Cluster Analysis

Johann Bacher

Chair of Sociology

University Erlangen-Nuremberg

Findelgasse 7-9

D-90402 Nuremberg

Nuremberg, 2002

Contents

Chapter 1:
Overview and Examples

1.1 Purpose and Techniques

The main idea of cluster analysis is very simple (Bacher 1996: 1-4):

- **Find K clusters (or a classification that consists of K clusters) so that the objects of one cluster are similar to each other whereas objects of different clusters are dissimilar.**

The following quotations should additionally illustrate this task:

"This monograph will be concerned with certain techniques for the analysis of multivariate data, which attempt to solve the following problem:
Given a number of objects or individuals, each of which is described by a set of numerical measures, devise a classification scheme for grouping the objects into a number of classes such that objects within classes are similar in some respect and unlike those form other classes. The number of classes and the characteristics of each class are to be determined".
(Everitt 1981: 1).

"The subject of classification is concerned with the investigation of the relationships within a set of 'objects' in order to establish whether or not the data can validly be summarized by a small number of classes (or clusters) of similar objects."
(Gordon 1999: 1)

Everitt's characteristic requires two notes:

- Clustering techniques can also be applied to cluster variables. Everitt only mentions cases!
- Clustering techniques can also be applied in a confirmatory way. Everitt's definition suggests that cluster analysis is an explorative technique.

The description of Gordon also needs some remarks:

- According to Gordon the classification must be valid.
- The number of clusters should be small.

Gordon formulates additional criteria: A cluster should contain similar objects and should satisfy additional criteria. Compared to Everitt, Gordon's definition portrays the development in cluster analysis. Until the 80s the discussion concentrated mainly on techniques. At the end of the 80s the whole process of clustering – starting with the selection of cases and variables and ending with the validation of clusters – became dominant. The steps in a clustering process are:

1. selection of appropriate cases, variables and methods
2. application of the methods
3. evaluation of the results.

This last step includes:

1. determination of the number of clusters, if unknown
2. substantive interpretation of clusters
3. test of stability
4. test of internal validity (model fit), relative validity and external validity

Techniques

Different techniques have been developed to cluster cases or variables. The lecture will discuss the most important ones:

- **Hierarchical clustering methods (see chapter 4).** They result in a hierarchy of classifications (partitions).
- **K-means clustering methods (see chapter 5).** They result in a classification with K clusters. A sequence of clusters containing a different number of clusters is not automatically generated.
- **Probabilistic methods (see chapter 7),** like latent class and latent profile methods or mixture models. These methods differ from the two approaches mentioned before (hierarchical and k-means techniques) in the assignment of objects to the clusters. Hierarchical and k-means techniques result in a deterministic assignment. An object can only belong to one cluster, e.g. object 1 belongs to cluster 2, object 2 to cluster 2, object 3

to cluster 1, and so on (see table 1-1). Probabilistic techniques assign objects with certain probabilities to the clusters, e.g. object 1 belongs with a probability of 0.1 to cluster 1, with a probability of 0.7 to cluster 2 and with a probability of 0.2 to cluster 3.

	deterministic clustering (hierarchical or k-means)				**probabilistic clustering**		
objects	**member-ship**	**cluster 1**	**cluster 2**	**cluster 3**	**cluster 1**	**cluster 2**	**cluster 3**
1	2	0	1	0	0.1	0.7	0.2
2	2	0	1	0	0.0	0.5	0.5
3	1	1	0	0	0.9	0.0	0.1
4	3	0	0	1	0.0	0.0	1.0

Table 1-1: Deterministic versus probabilistic assignment of objects

Most techniques are known for many decades. K-means was developed in the 70s by Forgy (Jain and Dubes 1988: 97). Hierarchical Methods were already developed in the 60s (see Everitt 1981: 28). However, the **implementation of clustering techniques** is still unsatisfactory in standard statistical software, especially in SPSS. SPSS (Version 10.0.5 for Windows) offers only two procedures:

- One module for hierarchical clustering (CLUSTER) and
- one module for k-means clustering (QUICK CLUSTER).

Statistical tests or criteria are not available. Probabilistic models are not covered, too. Therefore, a discussion of additional programme packages is necessary. We will concentrate on two programmes: ALMO and CLUSTAN (see chapter 4). LatentGold as a third one will also be mentioned (see chapter 7).

Clustering Cases or Variables

Agglomerative hierarchical techniques may be used to cluster cases **(clustering cases)** or variables **(clustering variables)**. **K-means and probabilistic methods only allow you to cluster cases.** Figure 1-1 visualizes the difference between clustering cases and variables.

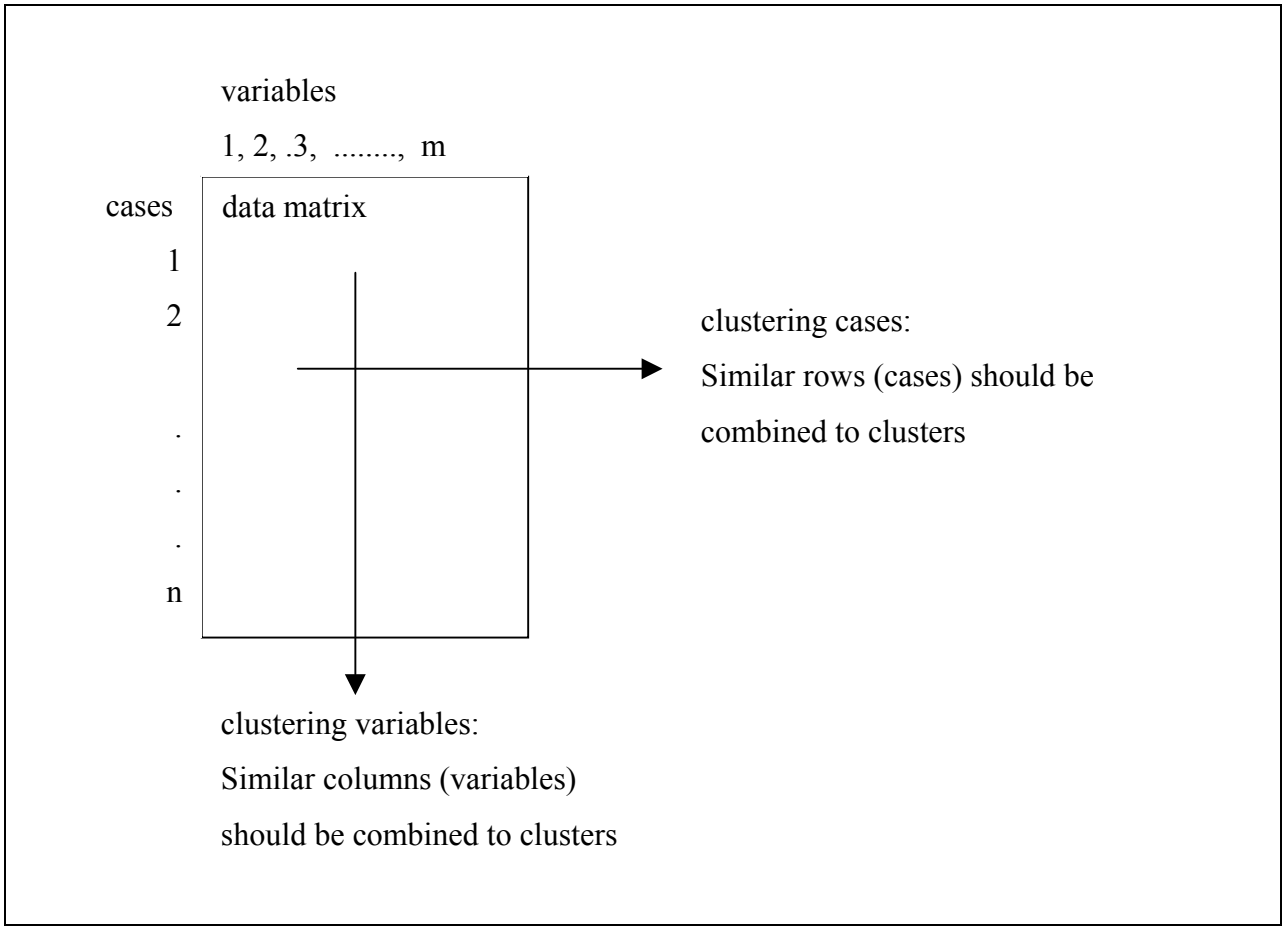


Figure 1-1: Clustering cases or variables

Example: A sample of n persons have been asked about their sympathy for m countries ('Here is a list of m countries. Can you select those countries you like?'). If a researcher wants to analyse the question 'which countries are similar?' he will cluster variables and he must apply agglomerative hierarchical techniques (or other techniques for clustering variables).

If the question is 'do the persons differ in their sympathy and can different patterns of preferences be identified?' cases are clustered. For this purpose, all three methods

(hierarchical techniques, k-means methods and probabilistic clustering) may be used in principle.

Agglomerative hierarchical techniques can be - dependent on your hardware - applied for small (e.g. n=50) and moderate sample sizes (e.g. n=500). K-means methods require at least a moderate sample size (e.g. n=300). K-means can be used for large sample sizes, too. Probabilistic techniques require large sample size (e.g. n=3000). The size of sample depends on the structure of the data. If well separated clusters exist the sample size can be smaller. Therefore, general threshold values cannot be given.

Explorative Cluster Analysis and Confirmatory Cluster Analysis

Clustering techniques are regarded as **explorative methods** in many text books. This is correct only to some extent. They only require a specificatoin of the variables and cases that should be used. A specification of the number of clusters is not necessary in advance. The number of clusters can be determined in principle. It is also not necessary to specify certain characteristics of the cluster (e.g. clustering variables: country A and B are in the same clusters; or clustering cases: cluster 1 prefers countries A, B and C, cluster 2 countries A and D, and so on). Therefore, clustering techniques may be used in an explorative way. But they can also be applied as confirmatory techniques. In this case the number of clusters and certain characteristics of the clusters are fixed. Figure 1-2 shows the differences between exploratory and confirmatory cluster analysis.

Exploratory cluster analysis	Confirmatory cluster analysis
• The number of clusters is unknown. => The number of clusters has to be estimated.	• The number of clusters is known. => The number of clusters shall not be estimated.
• The characteristics of clusters (e.g. cluster centres in k-means) are unknown. => Clusters have to be interpreted. Finding a substantive interpretation can be difficult.	• Characteristics of clusters are - at least partially – known. => Clusters already have a substantive interpretation.
• The fit to data is maximized.	• The fit to data may be poor.

Figure 1-2: Differences between exploratory and confirmatory cluster analysis

Confirmatory cluster analysis has **two advantages** (see figure 1-2):

- Confirmatory cluster analysis avoids the problem of determining the number of clusters. This problem is still unsolved.
- There is already a substantive interpretation for clusters. Finding such an interpretation can be difficult.

The **disadvantages** are:

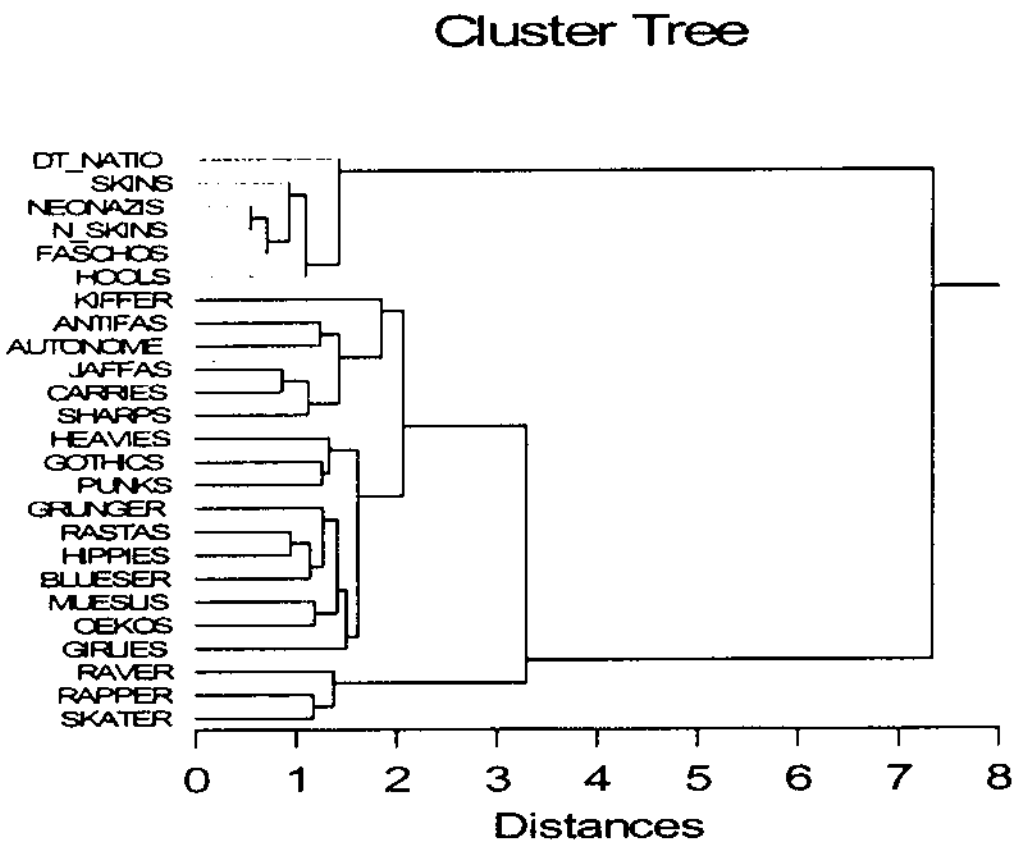
- Fit to data may be poor.
- Methods for confirmatory cluster analysis are not available in standard software. SPSS offers only a rudimentary confirmatory analysis. All starting values have to be fixed or freed for estimation. Linear or non linear restrictions or fixing some parameters is not possible.

Despite these advantages confirmatory techniques are rarely used.

1.2 Examples

Clustering Variables: Affiliation to Youth Cultures

Neumann et al. (1999) applied hierarchical techniques to analyse the perception of different youth cultures. Their analysis is based on a sample of 2.500 young Germans aged between 15 and 19 years. The survey was carried out in four different federal states ("Länder": Brandenburg, Schleswig-Holstein, Bayern and Thüringen) in 1996 and 1997. Figure 1-3 summarizes the results of their cluster analysis.



rechte Jugendkulturen (n=82)	harte Jugendkulturen (n=320)	weiche Jugendkulturen (n=111)	Popkulturen (n=250)
Deutsch-Nationale	Kiffer	Grunger	Raver
Skins	Heavies	Rastas	Skater
Neonazis	Gothics	Hippies	Rapper
Nazi-Skins	Punks	Blueser	
Faschos	Antifas	Müslis	
Hooligans	Autonome	Ökos	
	S.H.A.R.P.'s	Girlies	

Figure 1-3: An example of clustering variables (Neumann et al. 1999: 129)

The authors differentiate between four clusters:

- right wing youth cultures
- hard youth cultures

- soft youth cultures
- pop cultures

The results (in the Cluster Tree) show the relations between the four clusters: Soft and hard youth cultures are more similar to each other than the other clusters. They would be combined in the next step of agglomerative hierarchical cluster analysis. The next step would join soft and hard youth cultures with pop culture. These three clusters have large distances to the right wing youth cultures. Right wing youth cultures are seen as different from other youth cultures. This has two implications: 1. Right wing groups have difficulties in attracting juveniles with affiliations to other groups. 2. It is difficult to attract juveniles to other groups, who sympathize with right wing groups.

Unfortunately, the authors do not document the method and the similarity measure used. I suppose, it was an agglomerative hierarchical technique. Nonetheless, in chapter 4.2 we will try to reproduce the method used.

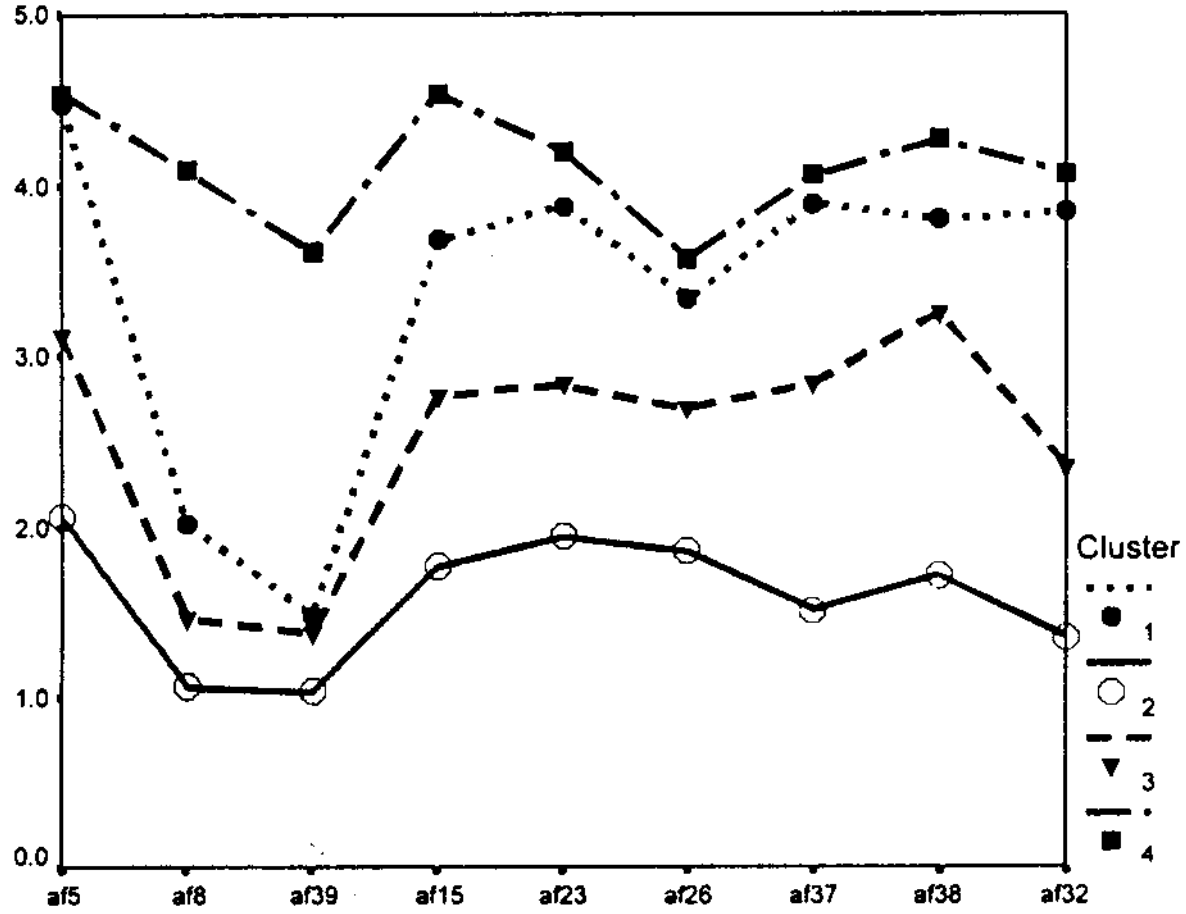
The procedure used by the authors can be applied in other research fields. Affiliations to political parties, to certain products or services, etc. can be analysed in a similar way.

Clustering Cases I: Analysing an Attitude Scale

Neumann et al. (1999) also applied cluster analysis to the analysis of an attitude scale on xenophobia. The used scale consists of nine items. The items capture different aspects of discrimination and assimilation of foreigners. The authors reported to have found four clusters. The clusters are (see figure 1–4):

- **"Romantic anti-racism"** (n=576; cluster 2). All items of xenophobia are rejected.
- **"Eliminatory" xenophobia** (n=334; cluster 4). Members of these clusters agree to all items. They support the exclusion of migrants and foreigners from Germany.
- **"Assimilatory" xenophobia** (n=292; cluster 1). Members of this cluster agree to all items except to those items, that demand the exclusion of foreigners from Germany (item af8 and af39).

- **The average** (n=811; cluster 3). The authors assume that fear is the motive underlying this cluster.



af5: Immigrants should practice their culture at home. They should adapt themselves to German culture when living in Germany.

af8: Only Germans should live in Germany.

af39: Germans should not marry immigrants.

af15: Immigrants provoke xenophobia by their behaviour.

af23: Most politicians take too much care of immigrants and do not take care of Germans.

af26: Some immigrants do not work hard enough. Otherwise they could enjoy the same standard of living like Germans.

af37: Male immigrants bother females more often than German males.

af38: Immigrants in Germany should not interlope if they are not liked.

af32: Immigrants have jobs that Germans should have.

Figure 1-4: Items used by Neumann et al. (1999) to measure xenophobia

In my opinion, the labels of the clusters are problematic, because anti-racism is regarded as a romantic attitude.

I selected the example, because it is unusual to analyse attitude scales with cluster analysis. The example raises the question whether or not cluster analysis is superior to factor analysis and allows to discover patterns that are not captured by factor analysis. Factor analysis assumes parallel response patterns in the uni-dimensional case. Because of this, factor analysis probably would not detect cluster 1. Factor analysis and cluster analysis will be compared in chapter 4.13.

