Autor: Dr. Ralf Gutfleisch, Stadt Frankfurt a. M., Bürgeramt, Statistik und Wahlen

Was ist eine Clusteranalyse, wann und wie wird sie angewendet?

Fragestellung

Drei Fragen stehen im Vordergrund dieser Einführung:

- Was ist eine Clusteranalyse?
- Wann wird eine Clusteranalyse angewendet?
- Wie wird eine Clusteranalyse angewendet?

Ziel ist es, auch Anfängern die Möglichkeit aufzuzeigen, eine Clusteranalyse zu rechnen. Die verschiedenen Verfahrensabläufe werden beschrieben und Prozesse erklärt, die meist im Programmhintergrund ablaufen. Um die geforderten Einstellungen vornehmen zu können, ist es notwendig, diese zu kennen. Handlungsempfehlungen werden gegeben, um Entscheidungen zu erleichtern, die an zentralen Stellen zu treffen sind.

Definition

Beim Clustern werden ähnliche Objekte zu Gruppen zusammengefasst

Die Clusteranalyse ist ein Gruppenbildungsverfahren. Das Ziel des Verfahrens besteht darin, ähnliche **Objekte** zu Gruppen **zusammenzufassen**. Wissenschaftlich exakt ausgedrückt lautet dies: Ziel der Clusteranalyse ist es, aus einer heterogenen Gesamtheit von Objekten homogene Teilmengen zu identifizieren. Wird die Etymologie des Wortstammes „Cluster“ betrachtet, wird die Verfahrensweise deutlich. Cluster, aus dem englischen Sprachraum kommend, lässt sich mit Haufen, Klumpen oder auch Ballung übersetzten. Also eine Zusammenballung von Objekten. Im Grimm’schen Wörterbuch findet man sogar das althochdeutsche Wort „Kluster“. Die Übersetzung hierfür: „**was dicht und dick zusammensitzt**“. So ist es auch nicht verwunderlich, wenn die Assoziation auf Sumo-Ringer fällt:

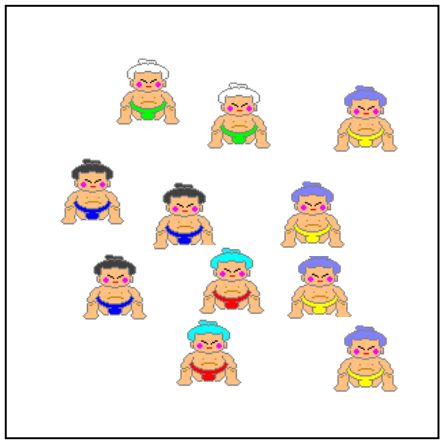


Abbildung 1: Sumo-Ringer sortiert

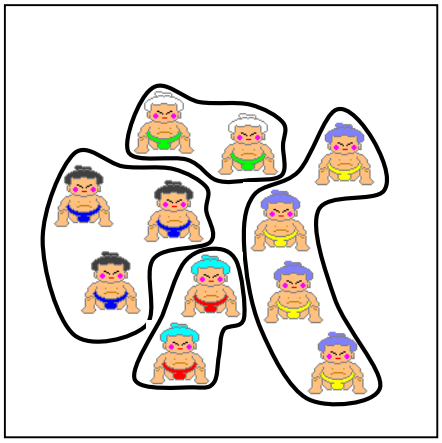


Abbildung 2: Sumo-Ringer gruppiert

Die Sumo-Ringer werden nach Haarfarbe und Schürze gruppiert

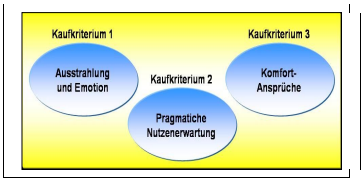
Betrachtet man die Sumo-Ringer in Abbildung 1, stellt man zunächst einen Haufen sitzender Kämpfer fest. Sie lassen sich nur schwer gruppieren, da gemeinsame Attribute auf Anhieb nur schwer erkennbar sind. Erst beim zweiten Blick wird deutlich, dass gruppenbildende Merkmale Haarfarbe und Schürze sind. Innerhalb einer Gruppe sind diese gleich eingefärbt, z.B. blaue Schürze und schwarze Haare. Damit unterscheidet sie sich auch von Anderen, z.B. den Weißhaarigen und Grünschürzigen.

