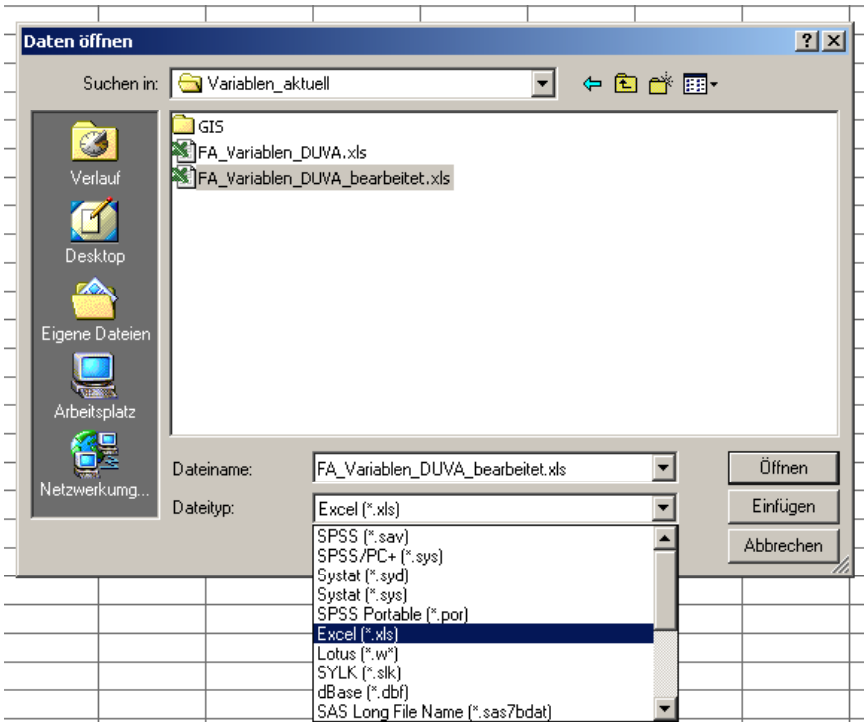
Autor: Thomas Nirschl, Amt für Stadtforschung und Statistik, Stadt Nürnberg

Clusteranalyse mit SPSS

Das Statistikpaket SPSS (aktuell in der Version 17 vorliegend) stellt dem Anwender eine große Vielfalt an nützlichen Features für statistische Auswertungen zur Verfügung. Bereits in den früheren Versionen des Programms waren diverse multivariate Methoden implementiert. Es ist somit auch nicht verwunderlich, dass zum Thema „Clusteranalyse“ auf einen ausgereiften und bewährten Werkzeugkasten zurückgegriffen werden kann. Das in diesem Leitfaden präsentierte und präferierte hierarchisch-agglomerative Verfahren ist in SPSS daher nur eine Möglichkeit unter vielen. Im Gegensatz zu den hierarchisch-divisiven Verfahren werden ausgehend von der „feinsten Partition“ die ähnlichsten Objekte schrittweise zusammengeführt, bis schließlich alle Objekte in einem Endcluster vereint sind („gröbste Partition“).

