20 Clusteranalyse

Hier wird die Anwendung der Two-Step-Clusteranalyse in SPSS-Version 17 und früher behandelt. Der Text entspricht dem in der 7. Auflage. Ab SPSS 18 sind sowohl die Dialogboxen als auch die Ergebnisausgabe neu gestaltet worden. In der 8. Auflage ist dies dargestellt.

20.2 Praktische Anwendung

20.2.3 Anwendungsbeispiel zur Two-Step-Clusteranalyse

Es sollen Kunden einer Telefongesellschaft in Kundengruppen segmentiert werden (Datei TELCOM.SAV).[1] Als Variable sind sowohl metrische (die Dauer der Gespräche in Sekunden von drei Gesprächsklassen: ORT, FERN, INTERNAT) als auch kategoriale Variable (TARIF_ORT mit den beiden Ortstarifen Pauschal und Zeitabhängig sowie TARIF_FERN mit den beiden Ferngesprächstarifen Normal und Rabatt) vorhanden. Die Gesprächsdauer hat für alle drei Gesprächsklassen eine linkssteile Verteilung. Um die Modellvoraussetzungen der Normalverteilung annähernd zu erfüllen, werden diese Variablen logarithmiert (Name der Variable: Voranstellen von Lg).[2]

Nach Öffnen der Datei TELCOM.SAV gehen Sie wie folgt vor:

- ▹ Wählen Sie die Befehlsfolge „Analysieren“, „Klassifizieren ▹“, “Two-Step-Clusteranalyse...“. Es öffnet sich die in Abb. 20.14 dargestellte Dialogbox.
- ▹ Übertragen Sie die Variablen TARIF_ORT, TARIF_FERN aus der Quellvariablenliste in das Feld „Kategoriale Variablen“ und die Variablen LGORT, LGFERN und LGINTERNAT in das Feld „Stetige Variablen“. Als Distanzmaß steht die Option „Euklidisch“ nicht zur Verfügung, da hier für die Clusteranalyse gemischte (kategoriale und metrische) Variablen genutzt werden.
- ▹ Im Feld „Anzahl stetiger Variablen“ werden 3 Variablen als „Zu standardisieren“ und 0 als „Als standardisiert angenommen“ angezeigt. Damit ist die z-Transformation gemeint (⇨ Kap. 8.5). Man standardisiert, um stetige Variable in ihrer Messskala vergleichbar zu machen.
- ▹ Im Feld „Anzahl der Cluster“ besteht die Wahl zwischen „Automatisch ermitteln“ und „Feste Anzahl angeben“. Meistens wird man die automatische Bestimmung der Clusteranzahl bevorzugen. Dafür gibt man eine Obergrenze an (hier: 11). In der zweiten Stufe des (agglomerativen hierarchischen) Clusterprozesses werden daher 11 Clusterlösungen (11 bis 1 Cluster) berechnet.
- ▹ Im Feld „Cluster-Kriterien“ wird ein Auswahlmaß zum Bestimmen der Clusteranzahl gewählt. Die Voreinstellung BIC wird man i.d.R. übernehmen.