Durch das Einkreisen in Abbildung 2 werden die vier Gruppen leicht erkennbar. Nach diesem einfachen Prinzip funktioniert auch die Clusteranalyse. Würde man dieses Beispiel clustern, wären die **zu gruppierenden Objekte** die Sumo-Ringer. Die **Merkmale unterschiedlicher Ausprägung** wären deren Haar- und Schürzenfarbe. Die Cluster wären die vier gebildeten Gruppen (z.B. schwarze Haare/blaue Schürze oder weiße Haare/grüne Schürze etc.). Das Ergebnis wäre idealtypisch:

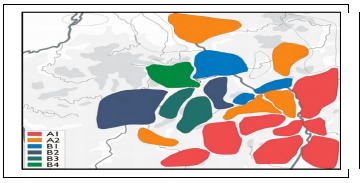
Objekte *innerhalb* einer Gruppe sind *homogen*,
Objekte *zwischen* den Gruppen sind *heterogen.*

In der Praxis werden natürlich selten Sumo-Ringer geclustert. Auffällig ist, dass seit den 90er Jahren in den verschiedensten Fachbereichen und zu den unterschiedlichsten Themen vermehrt geclustert wird. Zurückzuführen ist dies zum einen auf die bedienerfreundlicheren Programme, zum anderen auf die zahlreichen Erfahrungen mit der Methode.

Einige Beispiele seien hier genannt:



Marketing: Zusammenhang zwischen Selbstbild und Wahl einer Automarke



Archäologie: Kultureller Fingerabdrücke in der Kategorie „Schmuck“ in hallstattzeitlichen Siedlungen im Mittelrheingebiet

Botanik: Ein pflanzensoziologisches Modell der Schattentoleranz von Baumarten in den Bayerischen Alpen

Im Umfeld der Städtestatistiker wird die Clusteranalyse u.a. für Wahlanalysen, Analysen zur Luftverschmutzung und Lärmbelastung, Bürgerumfragen und Wirtschaftsanalysen eingesetzt. Der größte Anteil stellt sozioökonomische Analysen dar.

Voraussetzung

Ohne Vorüberlegungen geht es nicht

Bevor eine Clusteranalyse gerechnet wird, sind zahlreiche Vorüberlegungen notwendig. Die Auswahl und die Aufbereitung der Variablen spielen hierbei eine zentrale Rolle. Sie haben einen wesentlichen Einfluss auf das Analyseergebnis und die darauf aufbauende Interpretation.

Wichtige Punkte im Einzelnen:

- **Anzahl der Merkmale:** Für die Anzahl der Merkmale bzw. Variablen gibt es keine Begrenzung. Es sollte jedoch bei den Vorüberlegungen darauf geachtet werden, dass nur relevante themenspezifische Variablen einbezogen werden.
- **Skalenniveau:** Es ist kein spezielles Skalenniveau erforderlich. Es können alle Daten einbezogen werden.
- **Standardisierung:** Die Daten sollten vor dem Rechenbeginn standardisiert werden, um sie vergleichen zu können. Die Standardisierung erfolgt mithilfe einer z-Transformation. Sie gewährleistet, dass der Mittelwert gleich Null und die Standardabweichung gleich Eins ist.

- **Ausreißer:** Ausreißer sind Objekte, die im Gesamtvergleich Extremwerte aufweisen. Sie sollten ausgeschlossen werden, da der Fusionierungsprozess beeinflusst wird. Zusammenhänge können dadurch nur noch schlecht erkannt werden.
- **Korrelationen:** Hoch korrelierende Variablen sollten ebenfalls ausgeschlossen werden, wenn sie aus wichtigen inhaltlichen Gründen nicht einbezogen werden müssen. Das Ergebnis kann anderenfalls durch Überbewertungen verzerrt werden. Dies gilt es dann bei der Interpretation zu beachten.
- **Konstante Ausprägungen:** Variablen mit konstanten Ausgangswerten führen zu einer Nivellierung der Unterschiede und werden häufig aus der Analyse entfernt (z.B. ein Frauenanteil, der in allen Bezirken 50 % beträgt). Sie führen zu keiner Differenzierung und erschweren damit die Interpretation.
- **Anzahl Objekte:** Die Anzahl der Bezirke ist nicht begrenzt.