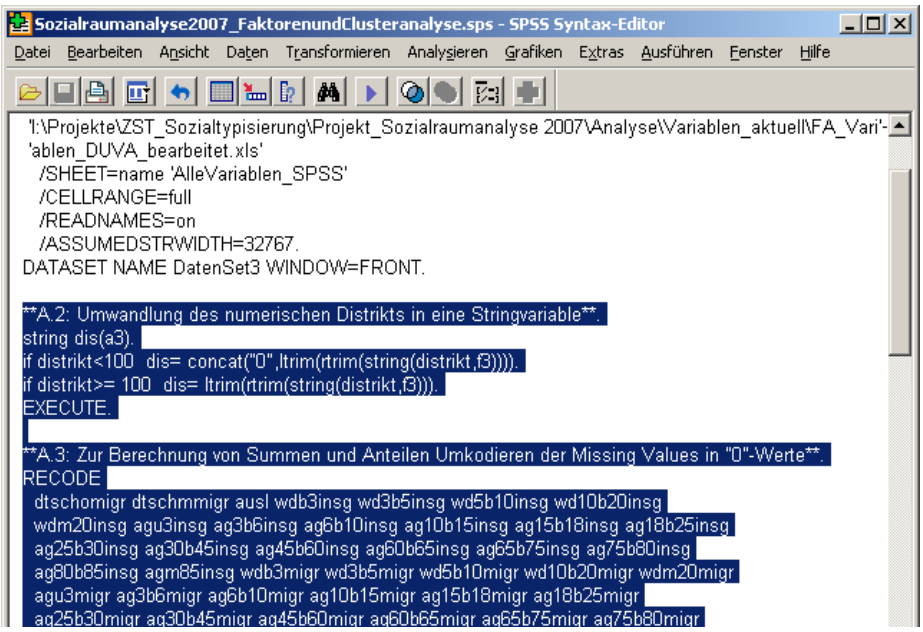
Abbildung 1: Datei öffnen



Im Folgenden wird davon ausgegangen, dass die zu analysierenden Merkmale bereits bekannt sind und in einer entsprechenden Form (hier als einfaches .xls-Sheet) vorliegen. In SPSS gelangt man über „***Datei-Öffnen***“ zur gewünschten Datei-Auswahl (s. Abbildung 1). Die Daten selbst können in den gängigsten Dateiformaten eingelesen werden. Ab diesem Schritt empfiehlt es sich, alle durchgeführten Schritte als sog. SPSS-Syntax mitzuprotokollieren; alle Kommandos werden über den Befehl „Einfügen“ in ein separates Fenster kopiert, wo sie zunächst – ohne Ausführung – als Code abgelegt sind; dieser Code muss als separate Datei (mit der Endung .sps) gespeichert werden (s. Abbildung 2). Dieses Verfahren hat entscheidende Vorteile gegenüber dem konventionellen Vorgehen:

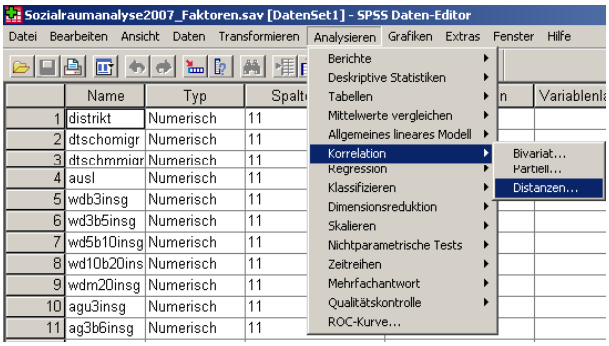
- Befehle können kommentiert werden und sind somit auch nach längerer Zeit noch nachvollziehbar – v.a. für einen selbst.
- Kommandos können modifiziert werden, ohne jedes Mal die Navigation durch die z.T. sehr verschachtelte Menüführung von SPSS erneut durchführen zu müssen; das Öffnen und Ausführen der Syntaxdatei genügt. Dies erfordert durchaus etwas fundiertere Kenntnisse von SPSS; die generell sehr gute Hilfe-Funktion des Programms gibt hier aber gute Hilfestellung.

Abbildung 2: Syntax-Editor



Sobald die Daten – sei es über die Syntax oder das Menü – eingelesen wurden, besteht die Möglichkeit die Daten noch einmal einer genaueren Analyse zu unterziehen und ggf. Daten auszuschließen. Über „***Analysieren-Korrelation-Distanzen***“ gelangt man zu einer Oberfläche, wo Art der Distanzberechnung (zwischen Variablen oder Fällen) und das Distanzmaß selbst (Ähnlichkeits- bzw. Unähnlichkeitsmaß) festzulegen sind. Hier werden die Daten nach den inhaltlich-konzeptionellen Überlegungen zur Definition des Sozialraums (s.o.) statistisch analysiert. Um Information über das Ausmaß des Zusammenhangs zwischen den Variablen zu erhalten, sind in diesem Beispiel die Distanzen zwischen den Variablen hinsichtlich ihrer Ähnlichkeit nach Pearson (Pearson-Korrelation) ermittelt worden (siehe Abbildung 3). Auch hier wird der Befehl zunächst in das Syntax-Fenster eingefügt und kommentiert. Nach Befehlsausführung erscheint in einem dritten Fenster – dem sog. Output – das Ergebnis dieser Berechnung (siehe Abbildung 4). Um die Daten detaillierter zu betrachten empfiehlt es sich SPSS zu verlassen, und die Tabelle in ein Tabellenkalkulationsprogramm zu kopieren (z.B. MS Excel, Calc). Dort sind die stark bzw. schwach korrelierenden Variablen über bedingte (Farb-)Formatierungen wesentlich leichter zu erkennen als in SPSS.