---

1 Die Datei beruht auf Daten von SPSS Training (SPSS GmbH Software München).

2 Da der Logarithmus für 0 nicht definiert ist, wird beim Logarithmieren 0 durch 1 ersetzt.

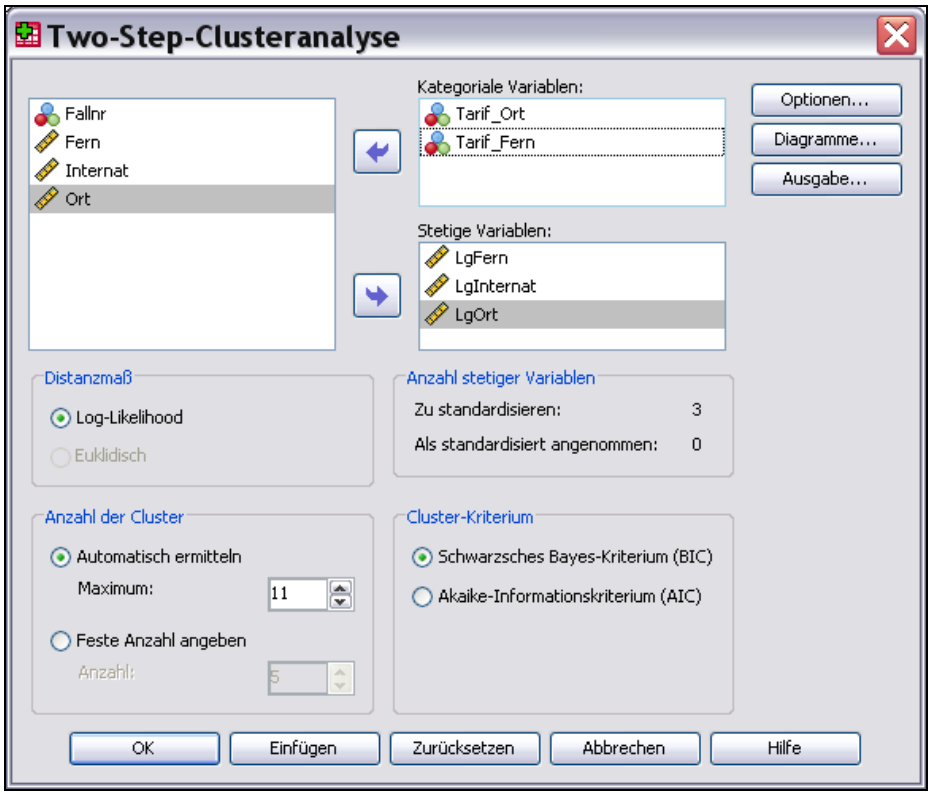

Abb. 20.14. Dialogbox „Two-Step-Clusteranalyse“

Wahlmöglichkeiten. Damit die Clusteranalyse mit der Berechnung startet, muss man in den Dialogboxen, die mit der Schaltfläche „Ausgabe“ oder der Schaltfläche „Diagramme“ in der Hauptdialogbox geöffnet werden, mindestens eines der Auswahlangebote anfordern.

① *Optionen*. Nach Klicken auf die Schaltfläche „Optionen“ (⇨ Abb. 20.14) öffnet sich die in Abb. 20.15 dargestellte Dialogbox „Two-Step-Clusteranalyse: Optionen“. Mit der Option „Rauschverarbeitung verwenden kann man anfordern, ob in der ersten Cluster-Stufe Ausreißer (noise, Rauschen in den Daten) ausgesondert werden sollen (⇨ Abb. 20.4). Es ist möglich, den voreingestellten Wert von 25 Prozent zu ändern. Mit einer Angabe x % wird festgelegt, dass maximal x % der Fälle im Blattknoten des CF-Baums mit der größten Fallzahl nicht in die Cluster einbezogen werden. Wir wählen die Voreinstellung. Im Feld „Speicherzuweisung“ kann man die voreingestellten 64 MB Speicherzuweisung für den Clusteralgorithmus verändern.

Im Feld „Standardisierung von stetigen Variablen“ kann man die Variablen, die schon standardisiert sind, von der Variablenliste „Zu standardisieren:“ in die Variablenliste „Als standardisiert angenommen:“ übertragen.

Klicken auf die Schaltfläche „Erweitert >>“ ergänzt die Dialogbox in Abb. 20.15 um einen Bereich am unteren Rand. Hier kann man die Voreinstellungen zum Aufbau des CF-Baums verändern. „Schwellenwert für anfängliche Distanzänderung“ bezieht sich auf den zu Beginn der Baumerstellung anfänglichen Schwellenwert, der darüber befindet, ob ein den Baum durchlaufender Fall in

einem ähnlichen Blatt landet oder ob der Fall sich zu stark von den Fällen in einem Blatt unterscheidet, so dass ein neuer Blattknoten gebildet werden muss. „Höchstzahl der Verzweigungen (pro Blattknoten)“ bezieht sich auf die Knoten in den Ebenen des Baums. Voreingestellt ist, dass von jedem Elternknoten 8 Kinderknoten abgehen. „Maximale Baumtiefe (Ebenen)“ bezieht sich auf die Höchstzahl der Knotenebenen des Baums. Voreingestellt ist 3 (⇨ Abb. 20.15). Für die jeweils gewählten Angaben wird die Anzahl der Knoten des CF-Baums angezeigt.[3]

Die Option „Aktualisierung des Clustermodells“ ermöglicht es, ein in früheren Analysen erzeugtes Clustermodell zu importieren und mit neuen Daten (die natürlich aus der gleichen Grundgesamtheit stammen müssen) zu aktualisieren. Die Eingabedatei enthält den CF-Baum im XML-Format (⇨ näheres im Hilfesystem). Im XML-Format kann ein Clusteranalysemodell gespeichert werden (siehe unten).