Methodik

Sind die Vorüberlegungen abgeschlossen, sind drei zentrale Ablaufschritte für die Berechnung der Analyse erforderlich:

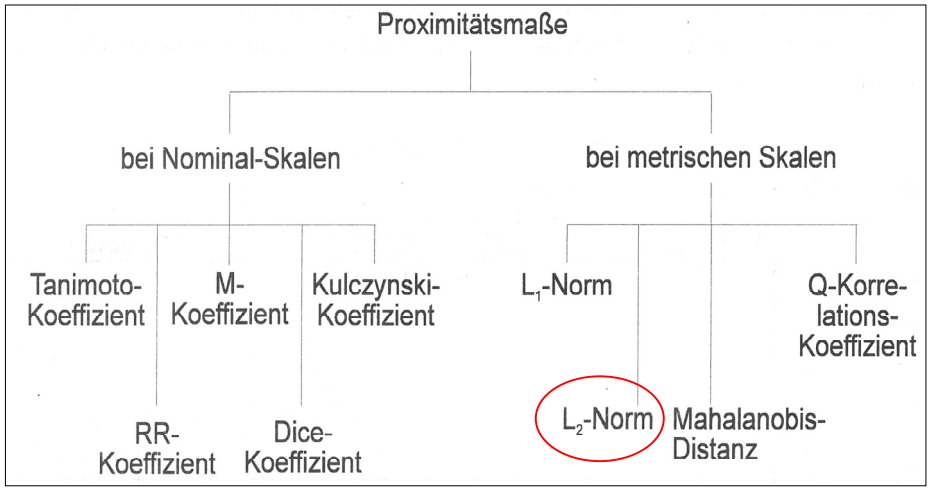
- **Bestimmung der Distanz durch die Proximitätsmaße**
- **Auswahl des Fusionierungsalgorithmuses**
- **Bestimmung der Clusteranzahl.**

Für Städtestatistiker überwiegend herangezogen: L2-Norm (Quadrierte Euklidische Distanz)

Proximitätsmaße

Die Proximitätsmaße geben die Distanz der Merkmale zwischen den einzelnen Objekten wieder. Es wurde eine Reihe von Maßen entwickelt, die abhängig sind vom Skalenniveau der Ausgangswerte. Da es sich bei den meisten von den Städtestatistikern eingesetzten Variablen um metrisch skalierte Daten handelt, kommen von den einsetzbaren Maßen überwiegend die **L-Normen** zum Einsatz. Bei der **L2-Norm**, der **Quadrierten Euklidischen Distanz,** werden die Differenzen zwischen den einzelnen Objekten am deutlichsten. Hier werden die quadrierten Differenzwerte addiert und aus der Summe die Quadratwurzel gezogen. Unterschiede zwischen ähnlichen und unähnlichen Objekten treten dadurch deutlicher hervor. Hieraus wird die **Distanzmatrix** berechnet.

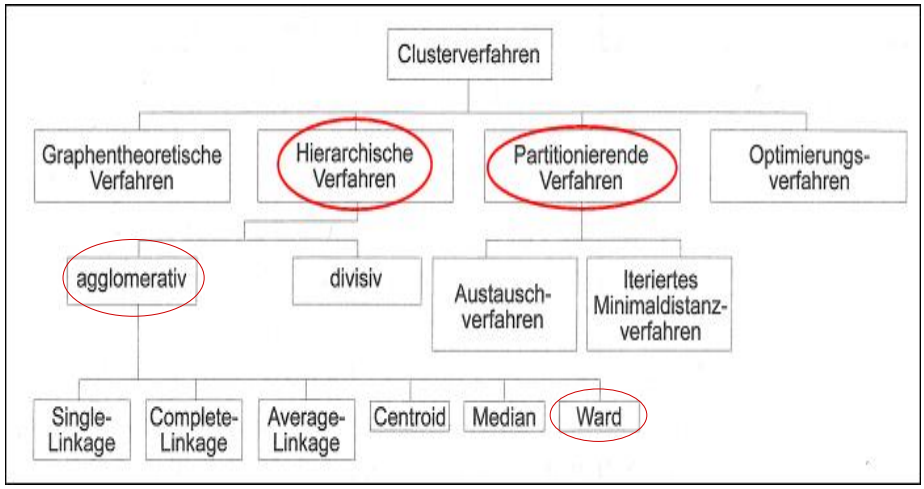
Abbildung 3: Überblick über ausgewählte Proximitätsmaße nach BACKHAUS (2006)