Abbildung 3: Korrelationen zwischen Variablen



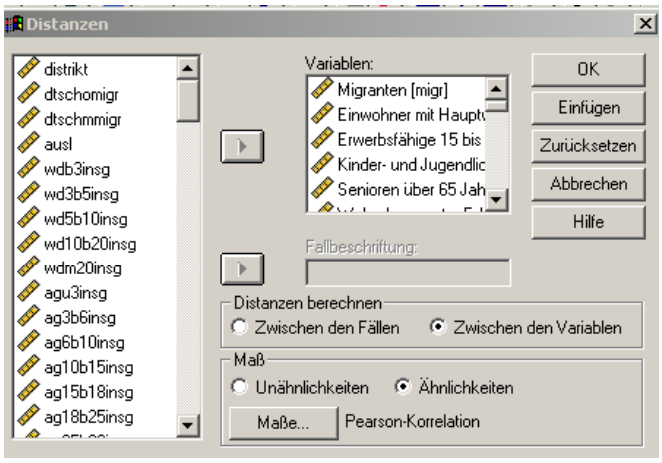
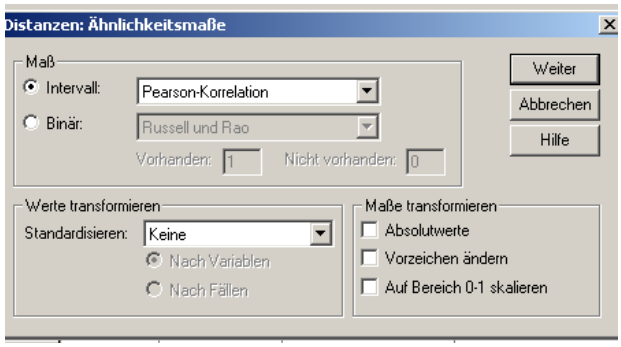


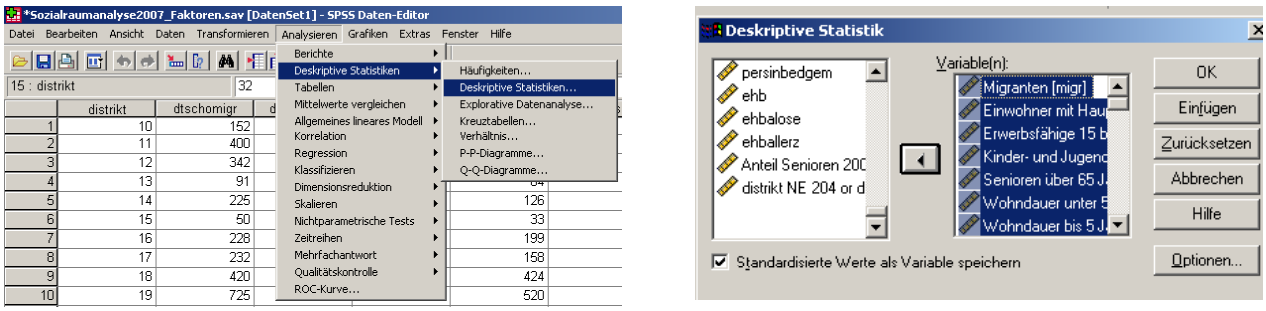
Abbildung 4: Ähnlichkeitsmaß und Output



	Migranten	Einwohner mit Hauptwohnsitz insgesamt	Erwerbsfähige 15 bis unter 65 Jahre	Kinder- und Jugendliche unter 15 Jahre
Migranten	1,000	,827	,850	,843
Einwohner mit Hauptwohnsitz insgesamt	,827	1,000	,993	,950
Erwerbsfähige 15 bis unter 65 Jahre	,850	,993	1,000	,940
Kinder- und Jugendliche unter 15 Jahre	,843	,950	,940	1,000
Senioren über 65 Jahre	,597	,889	,846	,781
Wohndauer unter 5 Jahre insgesamt	,908	,931	,945	,903
Wohndauer bis 5 Jahre Deutsch	,702	,924	,923	,849
Wohndauer bis 5 Jahre	,981	,709	,826	,821