② *Diagramme*. Nach Klicken auf die Schaltfläche „Diagramme“ (⇨ Abb. 20.14) öffnet sich die in Abb. 20.16 dargestellte Dialogbox „Two-Step-Clusteranalyse: Diagramme“.

Die Option „Prozentdiagramme in Cluster“ erzeugt für jede kategoriale Variable ein gruppiertes Balkendiagramm. Dieses zeigt in Balkenform die prozentualen Häufigkeiten der Variablenkategorien für jedes der Cluster in der 5-Clusterlösung (und auch für alle Fälle sowie für die Ausreißer). Abb. 20.17 links zeigt das Diagramm für die Variable TARIF_FERN. Durch Vergleiche der prozentualen Häufigkeiten in den Clustern mit denen für alle Fälle („Gesamt“) kann man sehr gut erkennen, dass in den Clustern 1 und 5 Telefonkunden enthalten sind, die den Ferngesprächstarif „Rabatt“ und in den Clustern 3 und 4 den Ferngesprächstarif „Normal“ haben.

Für jede metrische (stetige) Variable wird ein Fehlerbalkendiagramm (⇨ Kap. 27.3) erstellt, dass für jedes Cluster (und auch für die Ausreißer) einen Fehlerbalken für das arithmetische Mittel mit einem x%-tigem Konfidenzintervall zeigt. Als Bezugslinie dient der Mittelwert für alle Fälle. In Abb. 20.17 rechts ist ein Diagramm für die Variable LGFERN zu sehen. Es wird deutlich, dass im Cluster 2 der Mittelwert von LGFERN sehr niedrig ist. In diesem Cluster gibt es relativ wenig Kunden die Auslandsgespräche führen.

Die Option „Gestapeltes Kreisdiagramm“ zeigt ein Kreisdiagramm, das die prozentuale Verteilung der Fälle auf die Cluster darstellt.

Die Optionen im Feld „Wichtigkeitsdiagramm für Variablen“ erzeugen mehrere Diagramme, die die Bedeutung der einzelnen Variable für die einzelnen Cluster zum Ausdruck bringen. Dazu werden die Ergebnisse von statistischen Tests in Form von Balkendiagrammen aufbereitet.

[3] Siehe dazu die Erläuterungen zur Two-Step Clusteranalyse in Kap. 20.1.