Fusionierung

Die **Distanzmatrix** bildet nun den Ausgangspunkt für die **Fusionierung** mit Hilfe der Cluster-Algorithmen. Die Clusteranalyse bietet den Anwendern ein breites Spektrum an Algorithmen zur Gruppierung an.

Abbildung 4: Überblick über ausgewählte Cluster-Algorithmen nach BACKHAUS (2006)



Die häufigsten Anwendungen finden die **Partitionierenden** und die **Hierarchischen Verfahren**.
Die **Partitionierenden Verfahren** gehen von gegebenen Clustern aus. Wird die gewünschte Anzahl festgelegt, werden die einzelnen Objekte mit Hilfe eines Algorithmus so lange zwischen den Clustern umgeordnet, bis ein Optimum erreicht ist. Der Nachteil des Verfahrens liegt in der vorzugebenden Startpartition.

Bei den **Hierarchischen Verfahren** wird zwischen **agglomerativen** und **divisiven** Algorithmus unterschieden. Beim divisiven Algorithmus startet die Clusteranalyse mit der gröbsten Partition. Alle Objekte befinden sich in einem Cluster. Im Laufe des Rechenprozesses wird dieser Großcluster in immer kleinere Cluster aufgeteilt. Beim agglomerativen Verfahren findet der Prozess in umgekehrter Reihenfolge statt. Der Rechenprozess geht von der feinsten Partition aus. Hier stellt jedes Objekt ein Cluster dar. Objekte mit der geringsten Distanz werden miteinander verbunden. Das Verfahren läuft solange, bis ein Großcluster entstanden ist. Die hierarchische Beziehung zwischen den einzelnen Clusterlösungen besteht in der nicht mehr aufzulösenden Bindung. Im Gegensatz zu den partitionierenden Verfahren ist eine einmal vorgenommene Verschmelzung nicht mehr zu lösen.

Ward-Verfahren von großer Bedeutung

In der Praxis hat das **agglomerative Verfahren** eine größere Bedeutung und hier vor allem das **Ward-Verfahren**.

Der Unterschied zwischen den verschiedenen agglomerativen Verfahren liegt in der Bestimmung der Distanzen. Das **Ward-Verfahren** fasst nicht wie andere Verfahren diejenigen zusammen, die die geringste Distanz aufweisen. Es werden vielmehr die Objekte miteinander vereinigt, die das **Varianzkriterium (=Fehlerquadratsumme)** am wenigsten erhöhen.
Das Varianzkriterium stellt die Summe der Entfernungsquadrate zwischen den Clustern zum Gruppenmittelpunkt des Clusters dar. Vor jeder Zusammenführung mit einem „neuen“ Cluster wird die Gesamtvarianz des „alten“ Clusters berechnet und nur jene Cluster zusammengeführt, die die Varianz am wenigsten verändern. Je höher die Fehlerquadratsumme ist, desto heterogener sind die fusionierten Cluster. Zu Beginn werden dadurch kleine Cluster gebildet, die später zu großen fusioniert werden. Dieser Vorgang wird solange fortgeführt, bis alle Cluster miteinander vereint sind. Die Schwierigkeit besteht nun darin, eine geeignete Clusteranzahl zu bestimmen.