Clustering Cases II: Types of Juveniles

On the basis of the 12th Shell youth study Münchmeier (1997) extracts five clusters:

- **Kids:** They have no clear opinion about politics. An uncritical attitude predominates, the political interest is low. Kids are the youngest group.
- **Critical, but loyal juveniles:** They perceive social and economic problems, their perception of the future is pessimistic. They are interested in politics. Therefore, their political knowledge is above the average. They support institutionalized political actions avoiding conflicts.
- **Traditional juveniles:** They are convinced that politics is able to solve problems. They perceive fewer social and economic problems than the average. They feel less politically alienated and less pessimistic. Their political interest is high and they prefer institutionalized means of political participation (e.g. to vote). They are labelled traditional because they affiliated with SPD and CDU/CSU as traditional political parties.
- **Conventional juveniles:** Juveniles of this cluster are not interested in politics. They feel highly alienated. Politics is not able to solve problems in their opinion. However, they do not perceive economic and social problems above the average. Their willingness to participate is low.
- **(Not yet) integrated juveniles:** These juveniles feel alienated, too. But they do not withdraw to private live. They are willing to choose conflictual forms of political participation. The future is seen pessimistic. Economic as well as social problems are perceived.

Münchmeier used the variables shown in figure 1-4a.

• gender
• age
• federal state (Bundesland)
• pessimistic view of future
• politicians are not interested in juveniles
• political alienation
• political knowledge
• institutionalized political activities
• political activities avoiding conflicts
• conflictual activities
• perception of social problems in the society
• perception of economic problems in the society
• efficiency of politics
• motivation for political action: efficiency
• motivation for political efficiency: effectiveness
• basic orientation: assertiveness
• basic orientation: privatism
• perception of a generation conflict

Figure 1-4a: Variables used by Münchmeier

Münchmeier used socio-demographic information (gender, age, federal state) on the one hand and variables that are connected with politics (political interest, attitudes and behaviour) on the other hand. The last variables are scales derived from manifest items by scaling techniques. The scales are well documented. Figure 1-5 shows an example.

SKALA: ZUKUNFTSPESSIMISMUS / FRAGE 6

Item	Mittelwert
1. Technik und Chemie werden die Umwelt zerstören	2,73
2. Gewalttätige Konflikte werden das Leben zunehmend unsicherer machen	3,10
3. (i) Wir werden einen wirtschaftlichen Aufschwung erleben	2,23
4. (i) Es wird gelingen, die Umweltprobleme zu lösen	2,17
5. Die wirtschaftliche Krise wird sich verschärfen	2,88
6. Es wird immer weniger Arbeitsplätze geben, noch mehr Menschen werden arbeitslos werden	3,25
7. (i) Die Menschen werden wieder friedlicher und gewaltfreier zusammenleben	1,95
8. (i) Es wird für alle einen angemessenen Arbeitsplatz geben, die Arbeitslosigkeit wird verschwinden	1,59

Mittelwert Skala: 24,0 Minimum: 8
Theoretischer Mittelwert: 20,0 Maximum: 32

Reliabilität: Cronbach's Alpha r_{tt} = .77

Quelle der Skala: 10. Shell Jugendstudie, Bd. 1, S. 112 f.*

Abfragemodus:
4 = bestimmt, 3 = wahrscheinlich, 2 = wahrscheinlich nicht, 1 = bestimmt nicht.

(i) = Zur Errechnung des Punktwertes der Skala wird dieses Item invertiert.

* *Die Items 2 und 7 der Originalskala („Die Welt wird in einem Atomkrieg untergehen" und „In Europa werden die Atomwaffen auf beiden Seiten abgeschafft") wurden durch die neuen Items 2 und 7 ersetzt. Für einen Zeitvergleich kann deshalb nur auf die Werte der übrigen 6 Items zurückgegriffen werden.*

Figure 1-5: Example of a scale used by Münchmeier

The variables connected to politics have **different ranges**. Pessimism, for example, has a range from 8 to 32 points, political alienation from 5 to 20 points, and so on. The variables are

incommensurable. They must be transformed for cluster analysis. Cluster analysis (more precisely the tow deterministic methods) requires equal scales.

The variables have **different measurement levels,** too. Federal state and gender are nominal-scaled variables, all others are quantitative (interval-scaled) variables. This kind of **incommensurability** must also be handled if cluster analysis (more precisely the tow deterministic methods) should be used.

Different methods have been proposed for these problems. Some of them will be discussed in chapter 2 and chapter 7 (mixed variables). Münchmeier does not report the transformation he used. The clustering method he applied is not mentioned, too. Probably, he applied k-means because of the large sample size (n=2011). However, he reports that the technique he used computes similarities between all objects. This suggests that a hierarchical method was used.

Clustering Cases III: Analysing Life Styles

Lechner (2001) selects leisure time preferences of juveniles (in her analysis apprentices) to identify different life styles. The variables she used are:

- Preference for a special leisure time activity. Six factors were extracted by factor analysis: artistical (creative) activities, passive activities (consumption), going to parties/discos, visiting pop concerts and playing with computers, practising sports, driving motorcycles/cars.
- Preference for a special film category. Factor analysis differentiates five categories: action, horror, entertainment, classics and other films (like Western).
- Preference for a special music style. Factor analysis distinguishes between five categories: hard'n heavy, black roots, commerce, techno/house, grufties.

Lechner found seven clusters:

- Rockers: they prefer aggressive music, they like to repair and drive with cars and motor cycles.
- Female mainstreamers: they prefer consumption and pop music.

- Balanced consumers: they combine passive and active leisure time activities.
- Athletes: sport is the most important activity, athletes have no preference for a certain music, action films are preferred.
- The young savages: they want to make as much experiences as possible, prefer going to parties, playing computer games, etc.
- Grufties: they prefer wave/industrial music, favour a wide range of different activities, creative arts are important to them.
- Television fans: their most important activity is watching TV, they like all kinds of films except horror films.

Lechner used k-means. The starting values were computed with Ward's method. I refer to this example because data analysis is well documented.

Some General Remarks on Life Style Research

Life style typologies are based on different sets of variables. Lechner (2001; see above) uses preferences for special leisure time activities. Very often the interest is to find a more stable and general typology that allows you to predict behaviour in the different areas of life, e.g. politics, leisure and work. The SINUS-milieus, developed by the SINUS-Institute in Heidelberg (see for example Flaig 1997), are the most prominent examples of this general approach. The SINUS-milieus are based on general value orientations.

They are normally described by two dimensions: social status and value orientation. Social status constitutes the vertical dimension, the value orientation the horizontal one (see figure 1-6). Five categories of social status are distinguished: lower class, lower middle class, middle class, upper middle class and upper class. The classes are characterized by different value orientations. Six value orientations are distinguished: traditionalism, materialism with status/property and consumption as two subdimensions, hedonism, post-materialism and post-modernism. The characteristic orientations of these dimensions are: 'to preserve' for conservatism, 'to have' or 'to defend' for status/property, 'to buy' and 'to consume' for consumption, 'to indulge' for hedonism, 'to be' and 'to share' for post-materialism and 'to have, to sense and to indulge' for post-modernism. It is assumed that the value orientations build one dimension so that social milieus can be visualized in a two-dimensional space.

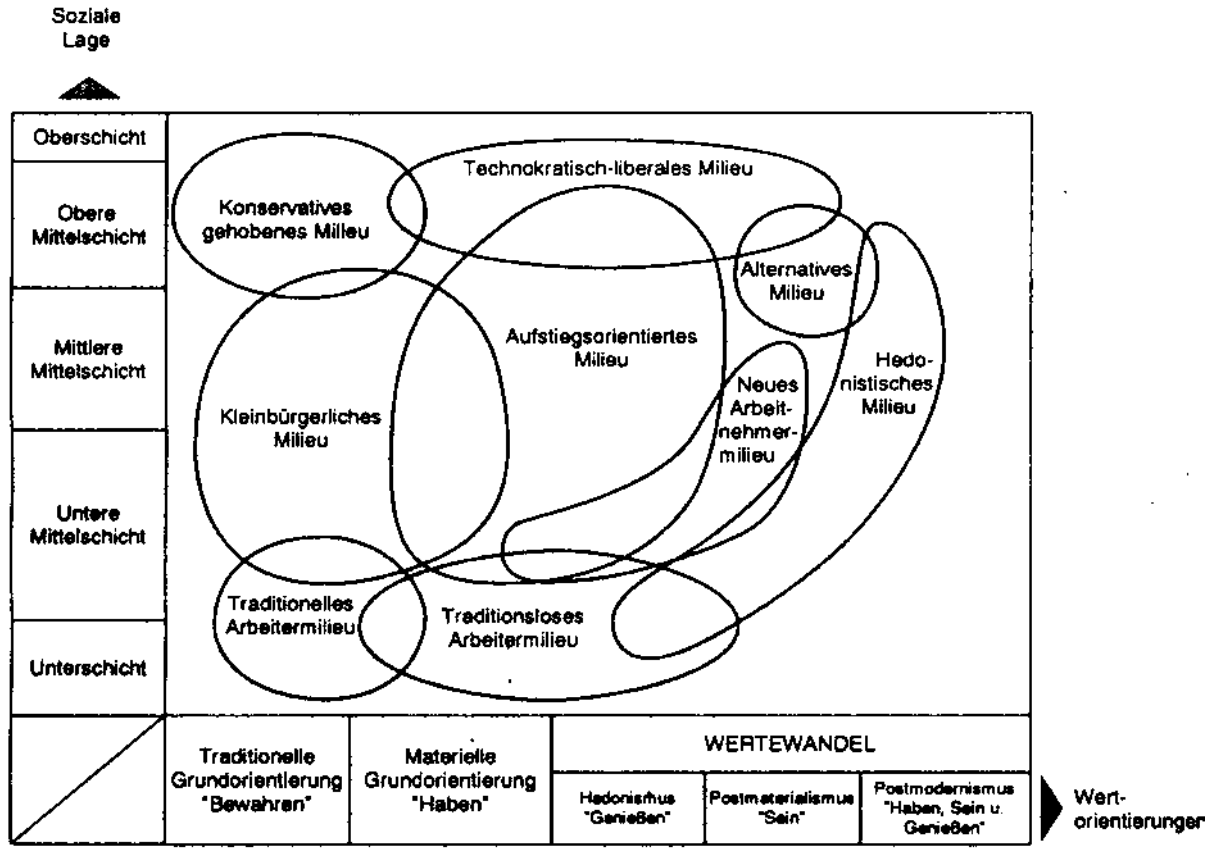


Figure 1-6: The SINUS-milieus (Flaig et al. 1997: 74)

Again, the clustering method is not documented. Indicators are published.

1.3 Criteria for a Good Classification

The main objective of clustering techniques is to compute a classification: The objects of one cluster should be similar (to each other), objects of different clusters should be dissimilar. A good classification should fulfil different criteria (Bacher 1996: 2-4, 150-154; Gordon 1999: 183-211):

Internal validity:

1. The clusters of the classification (the partition) should be homogenous: The objects that belong to the same cluster should be similar.
2. The clusters should be isolated from each other: Objects of different clusters should be different.
3. The classification should fit to the data: The classification should be able to explain the variation in the data.

Interpretability:

4. The clusters should have a substantive interpretation: It is possible to give names to the clusters. Ideally, these names should correspond to types deduced from a certain theory.

Stability:

5. The clusters should be stable (stability): Small modifications in data and methods should not change the results.

External validity:

6. The clusters should be valid (external validity): The clusters should correlate with external variables that are known to be correlated with the classification and that are not used for clustering.

Relative validity:

7. The classification should be better than the null model that assumes no clusters are present.
8. The classification should be better than other classifications (relative validity).

Further criteria:

9. Sometimes the size and the number of the clusters are used as additional criteria: The number of clusters should be as small as possible. The size of the clusters should not be too small.

The criteria are not completely independent. If clusters are isolated and homogenous, the fit to data will generally (but not necessarily) be high. If the fit to data is good, the null model (no clusters are present) will probably not fit (but not automatically).

Sometimes, one is interested only in some criteria, e.g. in criteria 1 to 4. Very often more than one good classification will be computed. One classification might fulfil criteria 1, 3, 4 and 7 best, the other partition meets criteria 1, 2, 4 and 6, and so on. However, all classifications have to fulfil criteria 4: They must be substantially interpretable.

1.4 Typologies without Cluster Analysis

To avoid misunderstandings: Not every typology is based on a cluster analysis. Moreover, well known typologies are constructed without using cluster analysis. A prominent example is Inglehart (1979). He distinguishes four types: pure materialistic persons, pure postmaterialistic persons, mixed materialistic persons and mixed postmaterialistic persons. The types are built by a simple mathematical transformation. In the short version of his scale four items (two materialistic and two postmaterialistic ones) are ranked from most important to least important. If the two materialistic items or the two post-materialistic items are ranked first and second a pure materialistic type or post-materialistic type is assumed. In the mixed materialistic case or post-materialistic case a materialistic item is ranked first and a post-materialistic item second or vice versa.

1.5 Further Applications of Clustering Techniques

Clustering techniques compute a classification. They can also be applied to other problems in data analysis. I would like to refer to the following:

- imputation of missing values (see chapter 6.3).
- data fusion (see chapter 6.3)
- statistical or optimal matching (see chapter 6.3)
- data mining (see chapter 6.4)

All these problems are intensively discussed nowadays. Among other methods, clustering techniques can be used to solve them. However, cluster analysis is ignored in some of these fields (imputation of missing values, optimal matching). In the other two fields cluster analysis is an important method.

References

Methods and Techniques

Bacher, J., 1996: Clusteranalyse [Cluster analysis]. Opladen. [only available in German].

Everitt, B., 1981: Cluster analysis. 2nd edition. London-New York. [4th edition available: Everitt, B., Landau, S., Leese, M., 2001: Cluster analysis. 4th edition. London-New York]

Gordon, A. D., 1999: Classification. 2nd edition. London and others.

Jain, A. K., Dubes, R.C. 1988: Algorithms for Clustering Data. Englewood Cliffs (New Jersey).

Substantive Examples

Flaig, B.B., Meyer, Th., Ueltzhöffer, J., 1997: Alltagsästhetik und politische Kultur [Aesthetic and political culture]. Opladen. [only available in German].

Inglehart, R., 1979: Wertewandel in den westlichen Gesellschaften: Politische Konsequenzen von materiellen und postmaterialistischen Prioritäten [Value change in Western societies: Political consequences of materialistic and postmaterialistic orientations]. In: H. Klages und P. Kmieciak (Hg): Wertewandel und gesellschaftlicher Wandel. Frankfurt a. M., 279-316 [only available in German].

Lechner, B., 2001: Freizeitverhalten von BerufsschülerInnen [Leisure time activities of apprentices]. Nürnberg. [only available in German].

Münchmeier, R., 1997: Jung – ansonsten ganz verschieden [young – and different]. In: Jugendwerk der Deutschen Shell (Hg.): Jugend 97. Opladen, 379-389. [only available in German].

Neumann, J., Frindte, W., Fuchs, F., Jacob, S., 1999: Sozialpsychologische Hintergründe von Fremdenfeindlichkeit und Rechtsextremismus. [Social psychological determinants of xenophobia and right-extremism]. In: F. Dünkel and B. Geng (Eds.): Rechtsextremismus und Fremdenfeindlichkeit. Godesberg, 111-138 [only available in German].

Chapter 2:
Transformation of Variables

Note: For a detailed discussion on this topic see Bacher (1996: 173-191). The problem is also discussed in Everitt (1981: 9-12) and Gordon (1999: 23-28).

2.1 Incommensurability

Cluster analysis requires commensurable variables. The variables must have equal scales. This assumption is violated in the following situations:

- All variables are quantitative, but have different scales (e.g. AGE and INCOME are used).
- The variables have different measurement levels (e.g. AGE (=quantitative), GRADE (=ordinal) and BRANCH OF STUDY (=nominal) are used).
- The variables are hierarchical. The occurrence of one variable depends on another variable (e.g. OCCUPATIONAL LEVEL depends on OCCUPATIONAL STATUS).

In addition, substantive consideration may result in the opinion that variables are incommensurable (Fox 1982).

This chapter concentrates on the first case. The second case will be discussed in chapter 6. If variables have different scales a transformation to an equal scale is necessary. Different approaches are possible:

- **Theoretical or empirical standardization**
- **Theoretical or empirical transformation to [0,1]**

Theoretical or Empirical Standardization (or z-transformation)

$$z_{gi} = \frac{x_{gi} - \bar{x}_i}{s_i} \; resp. \; z_{gi} = \frac{x_{gi} - \mu_i}{\sigma_i}.$$

$\bar{x}_i$ is the empirical mean of variable i, s_i the empirical standard deviation. The theoretical scale values are μ_i (mean) and σ_i (standard deviation). Note: μ_i and σ_i are not population parameters, they are derived from the property of the scale. Figure 2-1 shows the difference.

Theoretical or Empirical Transformation [0, 1] (or [0, 1]-transformation)

$$z_{gi} = \frac{x_{gi} - a_i}{b_i - a_i} \; resp. \; z_{gi} = \frac{x_{gi} - \alpha_i}{\beta_i - \alpha_i}.$$

a_i is the empirical minimum of variable i, b_i the empirical maximum. The theoretical scale values are α_i and β_i. The theoretical values are derived from the scale.

scale used in a survey ⇒	**results of a survey**
1 = strongly agree	strongly agree (1) = 35%
2 = agree	agree (2) = 45%
3 = disagree	disagree (3) = 20%
4 = strongly disagree	strongly disagree (4) = 0%
⇓	⇓
theoretical scale values	**empirical scale values**
mean = 2.5	mean = 1.850
standard deviation = 1.12	standard deviation = 0.73
minimum = 1	minimum = 1
maximum = 4	maximum = 3

Figure 2-1: Theoretical and empirical scale values

The formulas for computing the theoretical scale values are shown in figure 2-2. Three scale types are distinguished:

- **Scale type I:** Variables with continuous values in a given interval, e.g. percent of votes for party A.
- **Scale type II:** Variables with discrete equidistant values, e.g. response categories of an attitude scale.
- **Scale type III:** Variables with discrete, but not equidistant values, e.g. frequency for a certain leisure activity: 7.0 = daily, 3.5 = several times a week, 1.0 = weekly, 0.2 = less.

theoretical scale values	symbol	scale types		
		scale type I	scale type II	scale type III
minimum	α_i	immediately seen from the scale		
maximum	β_i	immediately seen form the scale		
mean	μ_i	$(\beta_i - \alpha_i)/2$	$(\beta_i - \alpha_i)/2$	$\sum X_{ij}/m_i$
variance	σ_i^2	$(\beta_i - \alpha_i)^2/12$	$\frac{[(m_i+1)\cdot(m_i-1)}{12}$	$\sum (X_{ij} - \mu_i)^2/n$
standard deviation	σ_i	$\sqrt{\sigma_i^2}$	$\sqrt{\sigma_i^2}$	$\sqrt{\sigma_i^2}$

m_i = number of categories of variable, X_{ij} = scale value of category j of variable i.

Figure 2-2: Computation of theoretical scale values

In total, **four methods can be differentiated**:

- theoretical standardization (theoretical z-transformation)
- theoretical normalization to [0,1] (theoretical [0,1]-transformation)
- empirical standardization (empirical z-transformation)
- empirical normalization to [0,1] (empirical [0,1]-transformation)

Empirical normalization to [0,1] has proved to be superior to other standardization methods in a simulation study done by Milligan and Cooper (1988). However, a theoretical and empirical [0,1]-transformation can be problematic. Categories with different content can get

equally transformed values (see chapter 5). Therefore, a general answer to the question 'which method should be used' is not possible.

A **theoretical standardization** has the following advantages:

- The standardized scale values have a clear interpretation. A positive or a negative sign indicates, for example, a positive or negative attitude. In the case of an empirical standardization a negative (or a positive) sign can indicate a positive (or a negative) attitude.
- Variables that separate clusters well are not lower weighted in the classification process. This unwanted effect occurs, if variables are empirically standardized.

However, a high variance in one variable does not necessarily imply good separation. Measurement errors can also result in high variances. Empirical standardization is to be prefered for these variables because they are lower weighted and measurement errors cannot destroy the structure. Another advantage of empirical standardization is the fact that some statistical tests require empirical standardized values.

Practically, it is difficult to decide whether high variances indicate good separation or high measurement errors. Being confronted with the two possibilities in practice (theoretical or empirical standardization), taking the right decision is not easy. However, the question 'theoretical or empirical standardization or normalization' should not be overestimated. Another simulation study by Milligan (1980) suggests that the negative effect of standardization (lower weighting of 'good' variables) is small.

Please take into consideration that a theoretical standardization or normalization is not applicable in all situations. It requires that the scale values, the minimum and maximum of a variable, are defined theoretically. This assumption can be violated, if, for example, economic growth is used. Economic growth has no theoretical minimum or maximum.

Technically, the transformations can be performed by COMPUTE statements in SPSS. The empirical scale values can be computed with the procedures FREQ or DESC. DESC enables you also to save empirically standardized variables (see chapter 5) and can be used to standardize variables instead of COMPUTE statements.

The CLUSTER procedure (agglomerative hierarchical techniques) offers different standardization methods, too (see figure 2-3). Therefore, it is not necessary to use COMPUTE statements or to run DESC in advance. Unfortunately, QUICK CLUSTER (k-means) does not provide such possibilities. CLUSTAN and ALMO provide standardization procedures for agglomerative hierarchical techniques and k-means. CLUSTAN is the most powerful one in this aspect. However, a syntax language is not integrated in CLUSTAN. Therefore, SPSS and ALMO are more flexible.