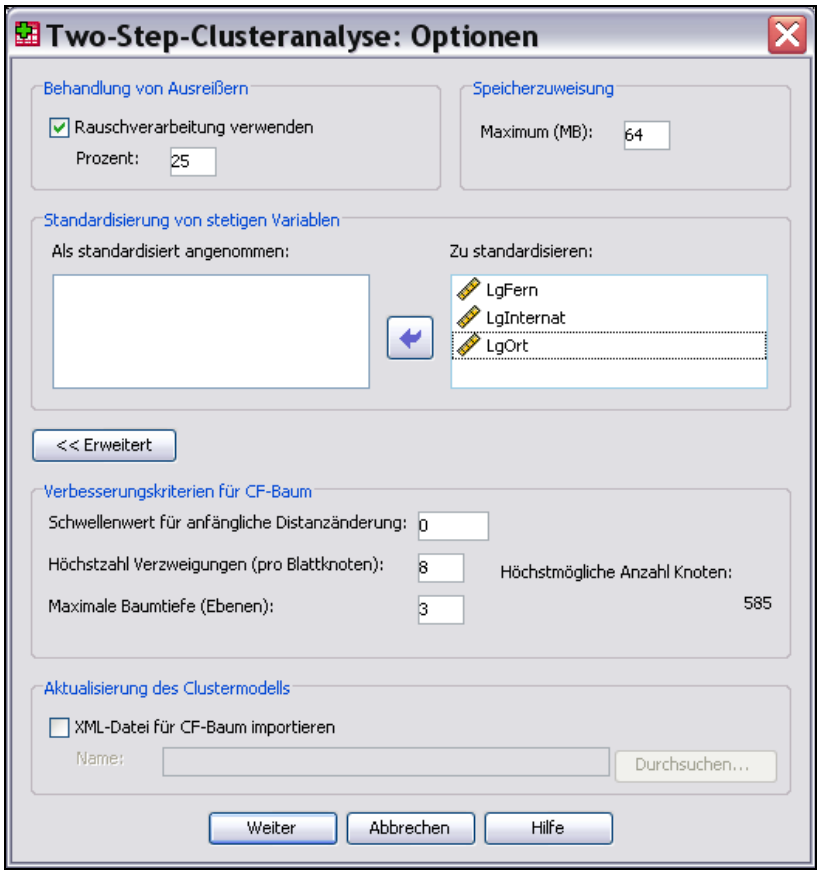

Abb. 20.15. Dialogbox „Two-Step-Clusteranalyse: Optionen“

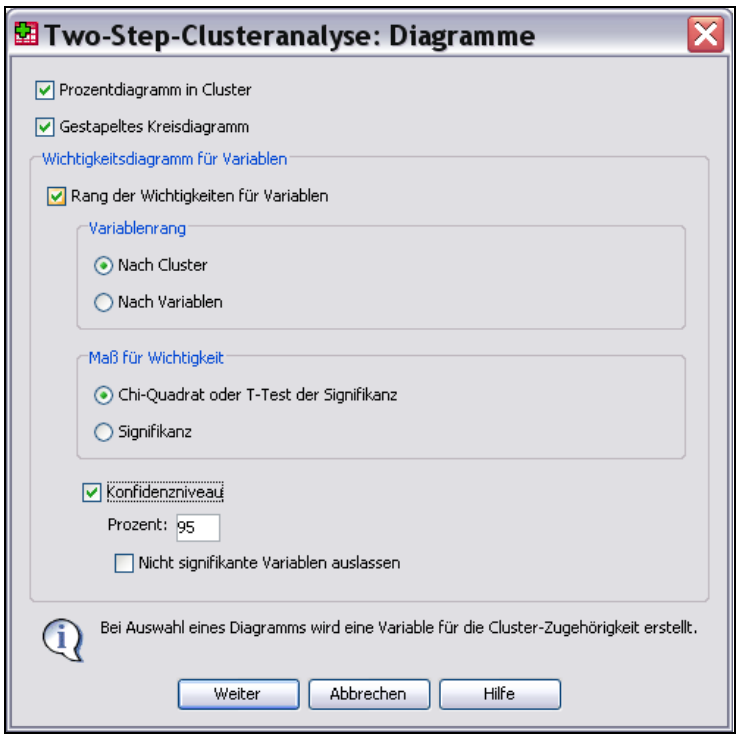

Abb. 20.16. Dialogbox „Two-Step-Clusteranalyse: Diagramme

Bei den kategorialen Variablen handelt es sich um einen Chi-Quadrat-Anpassungstest (⇨ Kap. 26.2.1). Für jede kategoriale Variable wird geprüft, ob die Häufigkeitsverteilung der Variable in einem Cluster sich signifikant von der Häufigkeitsverteilung für alle Fälle unterscheidet. Abb. 20.18 links zeigt das grafisch aufbereitete Testergebnis für die Variable TARIF_FERN (für die Option „Nach Cluster“ im Feld „Variablenrang“ und Wahl der Option „Chi-Quadrat oder t-Test der Signifikanz“ im Feld „Maß der Wichtigkeit“). In einer senkrechten unterbrochenen Linie ist der kritische Chi-Quadrattestwert (für ein vorzugebendes Signifikanzniveau) zu sehen. Je mehr der Balken eines Clusters diese Linie überschreitet, umso wichtiger ist die Variable für die Kunden in dem Cluster. Es wird deutlich, dass der Ferngesprächstarif für Kunden in allen Clustern mit Ausnahme des Clusters 2 eine wichtige Rolle spielt. In Abb. 20.17 rechts wurde schon sichtbar, dass im Cluster 2 nur relativ wenige Kunden Ferngespräche führen.