Clusteranzahl

Zur Bestimmung der geeigneten Clusteranzahl bestehen folgende Möglichkeiten:

- **Fehlerquadratsumme**
- **Elbow-Kriterium**
- **Dendrogramm.**

Fehlerquadratsumme: In der von den Programmen ausgegeben Zuordnungsübersichten werden die einzelnen **Fusionierungsschritte** angegeben (vgl. **Fehler! Verweisquelle konnte nicht gefunden werden.**). Die Spalte „Zusammengeführte Cluster“ gibt die Nummer der zusammengezogenen Objekte bzw. Clusters an. Der neu entstandene Cluster erhält als neue Identifikationsnummer immer die Nummer des zuerst genannten Clusters (Cluster 1). Zum Nachverfolgen werden ebenfalls das erste Vorkommen der Cluster und die nächste Fusionierung (=nächster Schritt) ausgewiesen.

Die Fehlerquadratsumme spiegelt die „Heterogenität“ der gebildeten Cluster wider

In der Spalte **„Koeffizient“** wird die am Ende einer Fusion entstandene Fehlerquadratsumme wiedergegeben. Am Ende der Tabelle stehen die Differenzen dieser Summen, die manuell zu berechnen sind (Diff. CL1:CL2). Bei größeren Fusionsschritten, bei denen die **Fehlerquadratsumme** einen „Sprung“ macht, sind Cluster zusammengeführt worden, die heterogen sind. Um dies zu vermeiden, wird an dieser Stelle die Fusion „angehalten“.

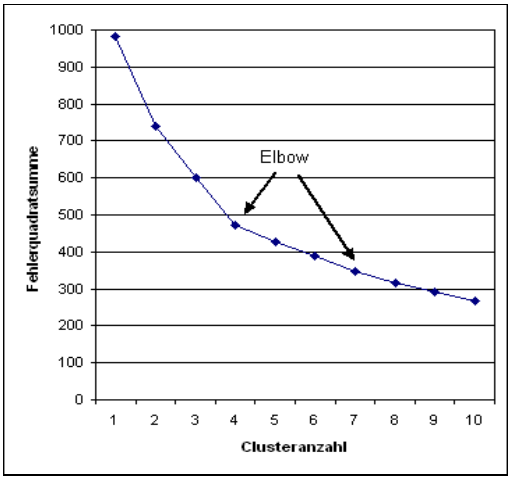
Abbildung 5: Einzelne Fusionsschritte in der Zuordnungsübersicht mit Differenzen

Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt	Diff. CL1:CL2	Clusteranzahl
	Cluster 1	Cluster 2		Cluster 1	Cluster 2			
100	25	62	267,35	99	91	103	21,8	10
101	14	18	291,56	94	97	105	24,2	9
102	3	13	317,16	96	0	108	25,6	8
103	**25**	**47**	**346,46**	**100**	**74**	**107**	**29,3**	**7**
104	**2**	**7**	**387,24**	**93**	**89**	**106**	**40,8**	**6**
105	14	65	428,42	101	98	107	41,2	5
106	**1**	**2**	**473,85**	**95**	**104**	**108**	**45,4**	**4**
107	**14**	**25**	**599,89**	**105**	**103**	**109**	**126,0**	**3**
108	1	3	738,35	106	102	109	138,5	2
109	1	14	981,00	108	107	0	242,7	1

Die **großen Sprünge** lassen sich in diesem Falle leicht erkennen. Sowohl zwischen dem 103. und 104. Schritt als auch zwischen dem 106. und 107. Bei insgesamt 109 Fusionsschritten entspricht dies schließlich einer Clusteranzahl von vier oder sieben.

Elbow-Kriterium: Beim Elbow-Kriterium werden die **Fehlerquadratsummen** in ein Diagramm abgetragen. Zeigt sich im Kurvenverlauf ein Knick („Ellbogen“), so kann dieser Wert als Entscheidungskriterium für die Clusteranzahl verwendet werden.

Abbildung 6: Elbow-Kriterium



Dendrogramm: Im Dendrogramm werden die einzelnen Fusionsschritte grafisch dargestellt. Die Schritte werden dabei auf einer Skala von 0 bis 25 normiert, wobei 25 immer der letzte Fusionsschritt darstellt. Alle Objekte befinden sich dann in einem Cluster und sind grafisch vereint.

Für die Anwender ist es hilfreich, die *nach* der gewünschten Clusteranzahl dargestellten **Fusionsschritte zu überdecken**. Die ausgewählten Cluster können leichter erkannt werden.