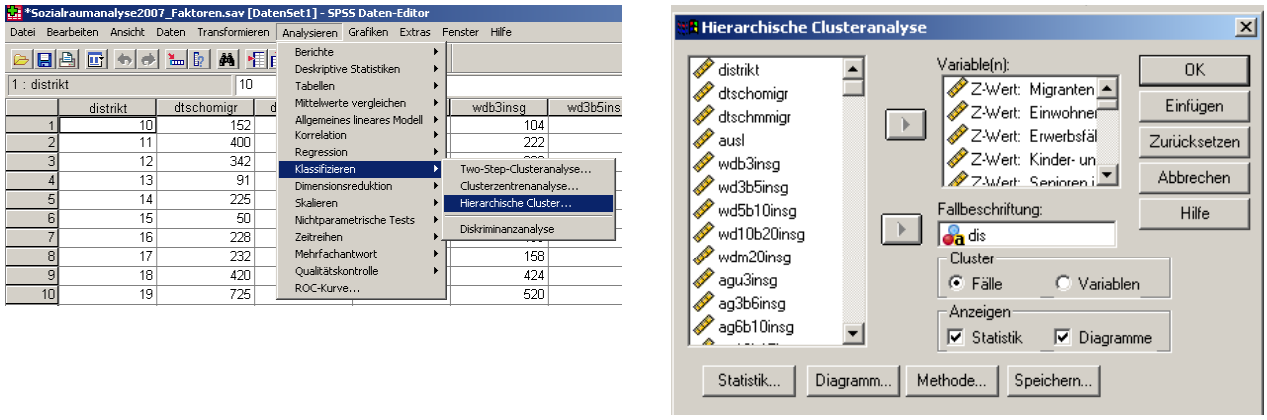
Sobald Variablen auf unterschiedlichen Maßeinheiten beruhen, kann es bei der Clusteranalyse zu einer Vergrößerung der Distanzen zwischen den Merkmalsausprägungen kommen. Um dieser „Unschärfe" im Ergebnis vorzubeugen, werden die Daten vorab standardisiert (genauer: z-transformiert, s. Abbildung 5). Bei der z-Transformation wird von jedem Wert der Mittelwert der entsprechenden Variablen abgezogen und die Differenz anschließend durch die Standardabweichung σ dividiert. Die derart standardisierten Variablen erhalten somit alle einen Mittelwert von „0" und eine Standardabweichung von „1". Die z-Transformation hat übrigens keine Einfluss auf die relativen Abstände zwischen den Werten. Sie bereinigt aber den ungewünschten Effekt der unterschiedlichen Messeinheiten; dadurch gehen alle Variablen mit derselben „Gewichtung" in die Analyse ein.

Abbildung 5: Speicherung z-transformierter Variablen



„*Analysieren-Klassifizieren-Hierarchische Cluster* ..." (s. Abbildung 6) öffnet ein Menü, welches alle relevanten Einstellungen zur Durchführung einer hierarchischen Clusteranalyse enthält.

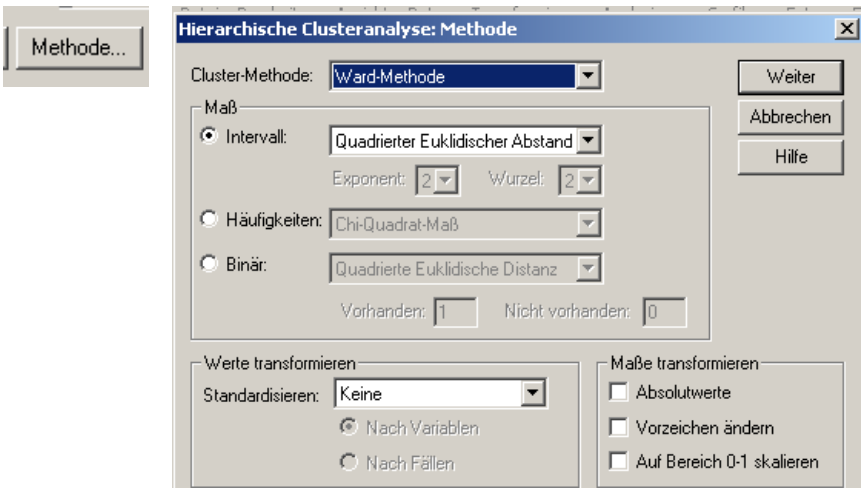
Abbildung 6: Hierarchische Clusteranalyse