	SPSS	**CLUSTAN**	**ALMO**
empirical standardization	yes, but only in CLUSTER	yes, for all techniques	yes, for all techniques
theoretical standardization	no	yes, via the definition of weights for all techniques	yes, via the definition of weights for all techniques
empirical normalisation	yes, but only in CLUSTER	yes, for all techniques	yes, via the definition of weights
theoretical normalization	no	yes, via the definition of weights for all techniques	yes, via the definition of weights
additional methods	yes, but only in CLUSTER	yes	yes
transformation via syntax possible	yes	no	yes

Figure 2-3: Transformation techniques offered by different software

2.2 Original Variables or Derived Variables?

Two approaches can be applied, if cases are clustered:

- to use original collected variables or
- to use derived variables.

The **second strategy** has the following advantages:

- Measurement errors are lower for derived variables. Because of that fewer irrelevant variables (like measurement errors) - eventually destroying the cluster structure - are entering the classification process.
- Interpretation is facilitated. Instead of p variables (e.g. 100) only q derived variables (e.g. q = 10) are used and have to be interpreted.

Sometimes, original variables are easier to interpret. So they can be used for interpretation as descriptive or passive variables, but they are not used for classification. Examples (see chapter 1): Münchmeier as well as the SINUS-Institute use variables to describe the clusters that were not used for clustering. This is as long legitimate as the user is informed which variables are used in classification and which variables are used only to describe the clusters.

However, the method used depends on the question analysed. If response patterns to one scale are to be analysed (as in the case of Neumann et al., see chapter 1) the original items have to be used.

Methods to obtain derived variables can be:

- Classical test theory (reliability analysis) for ordinal, quantitative and binary variables.
- Factor analysis for ordinal, quantitative and binary variables. In the case of binary variables, item difficulties should not vary too much.
- Scaling procedures for binary (or ordinal) variables, if items differ in difficulties, like Rasch scaling, Mokken scaling.

- Multiple correspondence analysis for nominal variables.

Reliability analysis, factor analysis and multiple correspondence analysis are available in SPSS. The procedures are: RELIABILITY, FACTOR and HOMALS. Rasch scaling, Mokken scaling or other scaling techniques for binary data are not provided. For example, they are available in ALMO. CLUSTAN does not provide scaling techniques.

References

Bacher, J., 1996: Clusteranalyse [Cluster analysis]. Opladen. [only in German]

Everitt, B., 1981: Cluster analysis. 2nd edition. London-New York.

Fox, J., 1982: Selective Aspects of Measuring Resemblance for Taxonomy. In: Hudson, H. C. (Ed): Classifying Social Data. New Applications of Analytic Methods for Social Science Research. San Francisco-Washington-London, 127-151.

Gordon, A. D., 1999: Classification. 2nd edition. London and others.

Milligan, G. W., 1980: An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. Psychometrika, Vol. 45, 325-342.

Milligan, G. W., Cooper, M.C., 1988: A study of standardization of variables in cluster analysis. Journal of classification. Vol. 5, 181-204.

Chapter 3:
Dissimilarity and Similarity Measures

Note:

For further details see Bacher (1996: 198-232). Similarity and dissimilarity measures are also discussed in Everitt (1980: 12-22), Gordon (1999: 17-23) and other textbooks.

3.1 Overview

Hierarchical methods (except Ward's method, median and centroid method; see next chapter) require the specification of a dissimilarity or similarity measure. In general, four groups of dissimilarity resp. similarity measures can be distinguished:

1. Correlation coefficients, also labelled association measures.
2. Distance measures.
3. Derived measures. They are derived from correlation coefficients or from distances.
4. Other dissimilarity or similarity measures. They have been developed for special purposes, mainly for binary variables.

Distance measures are defined as

$$d(q,r)_{g,g*} = \left[\sum_i \left| x_{gi} - x_{g*i} \right|^r \right]^{1/q}.$$

x_{gi} is the value of case g in variable i, x_{g*i} the value of case g* in variable i.

Correlation measures are defined as

$$r_{g,g*} = \frac{\sum_i (x_{gi} - \bar{x}_g) \cdot (x_{g*i} - \bar{x}_{g*})}{\left(\sum_i (x_{gi} - \bar{x}_g)^2 \cdot \sum_i (x_{g*i} - \bar{x}_{g*})^2 \right)^{1/2}}.$$

In general, correlation coefficients and derived measures based on correlation coefficients are less useful for clustering cases. Distance measures and derived measures based on distances are less useful for clustering variables.

The selection of a similarity or dissimilarity measure depends on the measurement level of variables.

3.2 Binary Variables

A large variety of coefficients has been developed for binary data. SPSS CLUSTER (SPSS 2001) provides 27 similarity or dissimilarity measures for binary variables. The measures for binary variables differ in the following aspects.

1. Conjoint presence (1,1) or (+,+) is weighted differentially.
2. Conjoint absence (0,0) or (-,-) is weighted differentially.
3. Mismatches (1,0) or (0,1) are weighted differentially.

Table 3-1 summarizes some measures using the following numbers and symbols:

		case g*	
		presence	absence
case g	presence (1) or (+)	a = conjoint presence	b = mismatch
	absence (0) or (-)	c = mismatch	c = conjoint absence

similarity coefficient	formula	example	properties
Jaccard's coeff. I	d/(d+b+c)	1/(1+1+1) = 1/3 = 0.33	Conjoint absence (0,0) is ignored.
Dice's coeff.	2d/(2d+b+c)	2·1/(2·1+1+1) = 2/4 = 0.50	Conjoint absence (0,0) is ignored, conjoint presence (1,1) is double weighted.
Sokal&Sneath's coeff. I	d/(d+2(b+c))	1/(1+2·(1+1)) = 1/5 = 0.20	Conjoint absence (0,0) is ignored, mismatches are double weighted.
Russel&Rao's coeff.	d/(d+a+b+c)	1/(1+1+1+1)= 1/4 = 0.25	Conjoint absence (0,0) is not evaluated as similarity, but used in the denominator.
simple matching coeff.	(d+a)/(d+a+b+c)	(1+1)/(1+1+1+1) = 2/4 = 0.50	Absence and presence as well as matches and mismatches have equal weights.
Sokal&Sneath's coeff. II	2(d+a)/(2(d+a)+b+c)	2·(1+1)/(2·(1+1)+1+1) = 4/6 = 0.67	Matches (conjoint absence and presence) are weighted double.
Rogers&Tanimoto's coeff.	(d+a)/(d+a+2(b+c))	(1+1)/(3+2+2(0+1)) = 5/7 = 0.71	Mismatches are weighted double.

Table 3-1: Similarity measures for binary variables (a, b, c and d equal 1)

Further measures are:

- Correlation coefficient Phi
- Coefficient kappa (Fleiss 1981: 217-225; Bacher 1996: 204-206)
- City block distance
- Euclidean distance
- Squared Euclidean measure

For binary variables the last three measures are equal and distance measures.

3.3 Nominal Variables

The following measures can be used for nominal variables, if cases are clustered:

- Simple matching coefficient
- Coefficient kappa for nominal variables (Fleiss 1981: 218-220; Bacher 1996: 212)
- City block metric
- Squared Euclidean distances

For clustering variables, Cramer's V and other association coefficients can be used, too.

3.4 Ordinal Variables

Measures for ordinal variables are:

- Correlation coefficients, like Kendal's tau or Gamma
- City block metric (has an ordinal interpretation)
- Coefficient kappa for ordinal variables

Special measures are:

Canberra Metric (dissimilarity coefficient):

$$Canberra_{g,g*} = \sum_i \frac{\left| x_{g,i} - x_{g*,i} \right|}{\left(x_{g,i} + x_{g*,i} \right)}$$

Jaccard's coefficient II (similarity coefficient):

$$Jaccard - II_{g,g*} = \frac{\sum_i x_{gi} + \sum_i x_{g*i} - 2\sum_i \min(x_{gi}, x_{g*i})}{\sum_i x_{gi} + \sum_i x_{g*i} - \sum_i \min(x_{gi}, x_{g*i})}$$

3.5 Quantitative Variables

Pearson's r can be used as a similarity or correlation measure.

Among the distance measures the following ones are used frequently:

$$\mathrm{CITY}_{g,g^*} = d(r=1,q=1)_{g,g^*} = \sum_i \left| x_{gi} - x_{g^*i} \right|.$$

$$\mathrm{EUKLID}_{g,g^*} = d(r=2,q=2) = \sqrt{\sum_i (x_{gi} - x_{g^*i})^2}.$$

$$\mathrm{QEUKLID}_{g,g^*} = d(r=2,q=1) = \sum_i (x_{g,i} - x_{g^*,i})^2.$$

$$\mathrm{CHEBYCHEV}_{g,g^*} = d(r=\infty,q=\infty)_{g,g^*} = \max_i \left| x_{gi} - x_{g^*i} \right|.$$

3.6 Mixed Levels

see chapter 6.

3.7 Symbolic Variables

Cases can be described by more general data. Gordon (1999: 136) refers to the following categories:

1. Variables can take more than one value or belong to more than one category.
2. Variables can be defined to belong to a specified interval of values.
3. Variables can be defined by continuous or discrete distribution.

Example: Households are clustered. Each household has a certain age distribution (case 3). The income can vary within a certain interval (case 2) and the household members can have different educational levels (case 1).

Similarity and dissimilarity measures for these cases are discussed in Gordon (1999: 136-142).

3.8 Missing Values

Methods to handle missing values are:

- **Listwise deletion** excludes a case from the analysis, if one or more variables are missing. If many variables are used to cluster cases, the number of cases may be reduced dramatically.
- **Pairwise deletion** uses all available information. A case is only eliminated, if the number of missing values exceeds a certain threshold.
- Estimating missing values with **imputation techniques** (see Rubin 1987, Little and Rubin 1987, Gordon 1999: 26-28).

Table 3-2 shows an example. Case g has a missing value in X4, case g* in X3. Both cases would be eliminated by listwise deletion of cases. In contrast to this, pairwise deletion of values would use the Variables X1, X2 and X5 and compute a mean or re-scaled similarity or dissimilarity measure using the following formula:

$$d_{g,g^*} = \sum_j w_{(g,g^*),j} \cdot d_{(g,g^*),j} \Big/ \sum_j w_{(g,g^*),j}$$

where $w_{(g,g^*)j}$ is equal to 1, if case g has a valid value in variable j. Otherwise $w_{(g,g^*)j}$ is equal to 0.
Using the city block metric, the distance between the two cases amounts to 2 (=6/3). The distance can be re-scaled to the original number of variables by multiplying the result with 5 or more generally by $w = \sum_j w_j$.

case	X_1	X_2	X_3	X_4	X_5	SUM
g	2	1	2	MIS	6	-
g*	1	5	MIS	1	5	-
w(g,g*),j	1	1	0	0	1	3
d(g,g*),j (a)	1	4	-	-	1	-
w(g,g*),j * d(g,g*),j	1	4	0	0	1	6
d(g,g*)=6/3=2						

(a) city block metric was used

Table 3-2: Pairwise deletion of missing values

Kaufman (1985) studied the effect of different treatments of missing values for Ward's method. Listwise deletion results in fewer misallocation of cases than pairwise deletion. However, the differences between the two methods were small. No simulation studies for other hierarchical methods or for k-means are known to me.

Perhaps, the following **two step cluster analysis** would result in fewer errors:

1. Use listwise deletion in a first step. Compute the clusters.
2. Assign the cases with missing values to the nearest cluster.

This algorithm has not yet been tested. Therefore, experiences of performance are not available. According to the simulation results of Kaufman, the algorithm should perform better than listwise or pairwise alone, because the proposed procedure combines the advantages of both methods. In the first step listwise deletion is used resulting in fewer misclassification. In the second step additional cases are assigned. Some of them will be assigned correctly.

3.9 Tests for the Absence of a Class Structure

This chapter describes a simple test for the absence of a class structure. The test uses the distribution of similarity and dissimilarity measures assuming the null model that all cases belong to the same population. The steps of the test are:

1. Pick up randomly a pair of cases g and g* and compute the similarity or dissimilarity measure. Delete the cases for further computation.
2. Repeat the first step q times.
3. Test, if the distribution of the computed similarities or dissimilarities differs significantly from the known null distribution. If this is the case, the null hypothesis 'all cases belong to the same population' resp. 'no class structure is present' can be rejected.

Known distributions for the null model are (see table 3-3):

- The squared Euclidean distances have a chi-square distribution in the case of quantitative standardized and independent variables.
- The Euclidean distance and the city block metric have a normal distribution in the case of quantitative standardized and independent variables.
- The city block metric and the simple matching coefficient have a binomial distribution in the case of binary independent variables.

	distribution	mean	variance
quantitative standardized variables			
city block metric (a)	normal	$1.14 \cdot m$	$0.73 \cdot m^2$
Euclidean distance	normal	$\sqrt{2m-1}$	1
squared Euclidean distance	chi square	$2 \cdot m$	$8 \cdot m$

m = number of variables

(a) Deduced from simulation results reported in Schlosser (1976: 126-128, 282-284).

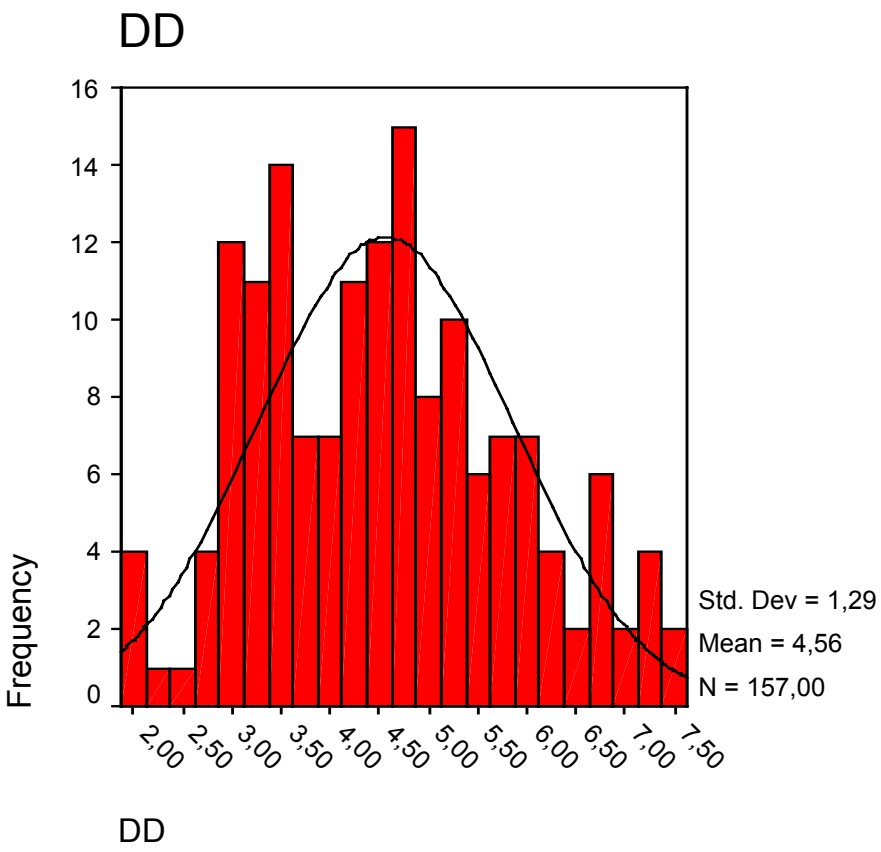
For further details see Bacher (1996: 208-209, 235).

Table 3-3: Distribution of distance measures

More general tests are described by Gordon (1999: 226). The test statistic mentioned above can be computed with a syntax programme in SPSS. The steps are:

1. Generate a random variable and sort the cases using the random variable.
2. Split the data matrix in two data matrices.
3. Match the two files.
4. Compute Euclidean distance (or another measure with a normal or binomial distribution) for each pair. Note: SPSS does not include a test for chi square distribution.
5. (Compute the frequency distribution) and test, whether the distribution is normal or binomial. Kolmogorov-Smirinov's one-sample test can be used for this purpose. Use the theoretical mean and the theoretical standard deviation according to table 3-3.

The syntax is reported in the appendix. Because of the randomised order of the cases your results may differ. The results are:



One-Sample Kolmogorov-Smirnov Test

		DD
N		157
Normal Parameters[a,b]	Mean	4,583
	Std. Deviation	1
Most Extreme Differences	Absolute	,133
	Positive	,133
	Negative	-,079
Kolmogorov-Smirnov Z		1,666
Asymp. Sig. (2-tailed)		,008

a. Test distribution is Normal.

b. User-Specified

The empirical distribution deviates significantly from the null model ('normal distribution'). The conclusion can be drawn that a cluster structure is present.

References

Bacher, J., 1996: Clusteranalyse [Cluster analysis]. Opladen. [only available in German].

Everitt, B., 1980: Cluster analysis. Second edition. New York.

Fleiss, J. L., 1981: Statistical Methods for Rates and Proportions. 2nd Edition. New York-Chichester-Brisbane-Toronto-Singapore.

Gordon, A. D., 1999: Classification. 2nd edition. London-New York.

Kaufman, R. L., 1985: Issues in Multivariate Cluster Analysis. Some Simulations Results. Sociological Methods & Research, Vol. 13, No. 4, 467-486.

Little, R.J.A, Rubin, D.B., 1987: Statistical Analysis with Missing Data. New York.

Rubin, D.B., 1987: Multiple Imputation for Nonresponse in Surveys. New York.

Schlosser, O., 1976: Einführung in die sozialwissenschaftliche Zusammenhangsanalyse [Introduction in Association Analysis in the Social Sciences]. Reinbek bei Hamburg. [only available in German].

SPSS Inc., 2001: Cluster. (SPSS Statistical Algorithms, http://www.spss.com/tech/stat/Algorithms.htm).

Appendix syntax: apriori.sps

```
get file="c:\texte\koeln\spss\km1.sav".

des var=pessi inter trust alien normless viol green spd cdu csu safe /save.

compute xx=rv.uniform(0,1000).
sort cases by xx.

compute set2=0.
compute set2a=lag(set2).
if (set2a eq 0) xx=lag(xx).
if (set2a eq 0) set2=1.
execute.

temp.
select if (set2 = 0).
save outfile="c:\texte\koeln\spss\data1.sav".
execute.
```

```
temp.
select if (set2 = 1).
compute zpessi2=zpessi.
compute zinter2=zinter.
compute ztrust2=ztrust.
compute zalien2=zalien.
compute znorm2=znormles.
compute zviol2=zviol.
compute zgreen2=zgreen.
compute zspd2=zspd.
compute zcdu2=zcdu.
compute zcsu2=zcsu.
compute zsafe2=zsafe.
save outfile="c:\texte\koeln\spss\data2.sav".
execute.

match files
    file="c:\texte\koeln\spss\data1.sav"
   /file="c:\texte\koeln\spss\data2.sav"
   /by xx
   /map.
execute.

compute dd =(zpessi - zpessi2)**2 +
            (zinter - zinter2)**2 +
            (ztrust - ztrust2 )**2 +
            (zalien - zalien2)**2 +
            (znormles -  znorm2)**2 +
            (zviol - zviol2)**2 +
            (zgreen - zgreen2)**2 +
            (zspd - zspd2)**2 +
            (zcdu - zcdu2)**2 +
            (zcsu - zcsu2)**2 +
            (zsafe - zsafe2)**2.
compute dd=sqrt(dd).

FREQUENCIES
  VARIABLES=dd
  /NTILES= 10
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN SKEWNESS SESKEW KURTOSIS SEKURT
  /HISTOGRAM  NORMAL
```

```
  /ORDER=  ANALYSIS .

NPAR TESTS
  /K-S(NORMAL,4.583,1)= dd
  /MISSING ANALYSIS.
```

Chapter 4:
Hierarchical Clustering Techniques -Part I

Note:

This chapter and the following one is based on Bacher (1996: 141-166, 238-278, 297-308). Hierarchical methods are also described in Everitt (1981: 24-34) or Gordon (1999: 78-90).

4.1 Basic Ideas, Algorithms and Methods

Hierarchical clustering techniques build clusters step by step. There are two main approaches:

- Divisive hierarchical techniques
- Agglomerative hierarchical techniques

Divisive methods start with the assumption that all objects are part of a single cluster. The algorithm splits this large cluster step by step until each object is a separate cluster. **Agglomerative** methods start inversely. Each cluster consists of one object. The clusters are combined step by step. In each step those two clusters with the smallest dissimilarity or the highest similarity are merged. Iteration continues until all objects are in one single cluster. Figure 4-1 visualizes the difference between the two methods.