Abbildung 7: Dendrogramm mit allen Fusionsschritten

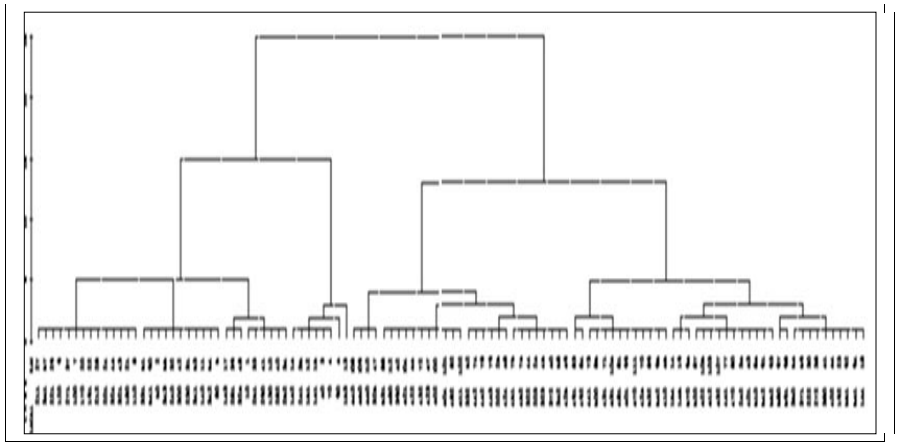


Abbildung 8:Dendrogramm mit vier ausgewählten Clustern

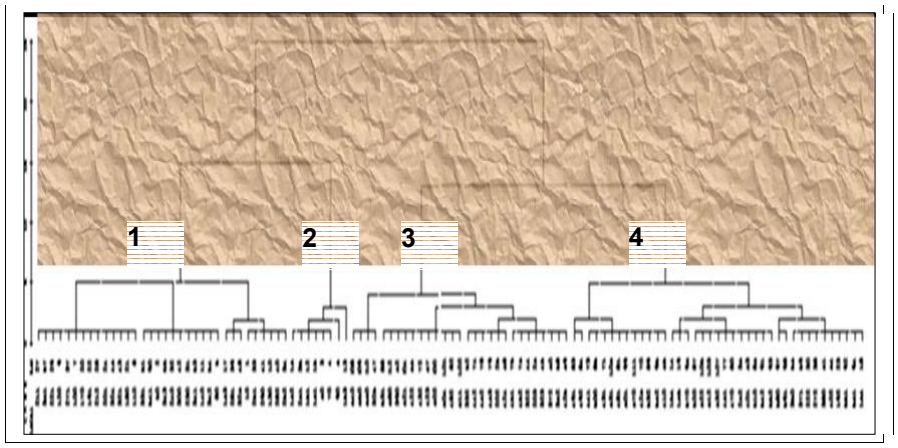
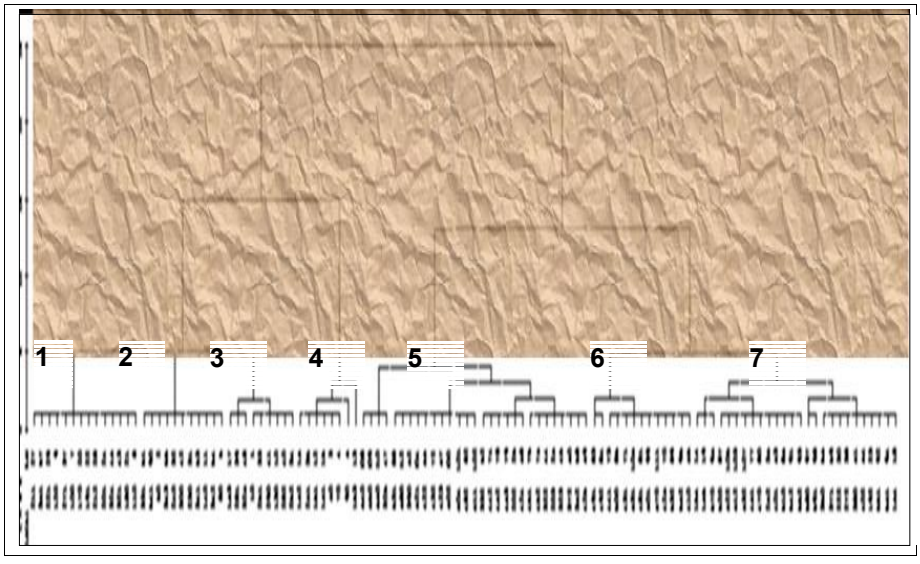