In der vorliegenden Beispieldatei liegen die Variablen auf der Ebene Statistischer Distrikte vor. Dementsprechend werden diese räumlichen Einheiten unter „Fallbeschriftung“ abgelegt, wohingegen die zu analysierenden Variablen nach „Variable(n)“ geschoben werden. Für jeden Fall bzw. jedes Cluster wird nun auf Basis der Merkmale und Merkmalsausprägungen die Distanz zu jedem anderen Cluster ermittelt. Die zwei Cluster mit der geringsten Distanz zueinander (d.h. mit der größten Ähnlichkeit) werden zu einem gemeinsamen Cluster vereinigt. Dieses Prinzip der Distanzberechnung und Verschmelzung wird nun sukzessive auf alle Fälle bzw. Cluster angewandt, bis alle Fälle zu einem einzigen Cluster zusammengefasst sind. Die Herausforderung an den Anwender ist es also, eine passende Anzahl von Clustern zu ermitteln (da ein einziges Großcluster keinen Erkenntnisgewinn mit sich bringt).

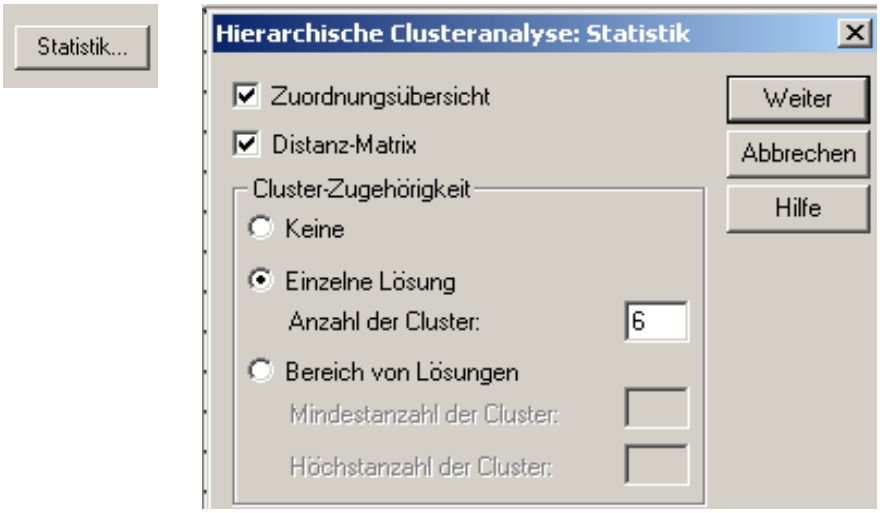
Unter „***Methode***“ (s. Abbildung 7) erfolgt die der Analyse zu Grunde liegende Auswahl der Cluster-Methode (hier: Ward-Methode). Hiermit wird der Fusionierungsalgorithmus bezeichnet, welcher die Entfernung von Clustern zueinander bestimmt. Da die Zusammenfassung von Objekten auf Basis von Distanzen erfolgt, ist die Wahl eines geeigneten Ähnlichkeits- bzw. Distanz-maßes unter „Maß-Intervall“ zwingend erforderlich (hier: Quadrierter Euklidischer Abstand). SPSS bietet für die drei potentiellen Skalenniveaus jeweils mehrere Maße zur Auswahl an. Zur Erinnerung: Die Distanz zwischen zwei Fällen ist gleich der Summe der quadrierten Differenzen für jede Eigenschaft für ein Objektpaar. Sofern die Daten noch nicht z-transfor-miert sind, ermöglicht der Punkt „Werte transfomieren-standardisieren“ eine nachträgliche Standardisierung.