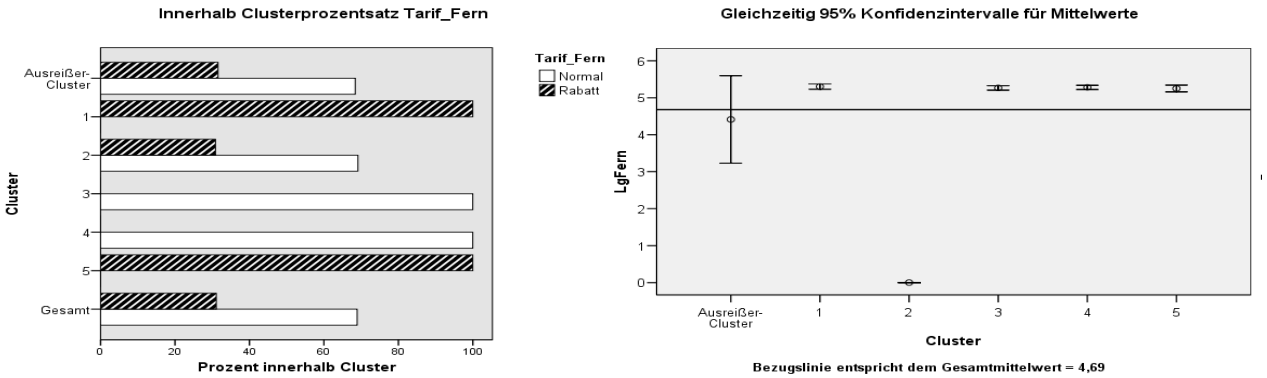

Abb. 20.17. Gruppiertes Balkendiagramm zur Darstellung der prozentualen Cluster-Häufigkeiten für die Variable TARIF_FERN (links) und Fehlerbalkendiagramm zur Darstellung von Cluster-Mittelwerten für die Variable LGFERN

Für jede metrische Variable wird per t-Test (⇨ Kap. 13.4) geprüft, ob der Mittelwert der Variable für Kunden in einem Cluster sich vom Mittelwert aller Kunden unterscheidet.[4] In Abb. 20.18 rechts wird das grafisch aufbereitete Testergebnis für die Variable LGFERN gezeigt (für die Option „Nach Cluster“ im Feld „Variablenrang“ und Wahl der Option „Chi-Quadrat oder t-Test der Signifikanz“ im Feld „Maß der Wichtigkeit“). Auch hier ist der kritische t-Wert für den Test (für ein vorzugebendes Signifikanzniveau) als senkrechte unterbrochene Linie dargestellt. Man sieht, dass es für Kunden im Cluster 2 keinen signifikanten Unterschied zwischen den Mittelwerten gibt. Kunden im Cluster 2 haben keine Präferenzen für Ferngespräche.

Mit der Option „Variablenrang“ wird festgelegt, ob die Diagramme für jedes Cluster („Nach Cluster“) oder für jede Variable („“Nach Variable) erstellt werden sollen.

Die Option „Maß für Wichtigkeit“ legt fest, welches Maß für die Wichtigkeit der Variablen in den Diagrammen dargestellt werden soll. Zur Auswahl stehen die Optionen „Chi-Quadrat oder t-Test der Signifikanz“ (diese Option wurde für das Bei-

4 Die Korrektur nach Bonferroni wird dabei berücksichtigt (⇨ Kap. 14.3).

spiel in Abb. 20.18 gewählt) und „Signifikanz“. Bei Wahl von Signifikanz wird der Wert –(log10(Wahrscheinlichkeit) abgebildet.[5]

Mit der Option „Konfidenzniveau.“ kann man das Signifikanzniveau für die Tests angegeben.