Abbildung 9: Dendrogramm mit sieben ausgewählten Clustern



Die drei wichtigsten Ablaufschritte hier nochmals zusammengefasst:

1. Bestimmung der **Distanz** durch die Proximitätsmaße
   Wahl: **Quadrierte Euklidische Distanz**
2. Auswahl des **Fusionierung**salgorithmuses
   Wahl: **hierarchisch, agglomerativ nach Ward**
3. Hilfe für die Bestimmung der **Clusteranzahl**
   Wahl: **Fehlerquadratsumme, Elbow-Kriterium und Dendrogramm.**

Interpretation

Zur Interpretation können verschiedene Indikatoren herangezogen werden:

- **Mittelwert der Variablenausprägungen in den jeweiligen Clustern**
- **Verortung der Cluster durch Übertragung in eine Karte**
- **Berechnung der F-Werte**
- **Berechnung der t-Werte.**

Je kleiner die F-Werte, umso homogener die Cluster

Die Berechnung der Mittelwerte und die Verortung der Cluster in eine Karte sind hinlänglich bekannt und müssen hier nicht näher erläutert werden. Die komplexeren Berechnungen der F- und t-Werte werden nicht automatisch von den Clusterprogrammen erstellt. Sie müssen vom Anwender selbst berechnet werden und werden dadurch seltener angewendet. Die F-Werte geben den Grad der Homogenität der Cluster an. Die Cluster sind als homogen anzusehen, wenn der Wert kleiner als Eins ist. Liegt der Wert über Eins, weist er eine größere Streuung als in der Grundgesamtheit auf. Mithilfe der t-Werte können die Cluster in ihrem Aussagewert näher bestimmt werden. Bei einem negativen t-Wert besitzt die Variable einen geringeren Anteil am Cluster und ist unterrepräsentiert, während ein positiver Wert einen höheren Anteil angibt und die Variable überrepräsentiert ist.

Sind die Werte errechnet und die Karte erstellt, kann die Interpretation beginnen. Wenn keine Analyse möglich ist, ist es sinnvoll, nochmals eine Clusterung durchzuführen. Spätestens dann sollte die Auswahl und Aufbereitung der Variablen nochmals überprüft werden.

Fazit

Festzuhalten ist, dass die Clusteranalyse ein komplexes, aber auch für weniger erfahrene Anwender durchaus handhabbares Verfahren darstellt. Die vorgestellten Verfahrensweisen stellen nur einen kleinen Ausschnitt aus der Vielzahl der Möglichkeiten dar. Sie bilden die Grundlage, eine erste Clusteranalyse zu rechnen und sich mit der Methode auseinander zu setzen. Experimentieren mit verschiedenen Einstellungen ist angeraten.
Die Clusteranalyse lässt dem Anwender hier genügend Spielraum.
Durch diese Entscheidungsfreiheit besteht leider jedoch auch die Gefahr, die Untersuchung zu manipulieren und die Ergebnisse zu beeinflussen. Wichtig ist es daher, die Vorgehensweise zu dokumentieren und einzelne Ablaufschritte darzustellen. Dazu gehören auch die Auswahl der Variablen und die konkrete Problemstellung der Untersuchung („was ist das Ziel der Untersuchung?“ und „welche Hypothese soll getestet werden?“). Wird dies beachten, ist die Clusteranalyse ein geeignetes und empfehlenswertes Verfahren für die Städtestatistiker.

Autor:
Dr. Ralf Gutfleisch
Stadt Frankfurt a.M.
Bürgeramt, Statistik und Wahlen
Zeil 3
60313 Frankfurt am Main
Tel.: 069 – 212 38 49 3
E-Mail: ralf.gutfleisch@stadt-frankfurt.de