Abbildung 7: Cluster-Methode und Proximitätsmaß



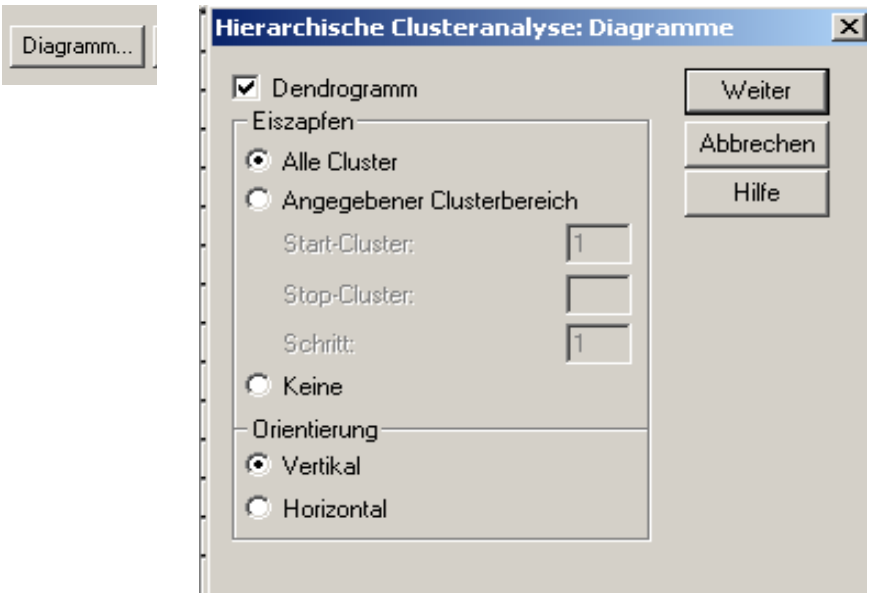
Nützliche Ausgaben zur weiteren Analyse der Cluster und –zugehörigkeiten werden im Punkt „***Statistik***“ (s. Abbildung 8) aktiviert. Dazu gehört eine „*Zuordnungsübersicht*“ sowie die „*Distanz-Matrix*“. Die Distanz-Matrix stellt die Distanzen zwischen jedem Pärchen von Fällen – und somit den Beginn der Clusteranalyse – tabellarisch dar. Hohe Werte in der Tabelle weisen auf große Unähnlichkeiten und kleine Werte auf eine relativ große Ähnlichkeit hin. Die Matrix zeigt die Distanzen vor dem Zusammenführen von Fällen zu Clustern und wird daher auch als sog. *Ausgangsdistanzmatrix* bezeichnet.

Abbildung 8: Statistik: Distanz-Matrix und Zuordnungsübersicht



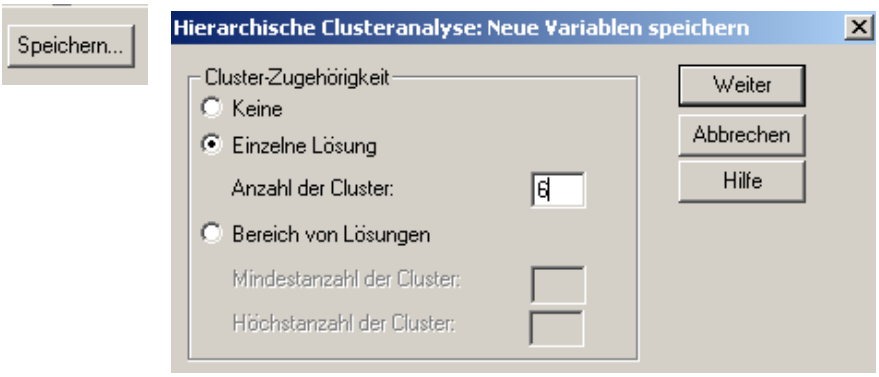
Unter „***Diagramm***“ (s. Abbilddung 9) besteht zudem die Möglichkeit, ein „*Dendrogramm*“ sowie „*Eiszapfen*“ ausgeben zu lassen. Gerade wenn man noch keine Vorstellung über Zahl und Zusammensetzung der Cluster hat, sind diese beiden graphischen Ausgaben hilfreiche Werkzeuge.