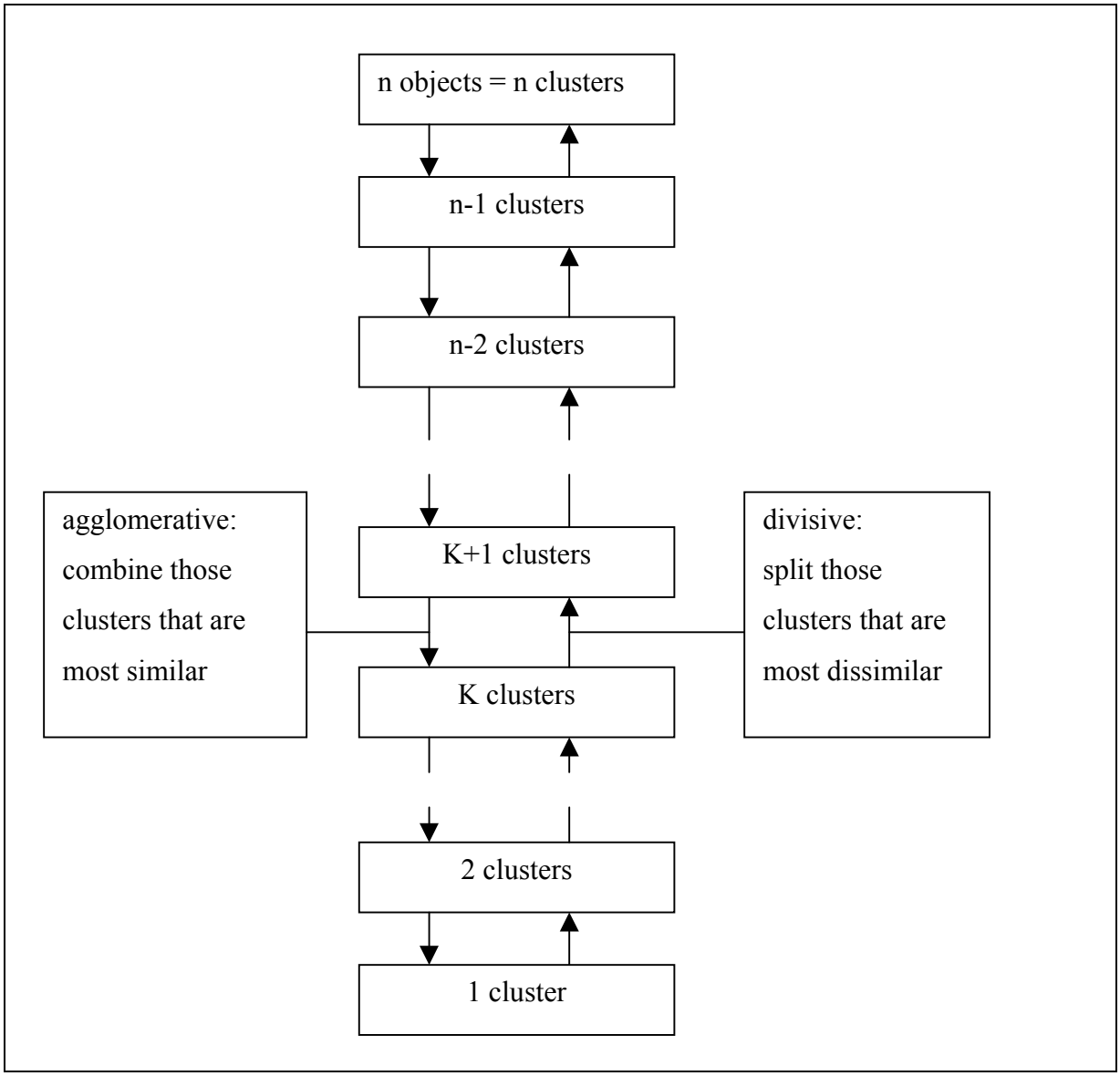


Figure 4-1: Main idea of hierarchical techniques

In general, SPSS and other statistical software packages offer hierarchical agglomerative procedures. Therefore, we will concentrate on these techniques.

Agglomerative hierarchical methods require a similarity or dissimilarity matrix of objects (if objects are clustered) or cases (if cases are clusters). For some methods one of a large variety of similarity or dissimilarity measures (see chapter 3) can be selected, for other methods one measure must be used. The **techniques with a fixed dissimilarity measure** are: Ward's, median and centroid linkage. These methods require squared Euclidean distances.

The methods differ in the way similarities or dissimilarities are re-computed after two clusters are merged. In figure 4-2 the formulas are summarized. It is not important to know the formulas in detail. They are only to show the basic ideas of the methods. The following notation was used:

$d^{new}_{(p+q),i}$ = dissimilarity between the new cluster (p+q) with cluster i

d_{pi} = dissimilarity between the cluster p and cluster i

d_{qi} = dissimilarity between the cluster q and cluster i

$s^{new}_{(p+q),i}$ = similarity between the new cluster (p+q) with cluster i

s_{pi} = similarity between the cluster p and cluster i

s_{qi} = similarity between the cluster q and cluster i

Method	**Formula for the re-calculation of dissimilarities and similarities (a)**
complete linkage	$d^{new}_{(p+q),i} = \max(d_{pi}, d_{qi})$ *res.* $s^{new}_{(p+q),i} = \min(s_{pi}, s_{qi})$
single linkage	$d^{new}_{(p+q),i} = \min(d_{pi}, d_{qi})$ *res.* $s^{new}_{(p+q),i} = \max(s_{pi}, s_{qi})$
average linkage (not available in SPSS)	$d^{new}_{(p+q),i} = \left(d_{pi} + d_{qi}\right)/2$
weighted average linkage (BAVERAGE in SPSS)	$d^{new}_{(p+q),i} = \left(n_p \cdot d_{pi} + n_q \cdot d_{qi}\right)/\left(n_p + n_q\right)$
within average linkage (b) (WAVERAGE in SPSS)	$d^{new}_{(p+q),i} = \left(n_p \cdot d_{pi} + n_q \cdot d_{qi}\right)/\left(n_p + n_q\right)$ $d^{new}_{(p+q),(p+q)} = \frac{\left(\frac{2}{n_p \cdot (n_p - 1)} \cdot d_{pp} + \frac{2}{n_q \cdot (n_q - 1)} \cdot d_{qq} + n_p \cdot n_q \cdot d_{pq}\right)}{\left(n_p + n_q\right) \cdot \left(n_p + n_q - 1\right)/2}$
median linkage	$d^{new}_{(p+q),i} = \frac{1}{2} \cdot d_{pi} + \frac{1}{2} \cdot d_{qi} - \frac{1}{4} \cdot d_{pq}$
centroid linkage	$d^{new}_{(p+q),i} = \frac{n_p}{n_p + n_q} \cdot d_{pi} + \frac{n_q}{n_p + n_q} \cdot d_{qi} - \frac{n_p \cdot n_q}{(n_p + n_q)^2} \cdot d_{pq}$
Ward's linkage	$d^{new}_{(p+q),i} = \frac{1}{n_p + n_q + n_i} \cdot \left[\left(n_p + n_i\right) \cdot d_{pi} + \left(n_q + n_i\right) \cdot d_{qi} - n_i \cdot d_{pq}\right]$

(a) Steinhausen and Langer (1977: 77), Gordon (1999: 78-79), Everitt (1981: 33-34)

(b) Within average linkage requires a further modification. In step k those two clusters are combined that minimize

$$d^{*}_{p,q} = \frac{\left(\frac{2}{n_p \cdot (n_p - 1)} \cdot d_{pp} + \frac{2}{n_q \cdot (n_q - 1)} \cdot d_{qq} + n_p \cdot n_q \cdot d_{pq}\right)}{\left(n_p + n_q\right) \cdot \left(n_p + n_q - 1\right)/2}$$

Figure 4-2: Re-calculation of dissimilarities and similarities for different hierarchical clustering techniques

As figure 4-2 shows, complete linkage and single linkage are extreme procedures with completely different properties. **Complete linkage** uses a max function. This results in a very strong definition of the homogeneity of clusters: The largest dissimilarity between all objects of one cluster should be less than a certain value. Due to this property complete linkage is labelled as **farthest or furthest neighbour method**. The farthest neighbour of each object

should have a distance less than a certain value. In contrast to these requirements **single linkage** only requires that the **nearest neighbour** is located within a certain distance. Therefore, the method is called nearest neighbour. Both methods can be used for similarity or dissimilarity measures.

The advantage of both procedures is their **invariance against monotonic transformation** of the dissimilarities or similarities. The results do not change, if the dissimilarities or similarities are – for example – squared, or if we take the log. The disadvantage is the extreme conception of homogeneity of a cluster. **Single linkage** leads to **chaining** (Everitt 1981: 67-68) and may result in too few large and heterogeneous clusters. **Complete linkage** results in **dilatation** (Gordon 1999: 88) and may produce too many clusters. **Averaging methods**, like average linkage, weighted average linkage, within average linkage, try to avoid these effects. They do not use the min or max function. They compute some kind of average. The following synonyms are used (Bacher 1996: 274):

- Simple average linkage, weighted average linkage, weighted pair group average method (=WPGMA) for average linkage
- Group average, unweighted pair group average method (=UPGMA), group (weighted) average, average linkage, between groups method (SPSS) for weighted average linkage
- Average-Linkage Within Groups (SPSS) for within average linkage

The **last three methods** (median, centroid and Ward) in figure 4-2follow a different logic. They make two assumptions:

- a data file (cases and variables) exists (all other methods require only a dissimilarity or similarity matrix that can be computed from a data file but that can be observed directly too)
- clusters can be described by their centres (means in the variables).

The centres are computed step by step so that a certain criteria is minimized or maximized. **Ward´s method** (also called incremental sum of squares method) minimises the within sum of squares, the **centroid method** and **median method** select in each step those clusters whose centres are closest. They differ in the way the centres are calculated. The methods are primarily designed for clustering cases. Squared Euclidean distances must be used. This

means: interval-scaled variables or variables that can be treated as interval-scaled are necessary and the distances are weighted implicitly. A larger distance in one variable has a higher weight than small distances in many variables.

The **results of hierarchical methods** are usually summarized in an **agglomeration schedule**. Figure 4-4 shows the structure of such a schedule. Five objects with the following dissimilarities were clustered with complete linkage (see figure 4-3).

	object 1	**object 2**	**object 3**	**object 4**	**object 5**
object 1	0.0				
object 2	1.0	0.0			
object 3	2.0	3.0	0.0		
object 4	8.0	9.0	10.0	0.0	
object 5	11.0	12.0	13.0	5.0	0.0

A higher value indicates a higher dissimilarity. Object 1 and 2 are more similar than object 1 and 5, for example.

Figure 4-3: Dissimilarity matrix for five objects

	Clusters combined		
Step or Stage	**Cluster 1**	**Cluster 2**	**Level or Coefficient**
1	1	2	v_1, e.g. 1.0
2	1	3	v_2, e.g. 3.0
3	4	5	v_3, e.g. 5.0
4	1	4	v_4, e.g. 13.0

Figure 4-4: The structure of an agglomeration schedule

The **agglomeration schedule** contains the following information:

- In the first step the cluster to which object 1 belongs is combined with the cluster to which object 2 belongs. The two clusters are merged at a level v1. In the example the level is 1.0.
- In the second step the cluster to which object 1 belongs is combined with the cluster to which object 3 belongs at a level of 3.0. Note that the schedule only enumerates the first

object of a cluster. Cluster 1 in step 2 actually consists of two objects, namely object 1 and 2 that have been merged in the first step.

- In step 3 the clusters to which object 4 and 5 belong are combined at a level of 5.0.
- In step 4 the clusters to which object 1 and object 4 belong are amalgamated. All objects are now in one single cluster because cluster 1 consists of the object 1, 2 and 3 and cluster 2 contains objects 4 and 5. The clusters are merged at a level of 13.0.

The schedule does not inform us which objects belong to a cluster. However, this information may be deduced from the schedule. Its main purpose is to inform about the process and to give some hints about the number of clusters. The agglomeration levels should continuously increase (if dissimilarities are used) or decrease (if similarities are analysed). However, not all techniques have this property. They may cause reversals **of the levels**: Dissimilarities increase until a certain step is reached, after this step the level decreases and increases again. Methods with reversals are: Median linkage, centroid linkage (Gordon 1999: 87) and within average linkage (Bacher 1996: 273).

Some **computer programmes** provide additional information. SPSS shows at which step a cluster appeared first and in which further step it will be merged with another cluster (see Figure 4-5).

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	2	1,000	0	0	2
2	1	3	3,000	1	0	4
3	4	5	5,000	0	0	4
4	1	4	13,000	2	3	0

Figure 4-5: SPSS output

Unfortunately, SPSS does not record the number of clusters, the increase of dissimilarities or the decrease of similarities and ties. **Ties** can influence the results (Gordon 1999: 78). They occur, if more than one pair of 'most similar' clusters are present in a certain step.

The interpretation of the agglomeration level depends on the method used. As shown in figure 4-6, average and median linkage are difficult to interpret. All other methods have a clear interpretation.

Method	**Interpretation of agglomeration level v_i**
complete linkage	The maximum dissimilarity between all objects of one cluster is v_i.
single linkage	For each objet in one cluster a second object with a maximum distance of v_i exists.
average linkage (not available in SPSS)	v_i is the dissimilarity between the two combined clusters if it is assumed that the clusters have equal size during the whole agglomeration process.
weighted average linkage (BAVERAGE)	v_i is the dissimilarity between the objects of the two combined clusters.
within average linkage (b) (WAVERAGE)	v_i is the average dissimilarity between the objects of the new built cluster. We additionally now that the average dissimilarities between the objects within all clusters are less or equal v_i.
median linkage	v_i is the squared Euclidean distance between the centres of the two combined clusters. The centres are computed under the assumption that the clusters are of equal size.
centroid linkage	v_i is the squared Euclidean distance between the centres of the two combined clusters.
Ward's linkage	v_i is the increase of the within cluster sum of squares. SPSS and other programmes print the within cluster sum of squares.

Figure 4-6: Interpretation of agglomeration levels

The agglomeration schedule can be visualized by a so called **dendrogram**. Figure 4-7 shows the dendrogram for the agglomeration schedule of figure 4-4.

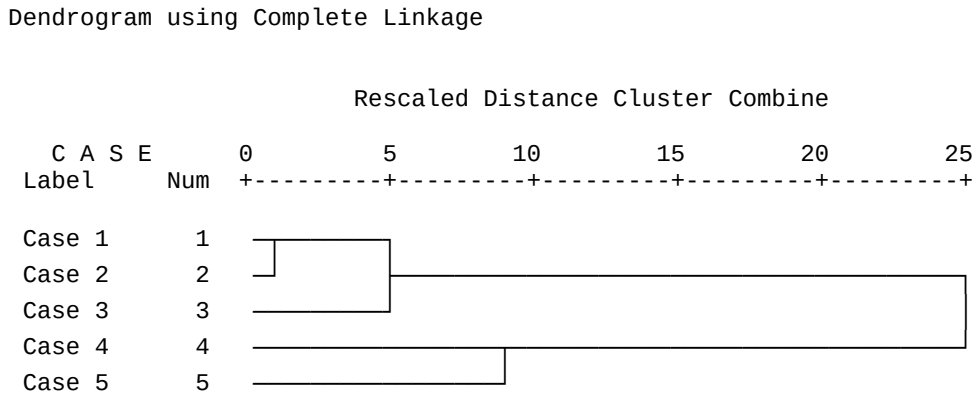


Figure 4-7: Dendrogram of the results of figure 4-4

The **dendrogram** shows which objects are combined in which step. In the example object (Case) 1 and 2 are combined in the first step. In the second step object 3 is merged into the cluster built by object 1 and 2. In the third step object 4 and 5 are combined. Finally all objects are combined to one cluster.

The question '**how many clusters are in the data?'** is difficult to answer. Frequently, the number of clusters is determined on the **basis of the dendrogram.** The user marks the number of 'small' hills (clusters) that combine objects at a low distance level. In figure 4-5 we see immediately one hill (Cluster built by case 1, 2 and 3). However, if also the second cluster (object 4 and 5) is a small hill, it is already difficult to decide. Nonetheless this method usually provides good results.

An **inspection of the agglomeration levels** is another frequently applied approach. The levels are read downwards starting with step 1. The user looks for a sharp increase (if dissimilarities are clustered) or decrease (if similarities are clustered) of the agglomeration levels. In figure 4-5 a considerable increase occurs between step 3 and 4. Therefore, step 3 is accepted as a best cut because step 4 increases the agglomeration level to much. Step 3 corresponds to the solution with two clusters. Cluster 1 contains object 1, 2 and 3, cluster 2 object 4 and 5.

The method described above can be applied graphically in the so called inverse scree test. A plot is constructed in this test (it is not a test in a statistical sense!). The x-axis contains the number of clusters, the y-axis the agglomeration levels. A sharp increase in the agglomeration

schedule results in an elbow knick. In our example a knick occurs at two clusters (see figure 4-6). Therefore two clusters will be selected.

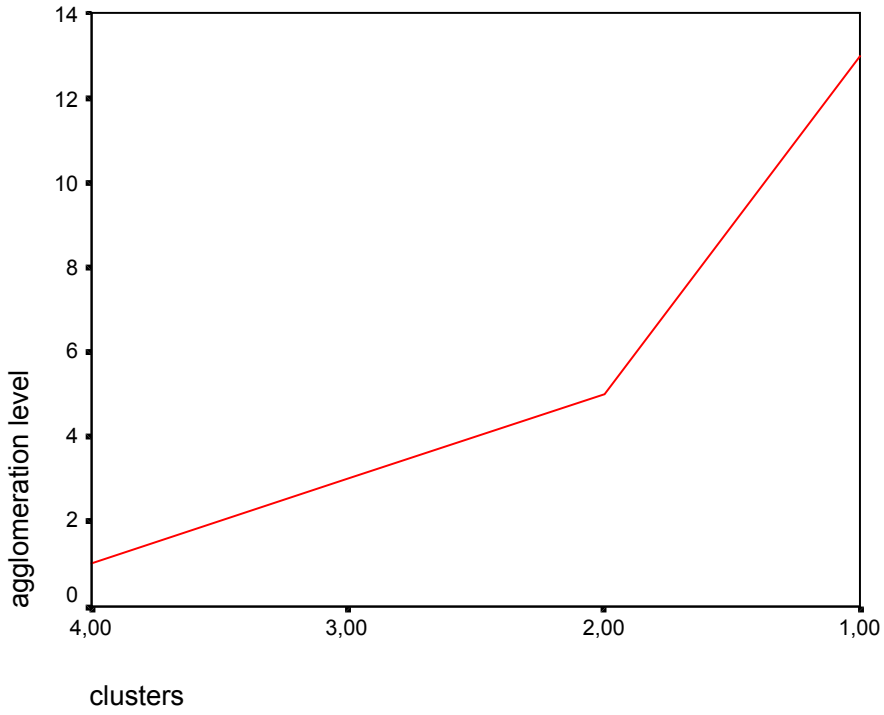


Figure 4-8: Scree diagram of the agglomeration schedule of figure 4-4

The methods usually do not result in a unique solution. It may be difficult to determine the number of hills or to decide if a cluster is a small hill. Very often two or more sharp increases or elbow knicks exist. In this situation it is difficult to select one cut. Strictly speaking, the first elbow should by chosen. But this decision may result in too many clusters. These problems will be discussed in Chapter 4-7.

Agglomerative hierarchical clustering techniques use a simple and easy to understand algorithm. Clusters are built step by step. Different methods with different properties exist. A general answer to the question, which technique should be used, cannot be given. The answer depends on the data used and the analysed question. An **important distinction** is the question whether cases or variables are clustered.

When students or users ask me, which methods they should apply for **clustering cases** I usually recommend Ward's linkage, if interval-scaled variables are analysed or if the variables

can be treated as interval scaled. There are the following reasons for my advice (see figure 4-9 for the following argumentation):

- Ward's linkage uses a criteria that is well known from other statistical procedures, namely within cluster sum of squares. But remember: Ward's linkage does not minimize within sum of squares for a certain solution, e.g. for 4 clusters. It minimizes the increase and may result in a solution with a within cluster sum of square greater than the minimum. This minimum may be reached by k-means clustering.
- There is a clear interpretation of the agglomeration levels. (Which is, for example, not the case for median linkage).
- Ward's linkages guaranties a continuous increase of the agglomeration level. (Which is, for example, not the case for centroid and median method).

In order to test the stability of Ward's linkage I recommend baverage, complete and single linkage and a split of the population.

Squared Euclidean distances must be used as a dissimilarity measure for Ward's method. If a user wants to select another similarity or dissimilarity measure I propose baverage linkage for the following reasons:

- Average linkage is difficult to interpret and the other average method may result in reversals.
- Single and complete linkage use too extreme definitions of cluster homogeneity resulting in too few heterogeneous or too much homogenous clusters. However, both methods should be used to test stability.

For clustering variables I also recommend baverage linkage.

I only would like to suggest single or complete linkage, if the results should be invariant against monotonic transformation. However, different philosophies exist and another expert may recommend single or complete linkage because of this invariance property. It is advisable to decide according to the weight of the different properties of the techniques (see figure 4-9). Gordon (1999: 99) reports a slightly different table.

Method	can be used for clustering cases	useful for clustering variables	different dissimilarity and similarity measures	agglomeration level has a clear interpretation	avoids chaining or dilatation	avoids reversals (d)	invariant to monotonic transformation
complete linkage	yes	yes	yes	yes	yes (a)	yes	yes (e)
single linkage	yes	yes	yes	yes	yes (b)	yes	yes (e)
average linkage	yes	yes	yes	no	no	yes	(yes (f))
weighted average linkage	yes	yes	yes	yes	no	yes	no
within average linkage	yes	yes	yes	yes	no	no	no
median Linkage	yes	no	no	no	(no) (c)	no	no
centroid Linkage	yes	no	no	yes	no	no	no
Ward's Linkage	yes	no	no	yes	no	yes	no

(a) chaining occurs (see text), (b) dilatation occurs (see text), (c) tendency to chaining (Gordon 1999: 67), (d) see text, (e) invariant to monotonic transformation of dissimilarities resp. similiarities (Bacher 1996: 146), (f) invariant to linear transformation of dissimilarities resp. similarities (Bacher 1996: 271)

Figure 4-9: Properties of hierarchical clustering techniques

4.2 A first Application – Clustering Variables

As a first application we are going to replicate Neumann´s et al. results on the perception of different youth cultures (see chapter 1.2). A survey on apprentices will be used. Eight groups (youth cultures) were presented in a list. The respondents were to mark the groups they like.