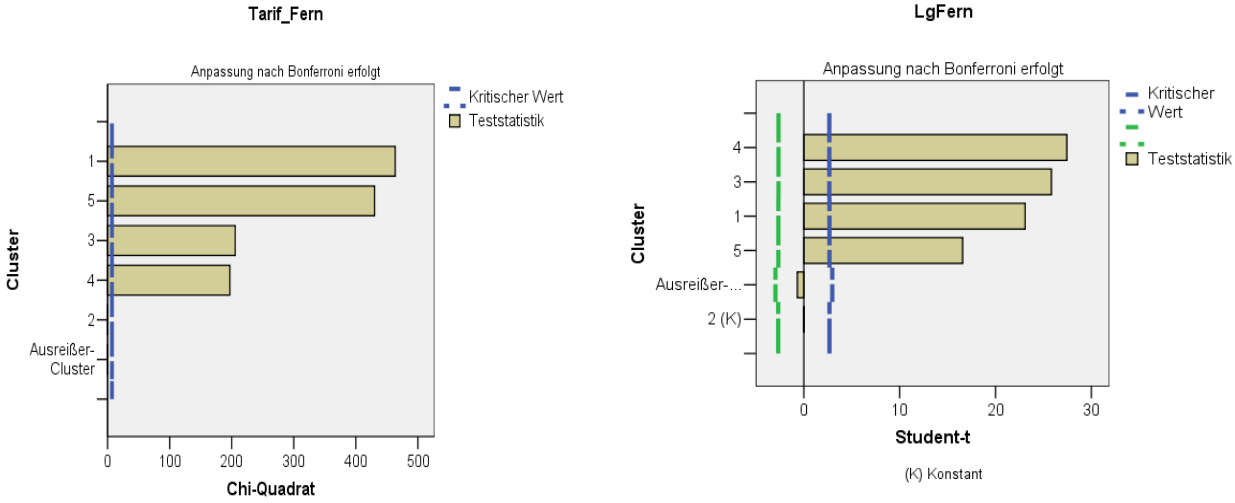

Abb. 20.18. Grafische Darstellung des Ergebnisses eines Chi-Quadrat-Anpassungstests für die Variable TARIF_FERN (links) und des Ergebnisses eines t-Tests für die Variable LGFERN (rechts)

Wenn die Option „Nicht signifikante Variablen auslassen“ gewählt wird, so werden Variablen, die für das angegebene Konfidenzniveau nicht signifikant sind, in den Wichtigkeitsdiagrammen nicht angezeigt.

③ *Ausgabe*. Nach Klicken auf die Schaltfläche „Ausgabe“ (⇨ Abb. 20.14) öffnet sich die in Abb. 20.19 gezeigte Dialogbox „Two-Step-Clusteranalyse: Ausgabe“. Mit der Option „Deskriptive Statistik nach Cluster“ im Feld „Statistik“ kann man sich für die kategorialen Variablen die absoluten und prozentualen Häufigkeiten und für die metrischen Variablen Mittelwerte und Standardabweichungen in einer Auswertung für die Cluster in das Ausgabefenster geben lassen. Anhand dieser Tabellen wird ein Profil der Cluster sichtbar. Die Option „Cluster-Häufigkeiten“ erzeugt als Output die Verteilung der Fälle auf die Cluster.

Die Option „Informationskriterium (BIC oder AIC) gibt bei Wahl der automatischen Ermittlung der Clusteranzahl für die Sequenz von Clusterlösungen das Auswahlkriterium BIC oder AIC (je nach Wahl in der Hauptdialogbox, ⇨ Abb. 20.14) und weitere zur Bestimmung der optimalen Clusteranzahl genutzte Gütemaße an. In Tabelle 20.7 sind 11 Clusterlösungen (wie in Abb. 20.14 ge-

[5] Da die Werte der Prüfverteilungen (t- bzw. Chi-Quadrat-Verteilung) im Wertebereich nicht begrenzt sind (sie gehen von 0 bis unendlich) und des Weiteren die Testgrößenwerte t und Chi-Quadrat nicht vergleichbar sind, sind die Wichtigkeiten metrischer und kategorialer Variablen nur relativ zueinander zu beurteilen. Daher wird empfohlen, als einheitliches Maß für die Wichtigkeit die „Signifikanz“ zu wählen. Sowohl für den t- als auch den Chi-Quadrat-Test werden dann die Werte als $-\log_{10}$(Wahrscheinlichkeit) ausgegeben [⇨ unveröffentlichtes Manuskript von Dr. Mathias Glowatzki (SPSS GmbH Software München)].