Abbildung 9: Diagramme (graphische Ausgaben)



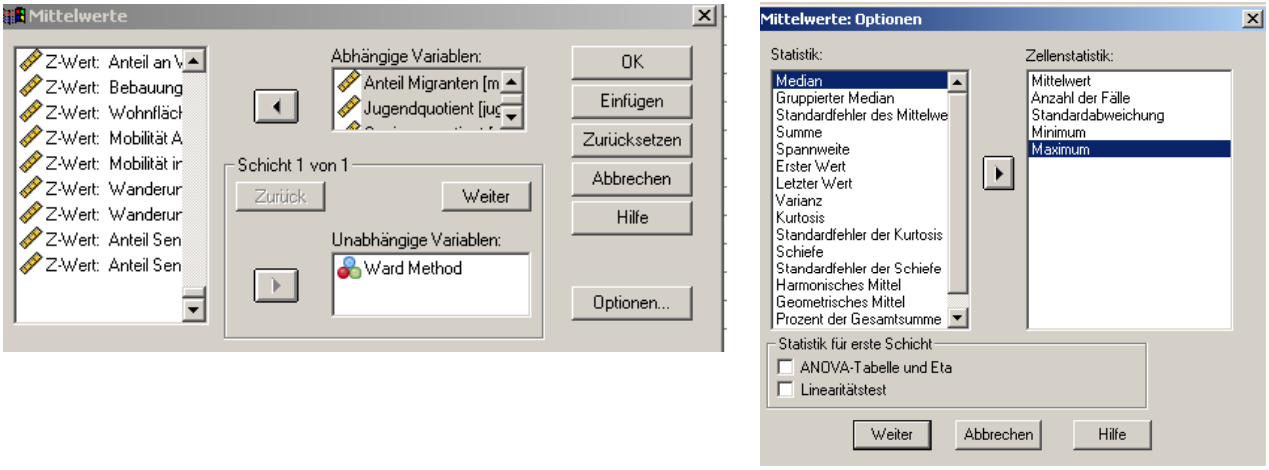
SPSS ermittelt standardmäßig keine Zugehörigkeit von Objekten zu einzelnen Clustern. Sofern man aber (unter Umständen erst nach mehreren Durchläufen) eine genaue Vorstellung über die Zahl der Cluster gewonnen hat, ist es möglich dem Programm unter „***Speichern***“ (s. Abbildung 10) die Zahl der Cluster vorzugeben (hier: „6“). Um einen Eindruck über die Verteilung von Objekten in verschiedenen Clusterlösungen zu erhalten, kann desweiteren auch ein „*Bereich von Lösungen*“ angegeben werden; die Analyse der verschiedenen Versionen und die Auswahl einer geeigneten Lösung steht dann aber immer noch aus.

Abbildung 10: Clusterlösungen



Nachdem die Cluster-Zugehörigkeit in eine neue Variable geschrieben wurde, kann nun damit begonnen werden die Cluster zu beschreiben. „***Analysieren-Mittelwerte vergleichen-Mittelwerte...***“ (s. Abbildung 11) liefert neben Mittelwerten auch Minima, Maxima, Standardabweichung σ etc. für die ermittelten Cluster und dient somit der Interpretation der Ergebnisse. Für den ein oder anderen Zweck macht es aber auch Sinn, eine separate und nach Clustern aggregierte Datei auszugeben.

Abbildung 11: Mittelwerte vergleichen



Es ist zwar durchaus möglich, alle graphischen Ausgaben (Tabellen, Diagramme) als separate Output-Datei zu speichern (.spo), allein die Bedienerfreundlichkeit und die Weiterverarbeitung der Informationen ist für die meisten Nutzer recht ungewöhnlich. SPSS bietet aber die Möglichkeit Daten in anderen Programmen (z.B. MS Excel, MS Word, Calc) zu verarbeiten; diese Option ist für den SPSS-Einsteiger durchaus zu empfehlen. Für die Darstellung der räumlichen Verteilung der Cluster und Clusterelemente in Form einer Karte muss SPSS ohnehin verlassen und auf ein GIS (z.B. ArcView) zurückgegriffen werden.

Autor:
Thomas Nirschl
Stadt Nürnberg
Amt für Stadtforschung und Statistik für Nürnberg und Fürth
Unschlittplatz 7a
90403 Nürnberg
Tel.: 0911 – 231 2842
E-mail: thomas.nirschl@stadt.nuernberg.de