77 respondents gave no answer. This reduces the valid number of cases from 620 to 543. Football fans or more general sport fans and motor cycle /car fans are the most popular (see table 4-1). In contrast to these groups, violent groups (autonomous groups/squatters/punks, neo-nazis/skins and hooligans) are the most unpopular ones.

variable	label	rel. frequencies (n=543)
v39.01	autonomous groups/squatters/punks	8,5%
V39.02	computerfreaks	21,4%
v39.03	neo nazis/skinheads	8,3%
v39.04	football fans, sport fans	38,4%
v39.05	hooligans	6,5%
v39.06	motorcycle and car fans	33,6%
v39.07	techno-fans	22,3%
v39.08	human right or enviromental activists	26,4%

Table 4-1: Popularity of different youth cultures

We intend to cluster variables, not cases. Each variable presents one object for classification. The aim of the analysis is to answer the question **'which groups are seen similar and belong to the same cluster (culture in this application)?'**.

On the basis of the results of Neumann et al. we expected five clusters (see table 4-2): a right wing youth culture (built by the two variables 'neo nazis/skins' and 'hooligans'), a hard youth culture (built by the variable 'autonomous groups/squatters/punks') and a soft youth culture (built by the variable 'human right and environmental activists'). In addition to Neumann et al. we assume additional clusters: one cluster formed by the two variables 'sport/football fan' and 'motor cycle/car fan' and another cluster combining the two variables 'techno fans' and 'computerfreaks'. The first cluster represents a traditional lower class culture, the second perhaps a culture that is labelled as @-cultures. In contrast to Neumann et al. the popularity of pop groups was not examined. So this cluster will not be reproduced.

Cluster	**variables building the cluster**
clusters reported in Neumann et al.	
right wing youth cultures	neo-nazis/skins (V39.03) hooligans (V39.05)
hard youth cultures	autonomous groups/squatters/punks (V39.01)
soft youth cultures	human right or environmental activists (V39.08)
(pop cultures	not asked in this scale)
additional clusters (youth cultures not asked in Neumann et al.)	
traditional lower class culture	sport/football fans (V39.04) motor cycle/car fans (V39.06)
@-cultures	techno fans (V39.07) computerfreaks (V39.02)

Table 4-2: Expected clusters

We decided to use

- Between average linkage (BAVERAGE) and
- Pearson's correlation r as similarity measure.

BAVERAGE was used as clustering technique in order to avoid chaining (single linkage) and dilatation (complete linkage). In contrast to other average methods the agglomeration levels can be interpreted clearly and reversals are avoided (see chapter 4.1).

Pearson's correlation r is equivalent to the phi-coefficient in this application because the variables are dichotomous ('yes' or 'no'). The correlation coefficient was selected because variables are clustered and we were not interested in size effects that arise in distance functions (see chapter 3).

After selecting the clustering technique and the similarity measure we specified the following syntax programme:

```
GET FILE="c:\texte\koeln\spss\jkult.sav".

CLUSTER  v39.01 v39.02 v39.03 v39.04 v39.05 v39.06 v39.07 v39.08
  /METHOD baverage
  /MEASURE= Correlation
  /PRINT SCHEDULE CLUSTER(2,8)
  /PRINT DISTANCE
  /PLOT DENDROGRAM.
```

The first command reads the data file. The second specifies the cluster analysis. BAVERAGE (between average linkage) is defined as clustering method and the correlation coefficient as similarity measure. The variables are V39.01 to v39.08. The last three subcommands define the output. SPSS will print: the agglomeration schedule, the membership of the objects in the case of 2 to 8 clusters, and the distance matrix (in our example the correlation matrix of variables). A dendrogram will be plotted, too.

The syntax can either be generated by some mouse clicks or by writing the commands immediately in the syntax window.

Stage	Clusters Combined		Coefficient	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	618	619	1,000	0	0	2
2	1	618	1,000	0	1	6
3	614	617	1,000	0	0	6
4	535	616	1,000	0	0	68
5	555	615	1,000	0	0	52
6	1	614	1,000	2	3	10
7	338	613	1,000	0	0	403
8	603	612	1,000	0	0	15
.	.					
538	1	20	,000	537	0	539
539	1	16	,000	538	0	540
540	1	8	,000	539	0	541
541	1	4	-,038	540	468	542
542	1	5	-,143	541	533	0

Table 4-3: Cluster analysis results (agglomeration schedule) obtained form a first specification

The results do not correspond to our expectation. SPSS reports a very long agglomeration schedule (see table 4-3). The reason: The programme clustered cases, not variables.

Two possible approaches solve this problem:

- Transposing the data file using the FLIP command.
- Using PROXIMITY. SPSS uses this solution automatically, if you select the option 'clustering of variables' in CLUSTER.

Both approaches will be discussed.

The SPSS FLIP command allows you to transpose the data file. The syntax in our example is:

```
FLIP
  VARIABLES=v39.01 v39.02 v39.03 v39.04 v39.05 v39.06 v39.07 v39.08
.
```

The new (transposed) data file consists of eight cases. Case 1 is the first variable (=V39.01), case 2 the second variable (=V39.03), and so on. The 'old' cases are the 'new' variables. The old data file contained 620 cases. Therefore, the new data file has 620 variables. VAR001 is the first case, VAR002 the second, and so on. SPSS additionally generates a variable labeled as CASE_LBL. This variable contains the names of the variables. CASE_LBL is used to identify the cases in cluster analysis.

We reran CLUSTER with the following specification:

```
CLUSTER  var001 to var620
  /METHOD baverage
  /MEASURE= Correlation
  /PRINT SCHEDULE
  /PRINT DISTANCE
  /ID case_lbl
  /PLOT DENDROGRAM.
```

The results are still disappointing. SPSS produces the WARNING that there are not enough cases left for cluster analysis and stops the analysis. The reason: SPSS CLUSTER deletes missing values listwise (see chapter 3.8). One case is eliminated, if there are missing values in one or more variables (cases in our example). Because of this a person (=variable in our example), who does not answer the question causes missing values in all variables (=cases in our example). This results in the elimination of all cases.

We reran the analysis with the following specification:

```
get file="c:\texte\koeln\spss\jkult.sav".

compute valid=sum.8(v39.01 to v39.08).
fre var=valid.
select if (valid >= 0).
```

```
fre var=valid.

FLIP
  VARIABLES=v39.01 v39.02 v39.03 v39.04 v39.05 v39.06 v39.07 v39.08.

CLUSTER  var001 to var541
  /METHOD Baverage
  /MEASURE= CORRELATION
  /PRINT SCHEDULE
  /PRINT DISTANCE
  /ID case_lbl
  /PLOT DENDROGRAM.
```

After reading the data file the sum of the eight variables is computed and stored in the variable VALID. VALID has values between 0 and 8 and SYSMIS, if one variable has a missing value. The first FREQ command reports the distribution. The cases with valid answers (no missing values) are selected by the SELECT-IF command in the next step. The next FREQ command checks, if SELECT IF works correctly. The frequencies should not include the category SYSMIS.

Figure 4-10 reports the final results. A first inspection shows that everything is correct now. The variables are clustered now.

The procedure PROXIMITY offers an alternative approach to the FLIP command. It enables you to compute a similarity or dissimilarity matrix of cases or variables. The matrix could be saved as a data file for CLUSTER or other procedures. The syntax for our example is:

```
get file="c:\texte\koeln\spss\jkult.sav".

Proximities v39.01 to v39.08
    /view = variable
    /measure=corr
    /matrix=out("c:\texte\koeln\spss\jk.mat").

CLUSTER  v39.01 to v39.08
  /METHOD Baverage
  /PRINT SCHEDULE
```

```
/PRINT DISTANCE
/ID = varname_
/matrix in("C:\texte\koeln\spss\jk.mat")
/PLOT DENDROGRAM.
```

PROXIMITY computes the dissimilarity resp. similarity matrix of variables (VIEW = VARIABLES). The correlation coefficient is used as similarity measure. The similarity matrix is stored in an external file ('c:\ koeln\spss\output\jk.mat'). CLUSTER reads this matrix as input file. The results are identical to those obtained for the FLIP command. The use of PROXIMITIES is more elegant and avoids the problem of eliminating missing cases. However, the SAVE subcommand in CLUSTER is not available. This subcommand may be needed for further analysis, e.g. for testing stability (see chapter 4.4).

Proximity Matrix

	Correlation between Vectors of Values							
Case	1:V39.01	2:V39.02	3:V39.03	4:V39.04	5:V39.05	6:V39.06	7:V39.07	8:V39.08
1:V39.01	1,000	,003	,028	-,117	,055	-,006	,028	,118
2:V39.02	,003	1,000	-,010	,033	-,045	-,008	,164	,045
3:V39.03	,028	-,010	1,000	,010	,520	-,001	,000	-,089
4:V39.04	-,117	,033	,010	1,000	-,006	,115	,042	-,058
5:V39.05	,055	-,045	,520	-,006	1,000	,020	,058	-,106
6:V39.06	-,006	-,008	-,001	,115	,020	1,000	,051	-,070
7:V39.07	,028	,164	,000	,042	,058	,051	1,000	,001
8:V39.08	,118	,045	-,089	-,058	-,106	-,070	,001	1,000

This is a similarity matrix

Agglomeration schedule using BAVERAGE

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	3	5	,520	0	0	6
2	2	7	,164	0	0	5
3	1	8	,118	0	0	7
4	4	6	,115	0	0	5
5	2	4	,029	2	4	6
6	2	3	,003	5	1	7
7	1	2	-,024	3	6	0

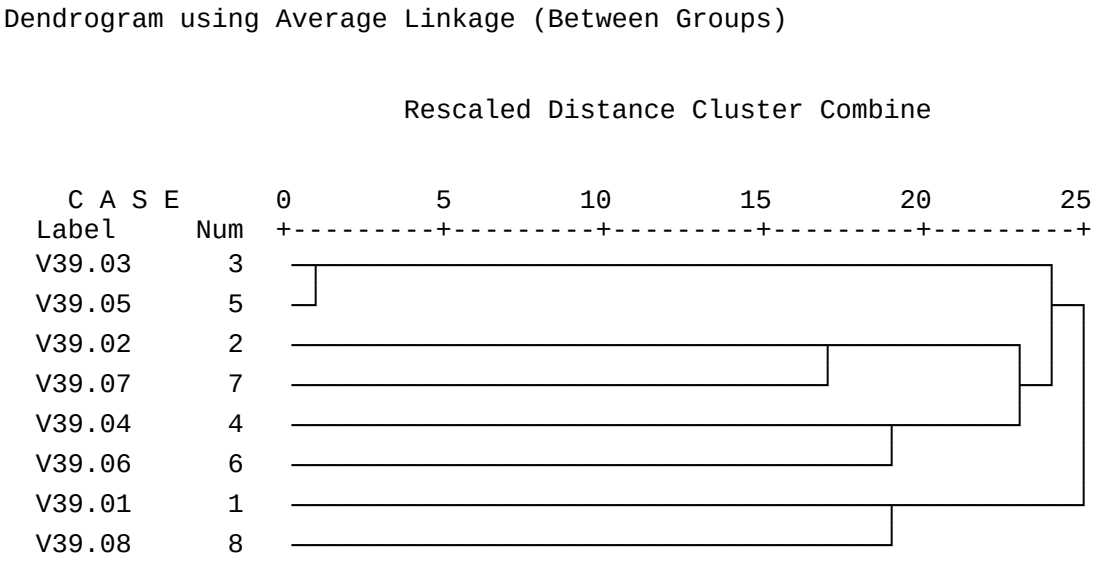


Figure 4-10: Results of SPSS CLUSTER

Variable V39.03 (neo nazis/skins) and V39.05 (hooligans) are agglomerated at a similarity level of 0.520. All other clusters are combined at a lower similarity (agglomeration) level. The next step merges V39.02 (Computerfreaks) and V39.07 (techno fans). The agglomeration level is 0.164. The agglomeration of V39.01 (autonomous groups/squatters/punks) and V39.08 (human right/environmental activists) and of V39.04 (football/sport fans) and V39.06 (motor cylce/car fan) follow at a level of 0.118 and 0.115. The clusters may be labelled as:

- right wing youth cultures (V39.03 and V39.05)
- @-cultures (V39.02 and V39.07)
- activists (V39.01 and V39.08)
- traditional lower class cultures (V39.04 and V39.06)

In contrast to our expectation and the findings of Neumann et al. environmental/human right activists and autonomous groups build one cluster. They belong to two different clusters in Neumann et al. (hard and soft youth cultures).

In the next step (stage 5) the @-cultures and the traditional lower class cultures are combined to one cluster. This cluster is merged in the next step with the right wing youth cultures. Finally, the activists are combined with this cluster.

Number of Clusters.

Chapter 4.1 described approaches to determine the number of clusters. One method is to look for 'small' hills in the dendrogram, another to look for sharp increases or decreases in the agglomeration schedule or equivalently for elbow knicks in the scree diagram.

The first method (looking for small hills) offers no unique solution in our example. It definitely reveals one small hill: the cluster build by V39.03 and V39.05. But it is difficult to decide, if V39.02 and V39.07, V39.04 and V39.06 on the one hand, or V39.01 and V39.08 on the other hand form three additional hills. Two possibilities exist:

- There are seven clusters. All variables – except neo-nazis/skins and hooligans - are separate clusters. The respondents differentiate between seven youth cultures. Only neo-nazis/skins and hooligans are seen to be similar.
- Four clusters exist. The respondents differentiate between right wing cluster, activists, traditional lower class culture and @-cultures.

An inspection of the agglomeration schedule also does not solve the problem. There is a first large decrease from 0.520 to 0.164 between the first and the second stage. This suggests a seven cluster solution. A second decrease occurs between step (stage) 4 and 5. The similarity level decreases from 0.115 to 0.029. This suggests a four cluster solution.

Additional arguments may convince us to accept the four cluster solution: The number of clusters is smaller and the solution corresponds to our theoretical expectations with one exception. However, the first sharp decrease between the first and second step is an argument against this decision. Therefore, we might be interested in further formal criteria. Some will be discussed in chapter 4.4ff.

Chapter 4:
Hierarchical Clustering Techniques - Part II

Note:

This chapter is based on Bacher (1996: 141-166, 238-278, 297-308). Hierarchical methods are also described in Everitt (1981: 24-34) or Gordon (1999: 78-90).

4.3 A second Application – Clustering Cases

Preferences for certain leisure time activities are one group of variables frequently used in life style research. Lechner (2000), for example, uses different leisure time activities in her study. We will use one subscale. In our survey of apprentices the respondents were questioned the activities they liked most. They were to mark the most preferred activities with a cross. A list of 18 activities was used. Table 4-4 summarises the frequencies for a random subsample of 150 cases.

Which leisure time activities do you like most?	**in % (n=150)**
repairing a car, motor cycle, bike and going about	39.3%
playing a computer game on the computer or a slot machine	33.3%
painting, photographing	31.3%
shopping	53.3%
reading	37.3%
listening to music	85.3%
lazing away	69.3%
practising sports	56.0%
going to a disco, dancing	61.3%
going to the cinema	60.7%
attending a pop concert, rock concert	24.0%
going to a party	70.7%
playing an instrument	16.7%
visiting a theater, a museum, going to classical concert etc.	13.3%
studying	11.3%
viewing television	67.3%
visiting a centre for young people	15.3%
doing something forbidden	18.7%

Table 4-4: Most preferred leisure time activities

The aim of the analysis is to answer the question '**can we identify clusters of respondents with different preferences?'**. We decided to use Ward's method, because dichotomous variables can be treated as interval-scaled.

Ward clustered 150 cases. So the agglomeration schedule consists of 149 steps. Only the last steps (e.g. the last 20 or 30 ones) are of interest. However, it is not possible to truncate the schedule in SPSS at a certain step. SPSS always reports all steps. If 500 cases are clustered, 499 steps are printed. In table 4-5 the first steps were deleted manually.

	combined clusters		**coefficient**	**stage cluster 1st appears**		**next stage**
Stage	**Cluster 1**	**Cluster 2**		**Cluster 1**	**Cluster 2**	
.						
.						
130	1	14	262,568	127	115	145
131	2	5	269,604	125	121	143
132	3	27	276,715	108	101	139
133	13	23	284,049	92	84	142
134	6	22	291,518	104	106	138
135	39	44	298,991	118	114	141
136	15	16	307,395	123	129	144
137	18	32	315,905	122	126	140
138	6	21	324,604	134	119	142
139	3	7	333,339	132	124	145
140	18	33	342,848	137	120	143
141	9	39	352,803	128	135	146
142	6	13	363,034	138	133	146
143	2	18	376,191	131	140	148
144	15	31	390,829	136	113	147
145	1	3	406,294	130	139	147
146	6	9	422,717	142	141	148
147	1	15	447,079	145	144	149
148	2	6	477,039	143	146	149
149	1	2	521,867	147	148	0

Table 4-5: Agglomeration schedule for clustering cases

It is difficult to realize a sharp increase in the SPSS output. Also the scree diagram reveals no elbow or very soft elbows depending on the scaling of the x axis (see figure 4-11a to figure 4-11c).

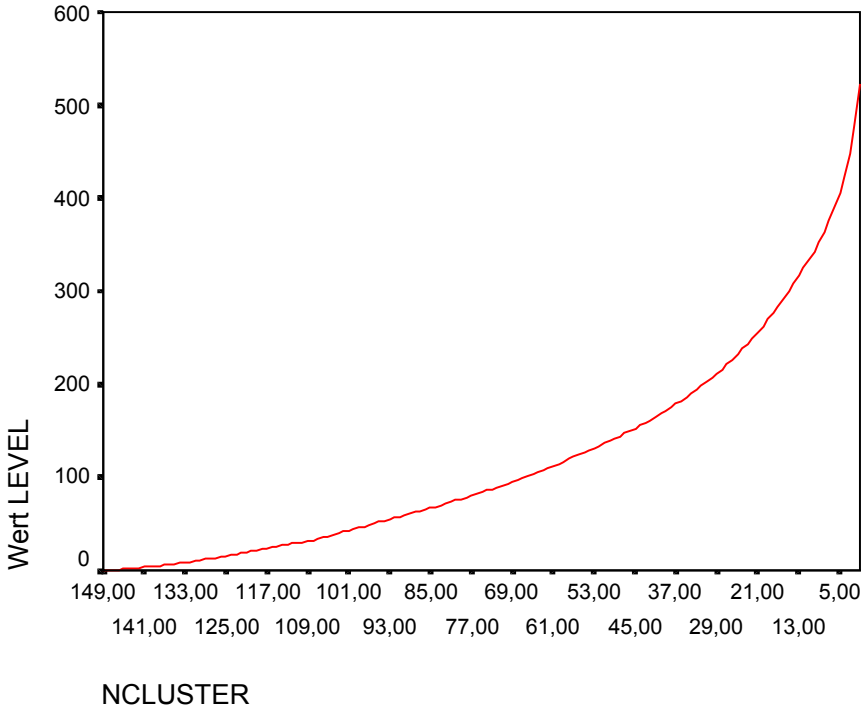
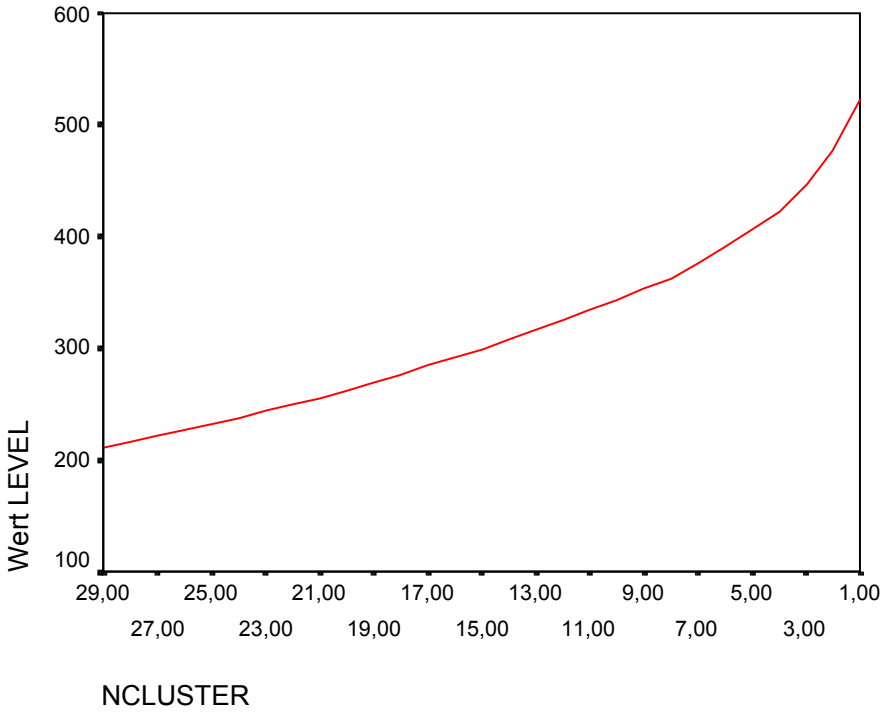


