übereinstimmungen werden gleich gewichtet.
Phi-4-Punkt-Korrelation. Dieser Index ist die binäre Entsprechung zum Korrelationskoeffizienten nach Pearson. Der Bereich liegt zwischen -1 und 1.
Lambda. Dieser Index ist der Lambda-Wert nach Goodman und Kruskal. Entspricht der proportionalen Fehlerreduktion, wobei ein Objekt verwendet wird, um das andere vorauszusagen (Voraussage in beiden Richtungen). Die Werte liegen im Bereich von 0 bis 1.
Würfel. Bei diesem Index werden gemeinsam fehlende Größen aus der Betrachtung ausgeschlossen und Übereinstimmungen doppelt gewichtet. Auch als Czekanowski- oder Sorensen- Maß bekannt.
Hamann. Dieser Index entspricht der Anzahl der Übereinstimmungen abzüglich der Anzahl der Nicht-Übereinstimmungen, geteilt durch die Gesamtanzahl der Einträge. Der Bereich liegt zwischen -1 und 1.
Jaccard. Bei diesem Index werden gemeinsam fehlende Größen aus der Betrachtung ausgeschlossen. Übereinstimmungen und Nicht-Übereinstimmungen werden gleich gewichtet. Auch als der Ähnlichkeitsquotient bekannt.

Unähnlichkeitsmaße für Häufigkeitsdaten:

Chi-Quadrat-Maß. Dieses Maß basiert auf dem Chi-Quadrat-Test der Gleichheit für zwei Häufigkeitsmengen. Dies ist die Voreinstellung für Häufigkeitsdaten.
Phi-Quadrat-Maß. Diese Maß entspricht dem Chi-Quadrat-Maß, das durch die Quadratwurzel der kombinierten Häufigkeit normalisiert wurde.