wählt) mit jeweils 1 bis 11 Cluster zu sehen. Das kleinste BIC liegt mit 711,660 bei 9 Cluster. „BIC-Änderung“ in der dritten Spalte der Tabelle gibt die Veränderung von BIC (= ΔBIC) bei Übergang von einer Clusterlösung zur nächsten an. Bei dieser Clusterlösung mit dem kleinsten BIC wird auch das „Verhältnis der BIC-Änderungen“ (in Fußnote 11 mit R_1 bezeichnet) mit -13,830/-765,940 = 0,018 am niedrigsten. Ausgehend von dieser Lösung wird die optimale Clusteranzahl bei der Lösung mit dem größten „Verhältnis der Distanzmaße“ (in Fußnote 14 mit R_2 bezeichnet) gefunden. Bei der Clusterlösung mit 5 Cluster ist dieser Wert mit 2,348 am größten.[6]

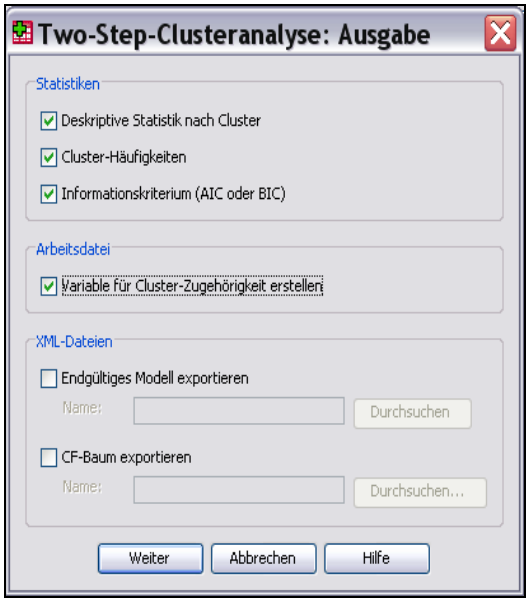

Abb. 20.19. Dialogbox „Two-Step-Clusteranalyse: Ausgabe“

Tabelle 20.7. Auswahlkriterien zur automatischen Auswahl der Clusteranzahl

Automatische Clusterbildung

Anzahl der Cluster	Bayes-Kriterium nach Schwarz (BIC)	BIC-Änderung[a]	Verhältnis der BIC-Änderungen[b]	Verhältnis der Distanzmaße[c]
1	2916,087			
2	2150,148	-765,940	1,000	1,148
3	1489,542	-660,606	,862	1,394
4	1030,630	-458,912	,599	2,001
5	827,938	-202,692	,265	2,348
6	772,163	-55,775	,073	1,374
7	746,079	-26,084	,034	1,074
8	725,490	-20,589	,027	1,101
9	711,660	-13,830	,018	1,368
10	715,885	4,226	-,006	1,455
11	735,438	19,553	-,026	1,787

a. Die Änderungen wurden von der vorherigen Anzahl an Clustern in der Tabelle übernommen.

b. Die Änderungsquoten sind relativ zu der Änderung an den beiden Cluster-Lösungen.

c. Die Quoten für die Distanzmaße beruhen auf der aktuellen Anzahl der Cluster im Vergleich zur vorherigen Anzahl der Cluster.

6 Siehe dazu die Erläuterungen zur Two-Step Clusteranalyse in Kap. 20.1.

Mit Wahl der Option „Variable für Clusterzugehörigkeit erstellen“ im Feld „Arbeitsdatei“ wird den Variablen der SPSS-Arbeitsdatei eine Variable hinzugefügt. Diese Variable enthält für jeden Fall die Clusterzugehörigkeitsnummer. Ausreißerfälle erhalten den Wert –1. Der Name dieser Variablen lautet TSC_ mit einer angehängten Zahl.

Im Feld „XML-Dateien“ gibt es zwei Optionen. Das endgültige Clustermodell und der CF-Baum sind zwei Arten von Ausgabedateien, die im XML-Format exportiert werden können. Der Name der Datei ist anzugeben. Der aktuelle Stand des Cluster-Baums kann gespeichert werden und später mit neuen Daten aktualisiert werden.