Figure 4-11a: Scree diagram for all solutions (number of clusters from 149 to 1)



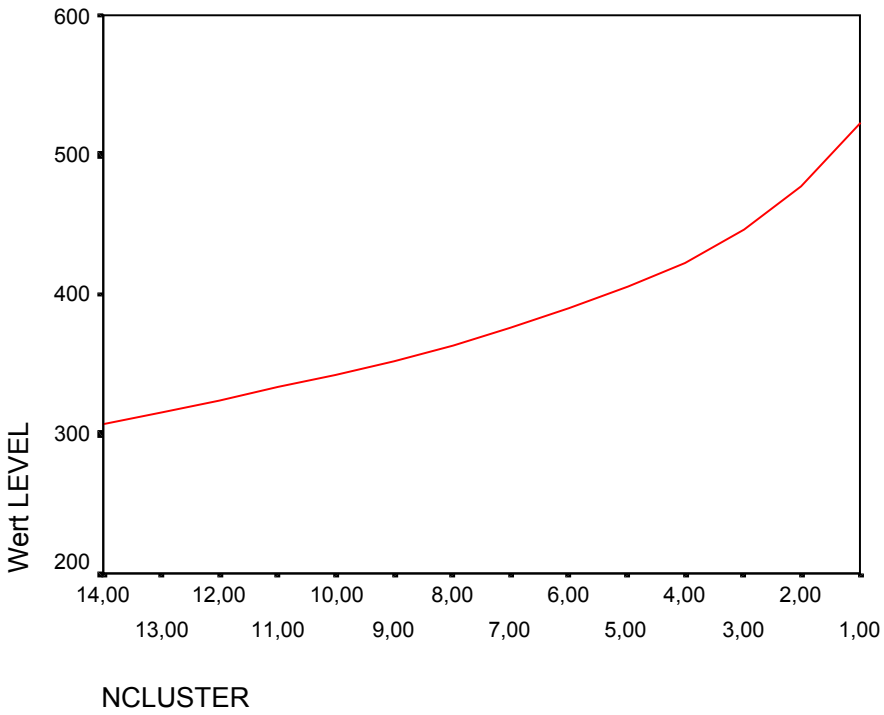


Figure 4-11b: Scree diagram for all solutions (number of clusters from 29 to 1)

Figure 4-11c: Scree diagram for all solutions (number of clusters from 14 to 1)

Interpretation is easier, if the differences between two consecutive steps are computed. The results are:

NCLUSTER	LEVEL	PLEVEL	DLEVEL
19,00	269,60	262,57	7,04
18,00	276,71	269,60	7,11
17,00	284,05	276,71	7,33
16,00	291,52	284,05	7,47
15,00	298,99	291,52	7,47
14,00	307,40	298,99	8,40
13,00	315,90	307,40	8,51
12,00	324,60	315,90	8,70
11,00	333,34	324,60	8,74
10,00	342,85	333,34	9,51
9,00	352,80	342,85	9,95
8,00	363,03	352,80	10,23
7,00	376,19	363,03	13,16
6,00	390,83	376,19	14,64
5,00	406,29	390,83	15,46
4,00	422,72	406,29	16,42
3,00	447,08	422,72	24,36
2,00	477,04	447,08	29,96
1,00	521,87	477,04	44,83

NCLUSTER is the number of clusters, LEVEL the agglomeration level (coefficient), PLEVEL the agglomeration level of the previous solution, DLEVEL is the difference. Example: the level of the three cluster solution is 447,08, the level of the previous four cluster solution is 422,72. Therefore, the difference has a value of 24,36.

Increases can be observed between 8 and 7 clusters, 4 and 3 clusters and 2 and 1 cluster. Therefore, a solution with 8, 4 or 2 clusters may be accepted. According to the scree test criteria the first increase is important favouring the 8 cluster solution. However, it is difficult to judge, if this increase is significant. If this is not the case the solution with 4 clusters would be the best one.

This is the syntax for those who want to reproduce the results:

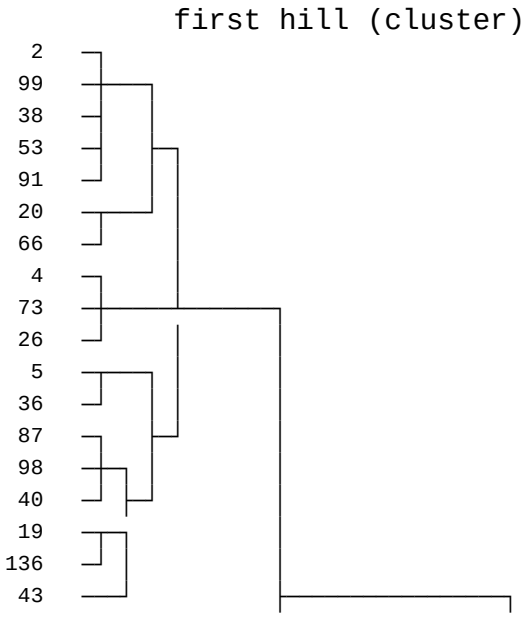
```
data list free/step cluster1 cluster2 level first1 first2 next.
begin data.
1     2     99    ,000  0     0     43
2     38    53    ,000  0     0     15
3     126   150   ,500  0     0     51
4     127   143   1,000 0     0     16
5     50    140   1,500 0     0     20
6     112   130   2,000 0     0     18
7     74    125   2,500 0     0     86
8     28    117   3,000 0     0     17
9     23    115   3,500 0     0     84
10    10    100   4,000 0     0     19
11    37    93    4,500 0     0     103
.
.
141   9     39    352,803    128   135   146
142   6     13    363,034    138   133   146
143   2     18    376,191    131   140   148
144   15    31    390,829    136   113   147
145   1     3     406,294    130   139   147
146   6     9     422,717    142   141   148
147   1     15    447,079    145   144   149
148   2     6     477,039    143   146   149
149   1     2     521,867    147   148   0
end data.

compute plevel=lag(level).
compute dlevel=level-plevel.
compute ncluster=150-step.

temp.
select if (ncluster < 20).
list var=ncluster level plevel dlevel.

GRAPH
  /LINE(SIMPLE)=VALUE( level) BY ncluster .

temp.
```

```
select if (ncluster < 15).
GRAPH
  /LINE(SIMPLE)=VALUE( level) BY ncluster
```

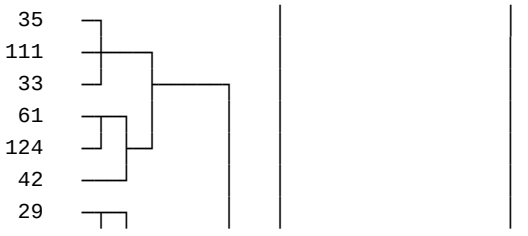
The procedure was: The agglomeration schedule was marked and copied to the pin board. Afterwards we opened a syntax window and pasted the schedule in the window. We defined the schedule as a new data file (DATA LIST ...). PLEVEL was computed using a LAG variable, NCLUSTER as 150 minus step and DLEVEL as the difference of LEVEL and PLEVEL. The variables were printed with the LIST command. Finally the scree diagrammes were plotted with the procedure GRAPH.

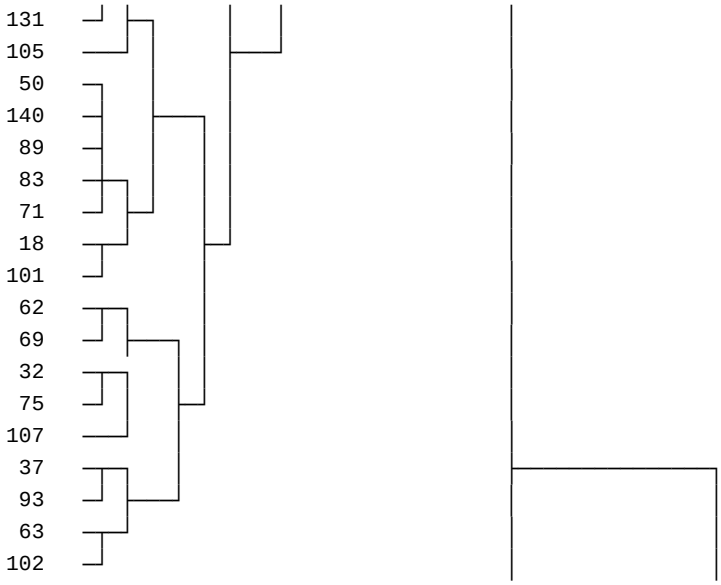
The dendrogram is difficult to interpret, too. It consists of more than one page (see figure 4-10). Unfortunately, SPSS does not offer an option to truncate the dendrogram for a large number of clusters so that the dendrogram only requires one page. This generally facilitates interpretation and allow to discover the number of clusters. Nonetheless, four hills can be distinguished. But perhaps there are more.

Dendrogram using Ward Method

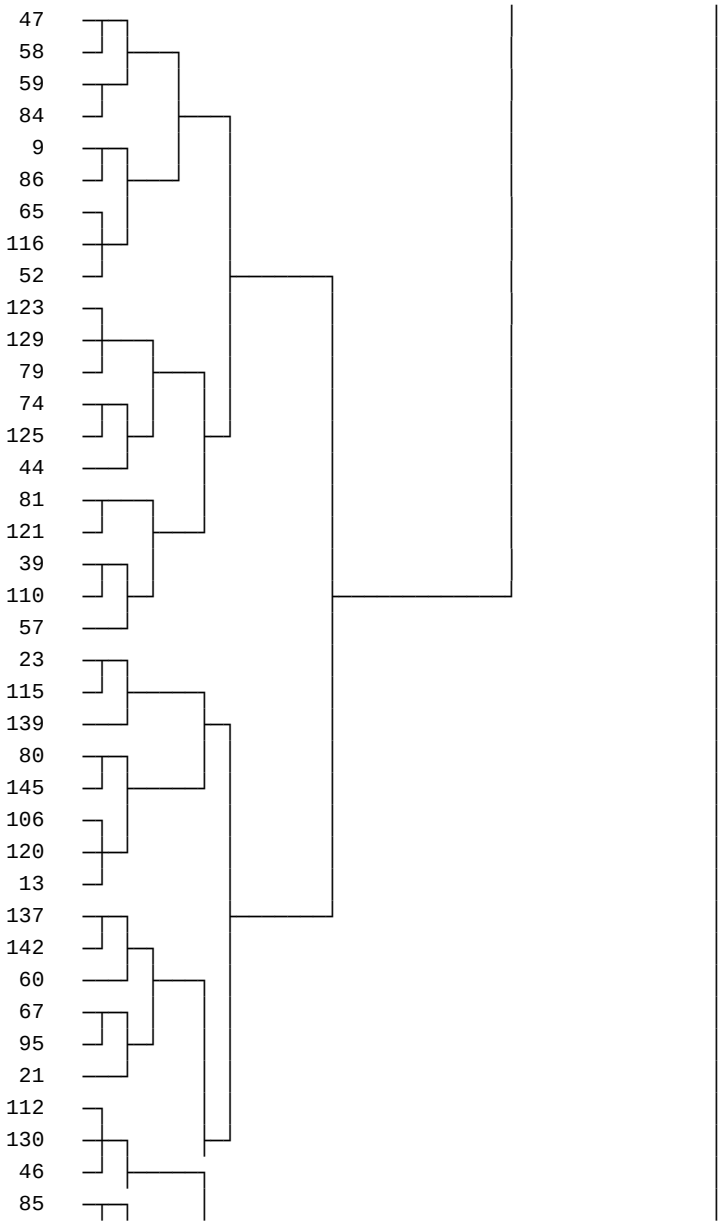


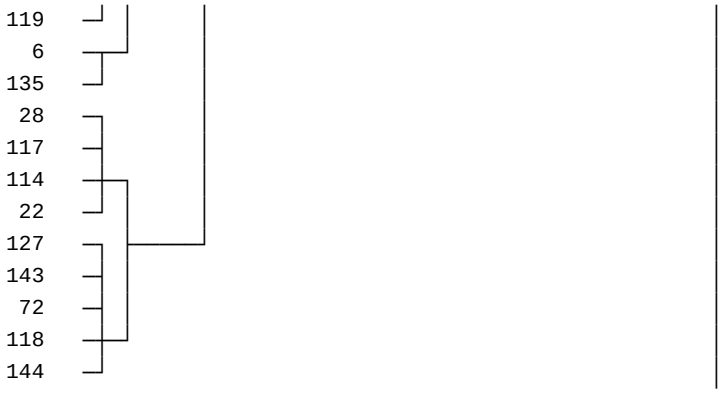
consists perhaps of two further hills (clusters)?



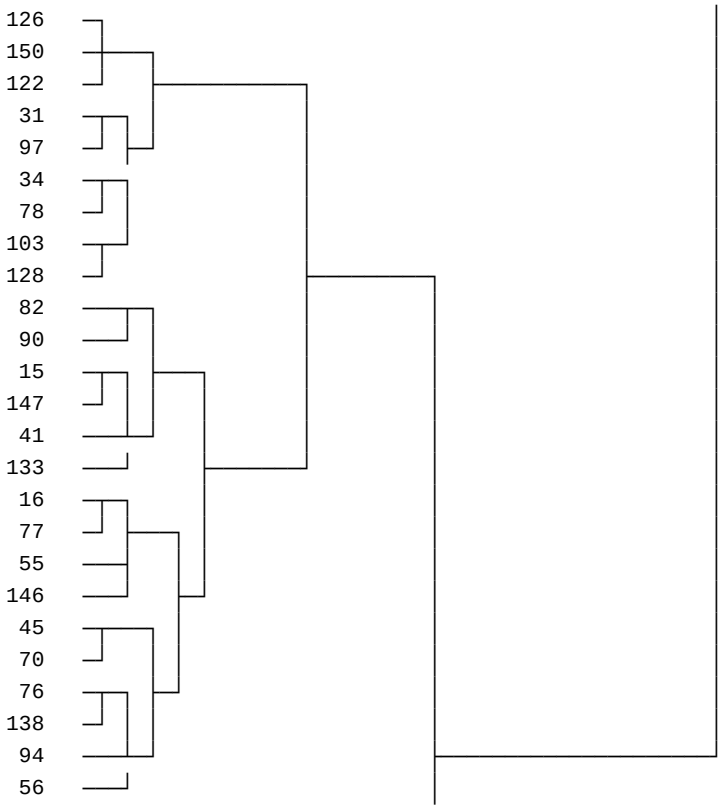


a second hill

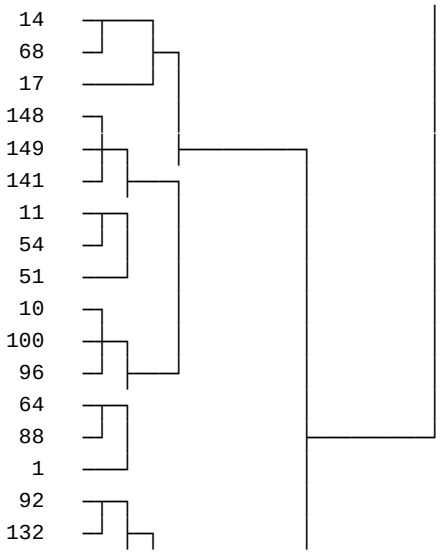


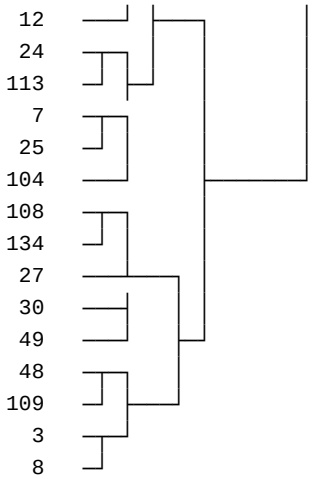


a third hill



a fourth hill





As already mentioned, interpretation becomes easier if the dendrogram is truncated.

Summarizingly, the application of CLUSTER could be facilitated, if SPSS offers some options to specify the output, to print and to plot only the last steps of the schedule and to report the increase and the number of clusters.

Additionally, a statistical test supporting the decision on the number of clusters would be an advantage. We will discuss such tests in chapter 4-7. However, note that the question 'can I substantially interpret the clusters?' is as important as formal aspects for fixing the number of clusters.

In the case of clustering cases we are very often not interested in the classification of cases (the information that respondent 136 is assigned to cluster 1 is of little value), but in some characteristics of the clusters we are, e.g. in cluster centres. CLUSTER does not compute cluster centres. They can be generated using the following procedure:

- Saving the membership of the objects to the clusters in CLUSTER
- Computing the centres by MEANS using the cluster membership as grouping variable.

The syntax corresponding to this procedure is:

```
CLUSTER  v31.01 v31.02 v31.03 v31.04 v31.05 v31.06 v31.07 v31.08
         v31.09 v31.10 v31.11 v31.12 v31.13 v31.14 v31.15 v31.16
         v31.17 v31.18
  /METHOD WARD
```

```
  /MEASURE= SEUCLID
  /PRINT SCHEDULE
  /SAVE CLUSTER(4)
  /PLOT DENDROGRAM.

MEANS
  TABLES=v31.01 v31.02 v31.03 v31.04 v31.05 v31.06 v31.07 v31.08
         v31.09 v31.10 v31.11 v31.12 v31.13 v31.14 v31.15 v31.16
         v31.17 v31.18  BY clu4_1
  /CELLS MEAN COUNT .
```

The SAVE command forces CLUSTTER to add a new variable to the data file. This variable is automatically labeled as clus4_1. (If we rerun the syntax without reading the data file again SPSS generates a variable clus4_2, and so on. If the number of clusters is changed, e.g. to 2, SPSS creates a variable, e.g. Clus2_1.)

The MEAN command computes the cluster centres (see table 4-6).

Ward's method		car, motor cycle	playing compu-ter game	painting	shopping	reading	listening music	lazing away	practi-cing sports	going to a disco, dancing
1	Mean	,63	,13	,44	,25	,44	,94	,53	,53	,38
	N	32	32	32	32	32	32	32	32	32
2	Mean	,70	,42	,28	,56	,12	,86	,86	,81	,98
	N	43	43	43	43	43	43	43	43	43
3	Mean	,16	,16	,40	,86	,66	,94	,76	,44	,60
	N	50	50	50	50	50	50	50	50	50
4	Mean	,28	,52	,08	,20	,08	,56	,56	,48	,36
	N	25	25	25	25	25	25	25	25	25
Total	Mean	,43	,29	,32	,53	,36	,85	,71	,57	,62
	N	150	150	150	150	150	150	150	150	150

ward's method		going to the cinema	going to a pop, rock concert	going to a party	playing an instrument	visiting a theater, ...	studying	viewing television	going to a centre for young people	doing something forbidden
1	Mean	,28	,03	,06	,47	,09	,22	,63	,09	,09
	N	32	32	32	32	32	32	32	32	32
2	Mean	,93	,40	,98	,12	,12	,26	,86	,21	,21
	N	43	43	43	43	43	43	43	43	43
3	Mean	,82	,18	,74	,08	,22	,06	,62	,08	,18
	N	50	50	50	50	50	50	50	50	50
4	Mean	,24	,12	,80	,08	,00	,12	,44	,12	,64
	N	25	25	25	25	25	25	25	25	25
Total	Mean	,64	,20	,67	,17	,13	,16	,66	,13	,25
	N	150	150	150	150	150	150	150	150	150

Table 4-6: Cluster centres for the solution with four clusters

Table 4-6 is difficult to interpret. The following strategies facilitate interpretation:

1. The variables are standardized before the MEAN procedure is used. The resulting means have positive or negative signs. The total mean is zero. A positive sign indicates a value above the total mean, a negative below the total mean. Values larger (smaller) than a certain threshold can be seen as important. Interpretation can focus on those values.
2. Instead of MEAN a discriminant analysis is used. The classification (membership) is used as the dependent variable, the variables used to cluster the cases are taken as the independent variables. Discriminant analysis computes information which variables separate the clusters best. The interpretation can concentrate on these variables.
3. The variables are reduced before the cases are clustered by factor analysis or another appropriate scaling method. The derived scales (e.g. factors) are used in CLUSTER and MEANS. This strategy facilitates interpretation because fewer variables are used in both procedures.
4. A typical case for a cluster is used instead of the means. These typical cases – also called cluster exemplars (Wishart 1999), mediods (Kaufman and Rousseeuw 1990: 69), leading

case (Hartigan 1975: 74-83) or centrotypes - have only empirically observed values, in our case 0 or 1, and are easier to interpret.

Strategy 1 and 2 may be combined, as well as strategy 2 and 3. Strategy 1 is indirectly applied in strategy 3, because derived scales are usually standardized as it is the case for factor analysis.

4.4 Stability of a Cluster Solution

A cluster solution is said to be stable, if a small modification of the method specified and the data used does not change the results too much. Clustering techniques and the dissimilarity (resp. similarity) measure may be modified in a stability analysis of methods. They may be changed in some application. In the example of chapter 4.2 (clustering variables) all other methods – except Ward, median and centroid – may be specified as clustering methods. Other coefficients - like the simple matching coefficient or Jaccard's coefficient (if positive and negative matches should have different weight) - may be used instead of the correlation coefficient. In the example of chapter 4.3 (clustering cases) all other clustering methods may be used, but squared Euclidean distances must be used for Ward´s method.

Practically, stability is tested by varying the specifications of CLUSTER and rerunning CLUSTER. Figure 4-12 summarizes the results of single linkage and complete linkage, if these two methods are used instead of BAVERAGE. Both methods lead to the same four cluster solution. We can formulate the hypothesis 'the four cluster solution is stable in reference to the clustering techniques' because single and complete linkage have extreme opposite properties and lead to the same solution for four clusters. This is not the case for the solution with three clusters. The solutions are:

	Cluster 1	Cluster 2	Cluster 3
single	{v39.03, v39.05, v39.02, v39.07}	{v39.01, v39.08}	{v39.04, v39.08}
complete	{v39.03, v39.05, v39.04, v39.06}	{v39.01, v39.08}	{v39.04, v39.08}
baverage	{v39.03, v39.05}	{v39.01, v39.08}	{v39.02, v39.07, V39.04, v39.08}

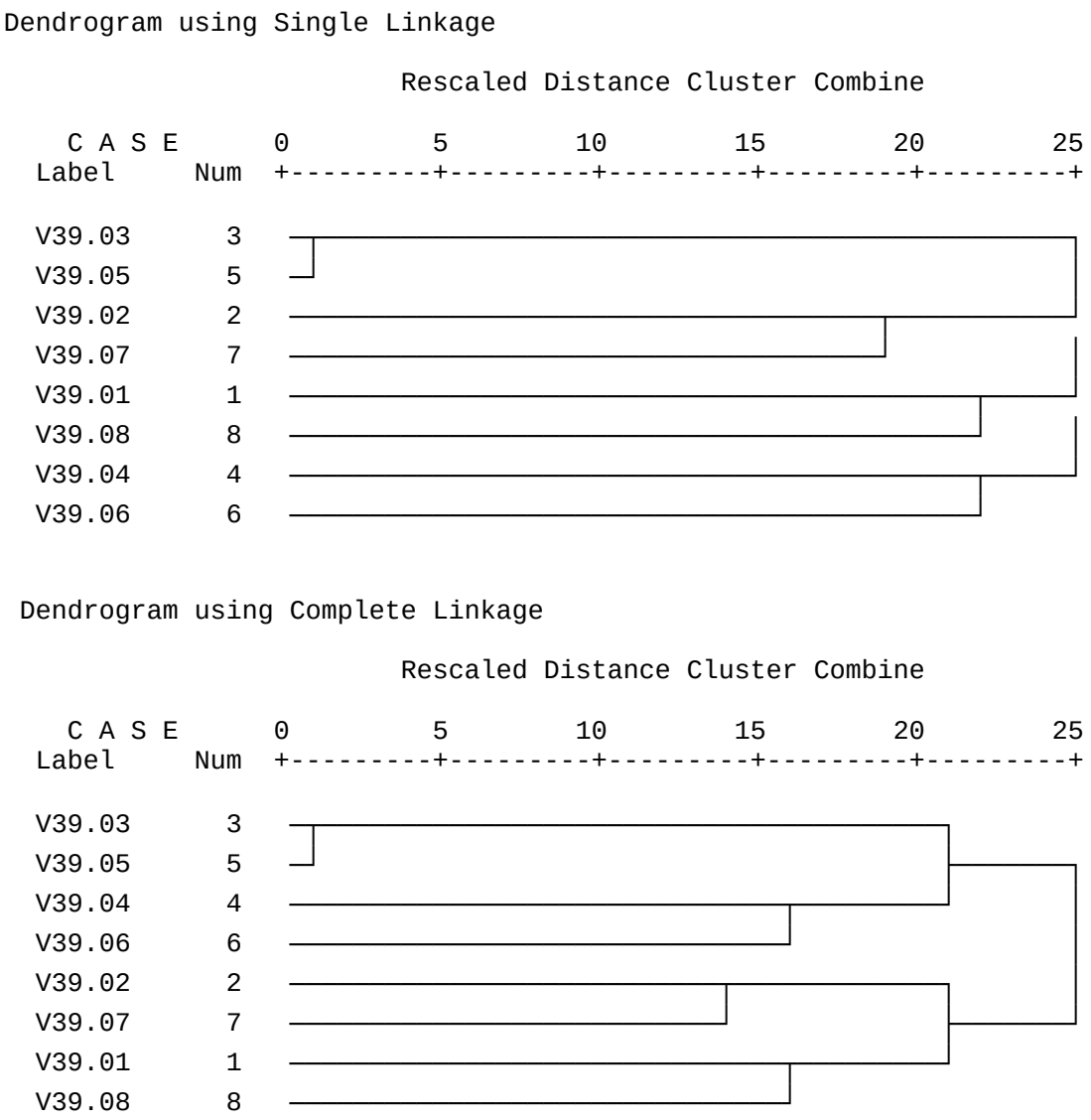


Figure 4-12: Results of complete and single linkage for the example of chapter 4.2

The procedure is straightforward, if stability according to the to similarity measure are to be tested. The effort increases, if both factors (similarity measures and clustering techniques) are being analysed. It is necessary to combine both factors. The design is shown in figure 4-13. The number of analyses is 12, if, for example, four similarity measures and three methods are allowed.

possible clustering techniques	similarity measure			
	measure 1	**measure 2**	**...**	**measure k**
complete	analysis 1	analysis 4	...	analysis 3.k-2
baverage	analysis 2	analysis 5	...	analysis 3.k-1
single	analysis 3	analysis 6	...	analysis 3.k

Figure 4-13: Concept for testing stability against method and similarity measure

Complexity increases further, if the survey population is to be studied as a third factor that might influence stability. If the results are stable, randomly drawn subsamples (or subsamples characterized by certain variables) should produce similar results. If this factor should also be analysed simultaneously, a factorial design with three factors is the result. For five subsamples, four similarity measures and three different methods 60 analyses must be compared. Beside the population, the variables are a second factor characterizing data. Stability according to variables may be tested, for example, by adding random variables.

One strategy to reduce this amount is to split the analysis of stability:

- testing stability against techniques and similarity measures in a first step
- analysing stability according to population in a second step
- analysing stability according to variables in a third step.

Statistical measures comparing classifications, dendrograms and cluster centres of different solutions are discussed in the next chapters. Note: Do not be influenced by these statistics to plan your analysis more untidy.

4.5 Comparing Classifications

Stability was tested visually by comparing the dendrograms for different solutions in the previous chapter. This technique is troublesome for a larger number of objects (cases). An

index or a test statistic for the similarity of different classifications could overcome this problems.

The **Rand index** (Rand 1971) is used for this purpose. The index compares two classifications of objects. The number of clusters may be different for the two solutions. The Rand index is

$$\text{Rand}(CL_i, CL_j) = \frac{2}{n \cdot (n-1)} \cdot \sum_{g} \sum_{g^*>g} r_{g,g^*}$$

where r_{g,g^*} is equal to 1, if two objects g and g* belong in both classifications CL_i and CL_j to the same cluster or in both clusters to different clusters. The Rand index measures the proportion of consistent allocations (two objects belong in both classifications to the same cluster or to different clusters). The number of consistent allocations is $n \cdot (n-1)/2$ and the Rand index is equal 1 for a perfect fit. Values greater than 0.7 are considered as sufficient (Dreger 1986, Fraboni and Salstone 1992). Among others Morey and Agresti (1984) proposed a modification of the Rand index. Hubert and Arabie (1985 quoted in Gordon 1999: 198) modified the index so that its maximum is 1 and its expected value is zero, if the classifications are selected randomly.

SPSS does not offer the Rand index. It may be computed with a syntax program (see appendix) or by using another software (like ALMO).

The Rand index may be used to compare classifications that result from clustering cases or variables.

4.6 Comparing Cluster Centres

These methods will be discussed within k means clustering. This method can be used to test the stability of cluster centres against method, similarity measure and population.

4.7 Statistical Tests of the Number of Clusters

Mojena (1977) formalized the method of the inverse scree test. He developed the following null models for statistical testing.

Model 1: The agglomeration levels are normally distributed by mean

$$v_k = \sum_{i=1}^{k} v_i / k$$

and standard deviation

$$s_k = \sqrt{\frac{1}{k-1}\sum_i (v_i - \bar{v}_k)^2}$$

The test analyses, if level k+1 is a sample point of this normal distribution. If this hypothesis does not fit to the data, a significant sharp increase occurs and the number of clusters is set equal to k. The test statistic is

$$t_1 = (v_{k+1} - \bar{v}_k)/s_k$$

According to Mojena (1977) the null hypothesis 'k+1 belongs to the same distribution' should be rejected for values of t_1 between 2,75 and 3,50. This rule corresponds to a (one sided) significance level of at least 99,7%.

Model 2: The agglomeration levels v_i up to step k can be described by a linear regression line

$$\hat{v}_k = a_k + b_k \cdot k.$$

The test analyses, if level k+1 can be predicted using this regression line. If this is not the case, a significant (sharp) increase is assumed and the number of clusters is equal to k. The test statistic is defined as

$$t_2 = (v_{k+1} - \hat{v}_{k+1})/s_k$$

whereas

$$\hat{v}_{k+1} = a_k + b_k \cdot (k+1)$$

According to Mojena a sharp (significant) increase occurs if t_2 is greater than 2,75. This again corresponds to a (one sided) significance level of 99,7%.

Both statistics are not available in SPSS. They may be computed by a syntax programme. The regression module can be used for t_2. Both approaches in SPSS are time consuming. An acceptable alternative might be the use of an alternative computer software.

Both statistics depend – like other statistics - on the number of cases. A small increase at a very early stage in the agglomeration may become significant because n is large. This may result in a large number of clusters. Conversely no significant increase may occur if n is small.

For our example of chapter 4-2 the values of t_1 and t_2 are:

```
Teststatstik zur Bestimmung der Clusterzahl
nach MOJENA (Regel 1)

Clusterzahl   Teststatistik   Signifikanz
     7             -              -
     6          -0.887            -
     5          -0.694          69.477
     4          -1.024          79.261
     3          -0.974          79.809
     2          -0.973          80.615
```

Teststatstik zur Bestimmung der Clusterzahl
nach MOJENA (Regel 2)

Clusterzahl	Teststatistik	Signifikanz
7	-	-
6	1.235	-
5	1.131	76.898
4	0.592	69.357
3	0.643	71.597
2	0.625	71.562

The values are not significant. One reason: The number of cases (n=8) is too small.

In contrast to this result t_1 is significant for 34 clusters if applied to the example of chapter 4.3, where 150 cases are analysed. Using t_2 a test value larger than 2,75 occurs for seven clusters.

Teststatstik zur Bestimmung der Clusterzahl nach MOJENA (Regel 1)

Clusterzahl	Teststatistik	Signifikanz	
.			
.			
35	2.722	99.625	
34	2.924	99.792	****
33	2.868	99.757	
32	2.798	99.707	
31	2.879	99.766	
30	2.808	99.715	
29	3.081	99.867	
28	3.059	99.858	
27	3.291	99.928	
26	3.154	99.892	
25	3.178	99.900	
24	3.119	99.882	
23	3.075	99.866	
22	3.207	99.909	
21	3.208	99.909	
20	3.336	99.938	
19	3.481	99.962	
18	3.601	99.975	
17	3.709	99.984	
16	3.583	99.974	
15	3.729	99.986	
14	4.064	100.000	
13	3.944	99.998	
12	3.785	99.990	
11	3.607	99.976	
10	4.036	100.000	
9	3.915	99.997	
8	4.056	100.000	
7	4.408	100.000	
6	5.287	100.000	
5	5.106	100.000	
4	6.355	100.000	
3	8.334	100.000	
2	11.280	100.000	

Teststatstik zur Bestimmung der Cluster nach MOJENA (Regel 2)

Clusterzahl	Teststatistik	Signifika	
35	1.028	84.688	
34	1.235	89.039	
33	1.186	88.101	
32	1.122	86.795	
31	1.209	88.540	
30	1.143	87.228	
29	1.420	92.094	
28	1.406	91.888	
27	1.646	94.877	
26	1.518	93.427	
25	1.551	93.827	
24	1.500	93.188	
23	1.461	92.676	
22	1.599	94.384	
21	1.607	94.475	
20	1.741	95.803	
19	1.894	96.990	
18	2.024	97.759	
17	2.143	98.308	
16	2.030	97.792	
15	2.186	98.474	
14	2.532	99.379	
13	2.429	99.178	
12	2.283	98.804	
11	2.115	98.196	
10	2.551	99.413	
9	2.442	99.208	
8	2.593	99.481	
7	2.957	99.816	***
6	3.853	99.994	
5	3.708	99.985	
4	4.983	100.000	
3	7.018	100.000	
2	10.066	100.000	

In our example t_2 seems to perform better. However, Whishart reports better results for t1. (Whishart 2001).

4.8 Measuring Homogeneity

The agglomeration levels provide only a vague idea of the homogeneity within the clusters. For some methods the levels are not well defined. For others the levels reflect extreme information as it is the case for single and complete linkage. So the user may be interested in **additional information on homogeneity**, even if the levels inform you about heterogeneity. Measures of homogeneity can be used in this situation. They assume that the average of the dissimilarities $\bar{u}_{in}$ within the clusters is smaller than the average of dissimilarities $\bar{u}_{zw}$ between the clusters for a good classification. Different approaches have been proposed to calculate the average (Klastorin 1983). We will discuss one. The dissimilarities within and between the clusters are computed in this approach as:

$$\bar{u}_{in} = \sum_{k} \bar{u}(k)_{in} / K,$$
$$\bar{u}_{zw} = \sum_{k} \sum_{k^*>k} \bar{u}(k,k^*)_{zw} / (K \cdot (K-1)/2),$$
$$\bar{u}(k)_{in} = \sum_{g \in k} \sum_{\substack{g^* \in k \\ g^* > g}} u_{g,g^*} / (n_k \cdot (n_k - 1)/2) \text{ and}$$
$$\bar{u}(k,k^*)_{zw} = \sum_{g \in k} \sum_{g^* \in k^*} u_{g,g^*} / (n_k \cdot n_{k^*}),$$

where K is the number of clusters. n_k is the number of cases in cluster k.

To test homogeneity the difference

$$g = \bar{u}_{zw} - \bar{u}_{in}$$

or the ratio

$$\bar{u}_{in} / \bar{u}_{zw}$$

can be computed. The first expression has the advantage that it is easy to construct a statistical test.

The standard normal distribution or Chebyschev´s inequality can be used to compute a significance level testing the null hypothesis 'the dissimilarities within clusters are equal to or larger than the dissimilarities between the clusters'. The statistic is (Klastorin 1983: 95)[1]

$z=(g-E(g))/\sigma(g)$,

where E(g) is the expected value of the null hypothesis that g is zero. s(g) is the corresponding standard deviation. Chebyschev´s inequality is more conservative and favors the null hypothesis.

The statistic results in the following values, if it is applied to the four cluster solution of the example of chapter 4-2:

```
G1-Homogenitaetsmass =  -0.235
Erwartungswert       =   0.000
Varianz              =   0.004
z-Wert               =  -3.535
Signifikanz          =  99.952
Fehler (Chebychev)   =   0.000
```

The g index is negative in our example because similarities are analysed. The z-value is –3.5. The null hypothesis 'the dissimilarities within the clusters are equal to or larger than the dissimilarities between the clusters' can be rejected. Standard normal distribution leads to a significance of 99,95% (error 1 = 0,05%). Chebychev's inequality results in an error level of 0.

Note: A higher g index does not necessarily refer to a higher homogeneity. Only the z values give evidence to this: A higher z value refers to a higher heterogeneity.

[1] Klastorin´s formula for variance has an error. Hubert und Levin (1977) report the correct formula.

4.9 Measuring the Fit of a Dendrogram

The dendrogram or equivalently the agglomeration schedule defines a hierarchy between the objects. Some objects are more similar than others. The agglomeration schedule resp. the dendrogramm enables you to compute a theoretical similarity or dissimilarity matrix. The theoretical or predicted dissimilarity (or similarity) between two objects g and g* combined at a certain step is set equal to the agglomeration level of this step. In the example of chapter 4.2 this rule results in the following predicted or theoretical similarities:

- V39.03 and V39.05 (object 3 and 5) are combined at a level of 0.520. So the theoretical similarity between these two objects is 0.520.
- V39.02 and V39.07 (object 2 and 7) are combined at a level of 0.164 in the next step. This results in a theoretical similarity of 0.164 between these two objects.
- V39.01 and V39.08 (object 1 and 8) are combined at a level of 0.118 in the next step. This results in a theoretical similarity of 0.118 between these two objects.
- V39.04 and V39.06 (object 4 and 6) are combined at a level of 0.115 in the next step. This results in a theoretical similarity of 0.115 between these two objects.
- The clusters that contain object 2 and 4 (v39.02 and v39.04) are combined. Note that only the first object in each cluster is enumerated. The cluster containing object 2 consists of the object 2 and 7, the cluster containing object 4 consists of object 4 and 8. The similarities between clusters (object 2 and object 4, object 2 and object 8, object 7 and object 4 and object 7 and 4) are set equal to the agglomeration level of 0.029.
- In the next step the clusters containing object 2 and object 3 are combined. The first cluster consists of the objects 2, 7, 4 and 8. The second one of the objects 3 and 5. All similarities between the two clusters (object 2 with object 3, object 2 with object 5, object 7 with object 5, and so on) are set to 0.003.
- Finally the clusters containing object 1 and object 2 are merged to one large cluster. The similarities between the two clusters are set to –0.024.

Figure 4-13 summarizes the values of the theoretical similarity matrix.

	V39.01	V39.02	V39.03	V39.04	V39.05	V39.06	V39.07
V39.01							
V39.02	-0.024						
V39.03	-0.024	0.003					
V39.04	-0.024	0.029	0.003				
V39.05	-0.024	0.003	0.520	0.003			
V39.06	-0.024	0.029	0.003	0.115	0.003		
V39.07	-0.024	0.164	0.003	0.029	0.003	0.029	
V39.08	0.118	-0.024	-0.024	-0.024	-0.024	-0.024	-0.024

Figure 4-13: Theoretical similarity matrix for the analyzed objects of chapter 4-2.

The matrix allows you to develop statistics to answer the question 'how good does the hierarchical structure fit to the data?'. The idea: Compute a statistic that compares the theoretical and the empirical similarity or dissimilarity matrix.

The **correlation gamma** was proposed as one criteria:

$$\Gamma = \frac{S(+) - S(-)}{S(+) + S(-)},$$

where S(+) is the sum of concordant relations: the objects (i,j) are more similar (or less similar) than objects (k,l) in both matrices. S(-) is the sum of discordant relations: the objects (i,j) are more similar (or dissimilar) than objects (k,l) empirically, but more dissimilar (or similar) than objects (k,l) theoretically. Only the elements below the diagonal (or the elements above) of both matrices are used for the computation.

The test statistic

$$z(\Gamma) = n \cdot \Gamma - 1.8 \cdot \ln(n)$$

is approximately standard normal distributed (Hubert 1974). An exact test distribution can be derived by simulation (Jain and Dubes 1988: 167-170).

The so called **cophenetic correlation** is another criteria. The cophenentic correlation is defined as Pearson´s r between the elements of the theoretical and the empirical similarity (or dissimilarity) matrix below (or above) the diagonal. The cophenetic correlation is

$$Coph(\mathbf{U},\tilde{\mathbf{U}}) = s(\mathbf{U},\tilde{\mathbf{U}}) \big/ s(\mathbf{U}) \cdot s(\tilde{\mathbf{U}}),$$

where $s(\mathbf{U},\tilde{\mathbf{U}})$ is the covariance between the elements of the theoretical and empirical similarity (or dissimilarity) matrix. $s(\mathbf{U})$ is the standard deviation of the elements of the empirical similarity (or dissimilarity) matrix, $s(\tilde{\mathbf{U}})$ the standard deviation of the elements of the theoretical similarity (or dissimilarity) matrix.

Further measures may be developed on the basis of some other goodness of fit index. For example one might apply the Stress coefficient of MDS .

SPSS does not provide both measures or other measures. They can be computed, if the empirical and the theoretical similarities are typed in the syntax or data file by the user. The user has to calculate the theoretical values manually, too.

The alternative ALMO software (see chapter 4.13) offers both measures. For our example ALMO computes the following gamma and cophenetic correlation for the baverage solution:

```
**********************************************************************

kophenetischer Korrelationskoeffizient =    0.930

Zahl der vorgeg.  Simulationen = 200
Zahl der erfolgr. Simulationen = 200
Erwartungswert        =    -0.01
Standardabweichung    =     0.16
Teststatistik         =     5.87
Schwellwert fuer      =     95 Prozent
                      =     0.27
----------------------------------------------------------------------
```

```
Gamma  =    0.554

Zahl der vorgeg.  Simulationen = 200
Zahl der erfolgr. Simulationen = 200
Erwartungswert        =     0.03
Standardabweichung    =     0.19
Teststatistik         =     2.73
Schwellwert fuer      =     95 Prozent
                      =     0.35
-----------------------------------------------------------------
```

4.10 Comparing Dendrograms

The methods described in the previous chapter can be used to compare dendrograms of different solutions. In contrast to the previous chapter two or more theoretical dissimilarities matrices are compared. (In chapter 4.9 the empirical and the theoretical similarity or dissimilarity matrices were compared.)

4.11 Measuring the Fit of a Classification

The methods described in chapter 4.9 can be used to compare the classification of different solutions, too. A theoretical dissimilarity (or similarity) matrix is computed for a certain cluster solution and compared with the empirical dissimilarity (or similarity) matrix. For more details see Bacher (1996: 253-254).

4.12 Graphical Presentation of a Cluster Solution

Sometimes cluster solutions are presented graphically on a two-dimensional space. Such a presentation can be obtained as follows:

- Run MDS for the dissimilarity (or similarity) matrix.
- Use the two-dimensional solution.
- Mark the clusters by circles.

This method is predominately used in the case of clustering variables. Figure 4-14 summarizes the results of the example of chapter 4-2. The circles were drawn manually (dotted lines = solution for four clusters; continuos line = clusters combined in the next step). The dimensional representation was calculated with Kruskal's nonmetric approach (Kruskal 1964a, 1964b). The statistic system ALMO (Holm 2000) was used to run MDS. The fit of the MDS is not important in this application. MDS is only used as an auxiliary tool. However, the fit should not be too poor. Otherwise, it my happen that objects of one cluster have a large distance. In the example the stress with the value of 0.077 was fair (Kruskal 19964a: 3).

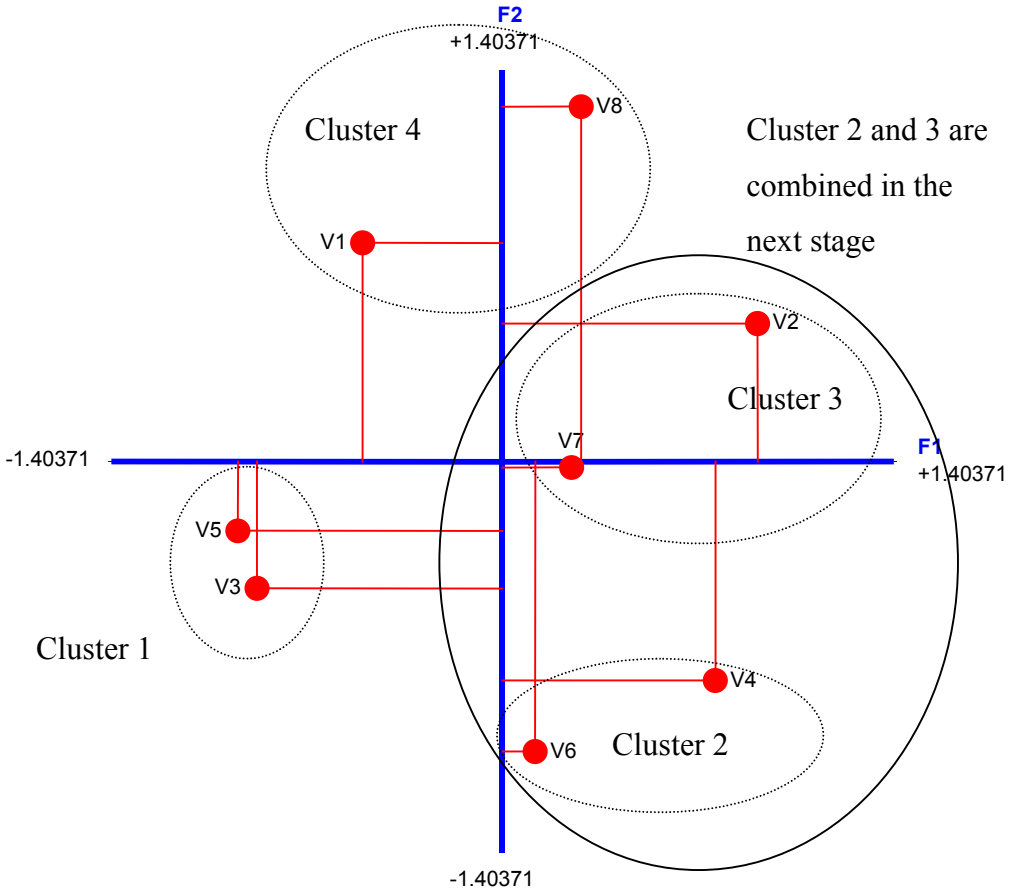


Figure 4-14: Graphical representation of a cluster solution

4.13 Comparison with other Methods

Clustering variables has become an unusual method. Factor analysis and multidimensional scaling are more frequently used to analyse the relations between variables.

Factor analysis has the following advantages:

- Approved rules to determine the number of factors exist.
- Groups of variables can be identified.

- Relations between these groups are defined by factor loadings and factor correlations.
- Scores measuring on interval scale level can be computed for the cases.

Factor analysis requires correlations (or covariances) as similarity measures. This may be a disadvantage. MDS and cluster analysis allow you to use other similarity or dissimilarity measures. This is their advantage.

Both methods (MDS and hierarchical cluster analysis) use different models. MDS assumes that objects are represented as points in a multidimensional space. Very often a two-dimensional Euclidean space is used. In contrast cluster analysis assumes that each object is part of a tree. Holman (1972) showed theoretically that these models are mutually exclusive and exhaustive. 'The distances among a set of n objects will be strictly monotonically related either to the distances in a hierarchical clustering systems, or else to the distances in a Euclidean space of less than n-1 dimensions, but not to both.' Therefore, either MDS with few dimensions or an hierarchy (computed with a cluster analysis) will fit to data.

The simulation study of Pruzansky et al. (1982) supports this conclusions. They used the program system KYST for metric and nonmetric multidimensional scaling and ADDTREE for clustering. ADDTREE is a clustering technique that generalizes hierarchical techniques. Their results show that the appropriate model fit the data better than the inappropriate model for all noise (error) levels. The two models were comparable: KYST explains plan data as well as ADDTREE explains tree data. If noise increases the differences between the models decreases.

The **following procedure** may be applied to select the appropriate model in practise:

- Analyse the data set with MDS and hierarchical cluster analysis. Use MDS, if MDS is able to reproduce satisfactorily the relations between the variables whereas the fit of cluster analysis is poor. Use cluster analysis, if the results are reciprocal (poor fit for MDS with few dimensions and good fit for hierarchical cluster analysis). The measures discussed in chapter 4.9 can be used for this purpose. Pruzansky et al. (1982) used correlation coefficients (the cophenetic correlation and a non-metric correlation coefficient) and distance measures (stress formula 2).

Additionally, the skewness of the distribution of distances can be computed for the model selection. Pruzansky et al. (1982) showed that skewness is positive for distance data and negative for tree data.

However, note: In practice, both methods can fit to the data well. This is the case if clusters exist in a dimensional space and if one is not interested in computing a hierarchy.

A disadvantage of MDS to some extent is the fact that MDS requires a larger number of objects (variables) to produce stable results (Bacher 1996: 83-84).

4.14 Alternative Software Products

The following two software packages offer advanced techniques for hierarchical cluster analysis:

Clustan Limited
16 Kingsburgh Road
Edingburgh EH12 6DZ
Scotland
Web: http:\\www.clustan.com
email: sales@clustan.com

ALMO Statistiksystem
Prof. Dr. Kurt Holm
Universtiät Linz
Web: http:\\www.uni-linz.ac.at
email: kurt.holm@jku.at

CLUSTAN is developed by Wishart (1999). The author programmed the cluster modules in ALMO.

additional features compared to SPSS	**ALMO**	**CLUSTAN**
weighting variables	• yes	• yes
additional hierarchical techniques	• complete linkage for over-lapping clusters • generalized nearest neighbour	• William´s flexible strategy • density method
information about ties	• yes	• yes
elaborated dendrogram	• yes, dendrogram can be truncated	• yes
information on cluster centres	• yes	• yes
Mojena´s rules	• yes	• yes (a)
measures for comparing classifications	• yes, Rand index	• no
additional homogeneity measure	• yes, index g	• no
measures for the fit of a dendrogram	• yes, gamma and cophenetic correlation	• no
measures for comparing dendrograms	• no, in preparation	• no
measures for the fit of a classification	• yes, gamma and cophenetic correlation	• no

(a) Mojena´s t1 is labelled as upper tail rule, Mojena´s t2 as moving average quality controll rule (Wishart 2000)

Figure 4-13: ALMO and CLUSTAN compared to SPSS CLUSTER

CLUSTAN offers a highly elaborated dendrogram technique. It allows you to cut a dendrogram and does not use ASCII codes as it is the case in SPSS and ALMO. Therefore, a

dendrogram of many cases is represented on one page. As a further advantage, CLUSTAN is able to cluster large data files by hierarchical methods.

Additional test statistics are the great advantage of **ALMO.** They were discussed in the previous chapters.

4.15 Factors Influencing the Results

A variety of simulation studies has analysed factors influencing the results of a cluster analysis (summarizing Bacher 1996: 164). Milligan (1980), for example, analysed the following factors:

1. outliers: 20 resp. 40 % outliers were added to the assumed cluster structure.
2. irrelevant variables: One or two random variables were added.
3. wrong dissimilarity measure: Catell´s similarity coefficient and the q correlation were used instead of Euclidean distances.
4. method: The following methods were used: single, complete, weighted average, within average, centroid, median, Ward, k means with random starting values, k means with starting values from weighted average linkage.

In total 9,7% of wrong classification occurred. This level of wrong classification is very low considering that the data file contains 30% outliers and 27% irrelevant variables (1.5 from 5.5) on the average. Among the factors analysed (outliers, irrelevant variables, wrong similarity measure, method) the irrelevant variables are most important. Two irrelevant variables increase the error from 9.7% to 18.0%. The differences between the methods are small. K-means performed best, if starting values of weighted average linkage are used. Ward´s method and complete linkage are sensitive to outliers. Punj and Stewart (1983) reached a similar conclusion in their review on twelve simulation studies. The selection of variables seems to be the crucial point (Milligan 1980; Milligan and Cooper 1987; Green et al. 1990).

However, it is difficult to generalize simulation results. They depend on the specification of the experiment and the evaluation criteria used. Most simulation studies were done in the 70s

and 80s. They do not include new methods. Nonetheless, they give some information on the importance of factors influencing clustering methods.

4.16 Further Developments

Further developments of hierarchical cluster analysis took place in the following fields:

- generalization of neighbour methods (e.g. Bacher 1996: 264-270; Jain and Dubes 1988: 128-130; Gowda and Krishna 1978)
- methods for overlapping clustering (e.g. Bacher 1996: 261-262, Opitz and Wiedemann 1989)
- methods that allow restrictions (e.g. Ferligoj and Batagelj 1982, 1983)
- methods to compare dendrograms (e.g. the so-called consensus methods, Gordon 1999)

References

Bacher, J., 1996: Clusteranalyse [Cluster analysis]. Opladen. [only available in German].

Dreger, R. M., 1986: Microcomputer Programs for the Rand Index of Cluster Similarity. Educational and Psychological Measurement, Vol. 46, 655-661.

Everitt, B., 1981: Cluster analysis. 2nd edition. London-New York.

Ferligoj, A., Batagelj, V., 1982: Clustering with Relational Constraints. Psychometrika, Vol. 47, 413-426.

Ferligoj, A., Batagelj, V., 1983: Some Types of Clustering with Relational Constraints. Psychometrika, Vol. 48, 541-552.

Fraboni, M., Salstone, R., 1992: The WAIS-R Number-of-Factors Quandary: A Cluster Analytic Approach to Construct Validation. Educational and Psychological Measurement

Gordon, A. D., 1999: Classification. 2nd edition. London and others.

Gowda, K. C., Krishna, G., 1978: Agglomerative Clustering Using the Concept of Mutal Nearest Neighborhood. Pattern Recognition, Vol. 10, 105-112.

Green, P. E., Carmone, F. J., Kim, J., 1990: A Preliminary Study of Optimal Variable Weighting in K-Means-Clustering. Journal of Classification, Vol. 2, 271-285.

Hartigan, J. A., 1975: Clustering Algorithms. New York-Chichester-Brisbane-Toronto-Singapore.

Holm, K., 2000: Almo Statistik System. Handbuch. Sozialwissenschaftliche Skalierungsverfahren. Linz.

Holman, E., W., 1972: The Relation between Hierarchical and Euclidean Models for Psychological Distances. Psychmoetrika, Vol. 37, 417-423.

Hubert, L. J., Levin, J. R., 1977: Inference Models for Categorical Clustering. Psychological Bulletin, Vol. 84, 878-887.

Hubert, L., 1974: Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarchical Clustering Procedures. Journal of the American Statistical Association, Vol. 69, 698-704.

Jain, A. K., Dubes, R. C., 1988: Algorithms for Clustering Data. Englewood Cliffs (New Jersey).

Kaufman, L., Rousseeuw, P.J., 1990: Finding Groups in Data. An Introduction to Cluster Analysis. New York-Chichester-Brisbane-Toronto-Singapore.

Klastorin, T. D., 1983: Assessing Cluster Analysis Results. Journal of Marketing Research, Vol. 20, 92-98.

Kruskal, J. B., 1964a: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika, Vol. 29, 1-27.

Kruskal, J. B., 1964b: Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika, Vol. 29, 115-129.

Milligan, G. W., 1980: An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. Psychometrika, Vol. 45, 325-342.

Milligan, G. W., Cooper, M. C., 1987: Methodology Review: Clustering Methods. Applied Psychological Measurement, Vol. 11, 329-354.

Mojena, R., 1977: Hierarchical Grouping Methods and Stopping Rules: An Evaluation. Computer Journal, Vol. 20, 359-363.

Morey, L. C., Agresti, A., 1984: The Measurement of Classification Agreement: An Adjustment to the Rand Statistics for Chance Agreement. Educational and Psychological Measurement, Vol. 44, 33-37.

Opitz, O., Wiedemann, R., 1989: An Agglomerative Algorithm of Overlapping Clustering. In: Opitz, O. (Ed.): Conceptual and Numerical Analysis of Data. Berlin, 201-211.

Pruzansky, S., Tversky, A., Carroll, J.D., 1982: Spatial versus Tree Representations of Proximity Data. Psychometrika, Vol. 47, 3-19.

Punj, G., Stewart, D. W., 1983: Cluster Analysis in Marketing Research: Review and Suggestions for Applications. Journal of Marketing Research, Vol. 20, 134-148.

Rand, W. M., 1971: Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, Vol. 66, 846-850.

Steinhausen, D., Langer, K., 1977: Clusteranalyse. Einführung in Methoden und Verfahren der automatischen Klassifikation. Berlin-New York.

Wishart, D., 1999: ClustanGraphics Primer. Edinburgh.

Wishart, D., 2001: Optimal Tree Partition. Http://www.clustan.com/best_cut.html (22.11.2001)

Appendix: Rand.sps

```
DATA LIST free /a  b  c d.
BEGIN DATA
7 1 0 0
0 0 3 0
0 0 0 3
0 0 0 1
0 0 0 0
END DATA.

compute e=0.

* read matrix data from a regular spss sav file.
MATRIX.
GET M /VARIABLES= a TO e.
PRINT M /TITLE='Original matrix'.

* nr=number of rows, nc=number of columns.
Compute nr=4.
compute nc=4.

loop #i=1 to nr.
+  compute m(#i,nc+1)=0.
+  loop #j=1 to nc.
+     compute m(#i,nc+1)=m(#i,nc+1)+m(#i,#j).
+  end loop.
end loop.

loop #j=1 to nc+1.
+  compute m(nr+1,#j)=0.
+  loop #i=1 to nr.
+     compute m(nr+1,#j)=m(nr+1,#j)+m(#i,#j).
+  end loop.
end loop.


print m/title='matrix with row and column sums'.

compute a=m.
print a.


loop #i=1 to nr+1.
+  loop #j=1 to nc+1.
```

```
+     compute a(#i,#j)=a(#i,#j)*(a(#i,#j)-1)/2.
+  end loop.
end loop.

print a/title='matrix with elements (m(i,j) over 2)'.

compute S1=a(nr+1,nc+1).

compute S2=0.
loop #i=1 to nr.
+  loop #j=1 to nc.
+     compute S2 = S2 +  a(#i,#j).
+  end loop.
end loop.

compute S3=0.
loop #i=1 to nr.
+  compute S3=S3 + a(#i,nc+1).
end loop.

compute S4=0.
loop #j=1 to nc.
+  compute S4=S4 + a(nr+1,#j).
end loop.


compute rand=(S1+2*S2-(S3+S4) ) / S1.

compute adjrand=(S2- S3*S4/S1) / ( (S3+S4) /2 - (S3*S4) / S1).

print S1.
print S2.
print S3.
print S4.
print rand/title 'Rand Index'.
print adjrand/title 'adjusted Rand Index'.

END MATRIX.
```

Chapter 5:
K-Means

Note:
This chapter is based on Bacher (1996: 308-348)

5.1 Basic Idea, Algorithm and Methods

The idea of k-means is very simple:

- **Assign a set of cases into K clusters so that the within cluster sum of squares ('error') is minimized.**

If x_{gj} denotes the value of case g in variable j and $\bar{x}_{kj}$ the mean (centre) of cluster k in variable j, the function to minimize is

$$SQ_{in}(K) = \sum_{k} \sum_{g \in k} \sum_{j} \left(x_{gj} - \bar{x}_{kj}\right)^2 \rightarrow \min .$$

$\sum_{j} \left(x_{gj} - \bar{x}_{kj}\right)^2$ is the squared Euclidean distance $d^2_{g,k}$ of case g to the mean of cluster k.

Therefore, the function can be written as

$$SQ_{in}(K) = \sum_{k} \sum_{g \in k} d^2_{g,k} \rightarrow \min .$$

The cluster centres are computed iteratively. The **algorithm** is:

1. Choose an initial partition of the cases into k clusters.
2. Compute for each case the squared Euclidean distance to cluster centres. Assign each case to its nearest cluster.
3. Re-compute the cluster centres after all cases are checked.
4. Repeat steps 2 and 3 until cluster membership, cluster centres or the error sum of squares do not change.

K-means clustering procedure belongs to a more general group of clustering techniques **labelled as partitioning or optimization methods**. In contrast to hierarchical techniques, the number of clusters must be defined in advance. The idea of optimization methods is shown in the following figure:

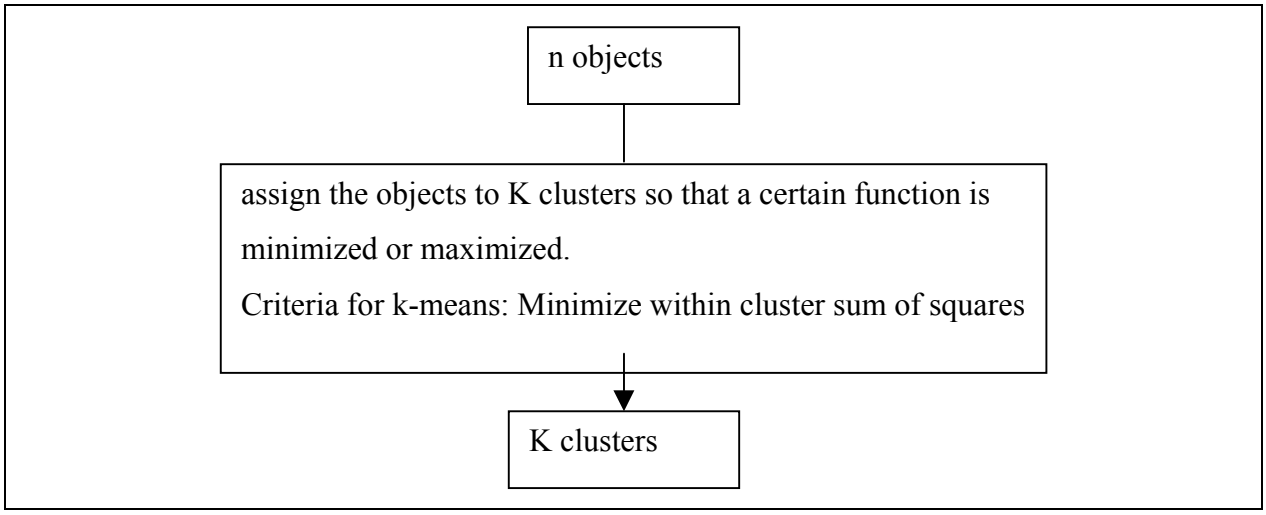
Different methods exist to select an **initial partition**. Some of them are:

1. The cases are randomly assigned to K clusters.
2. The user specifies starting values for the K cluster centres (means).
3. The starting values for the K cluster centres are obtained from a hierarchical procedure. If the sample size is large, a randomly selected subsample is analysed using hierarchical techniques.

The third approach has the advantage that information about the number of clusters is provided, too. The techniques for hierarchical cluster analysis to determine the number of clusters (see chapter 4) can be used in this case (and these techniques can perform better than those for k-means). The second approach is some kind of a rudimentary confirmatory analysis. The starting partition is defined. In contrast to a full confirmatory analysis, no values are fixed or constrained.

SPSS (procedure QUICK CLUSTER) includes none of these methods. They can be implemented easily using syntax commands. SPSS selects starting values with a 'leader' algorithm. The K cases with the largest distances to each other as starting points are selected.

According to simulation studies (e.g. Milligan 1980, Hajanal and Loosveldt 2000) the conclusion can be drawn that randomly generated starting values perform worst among all studied methods. However, as in all simulation studies caution is necessary. As far as I know the QUICK CLUSTER starting procedure has not yet been evaluated.



5.2 A first Application – Reproducing the Typology of the Shell Youth Survey 1997

As a first example we are going to replicate Münchmeier's typology of juveniles (see chapter 1) for our survey of apprentices. The question analysed is '**does this typology fit to our survey on apprentices?**'

The first problem of our analysis is the fact that our survey does not contain all variables Münchmeier used. Only a small overlap exists in the two studies. Therefore, we tried to find similar indicators and added variables. Figure 5-1 shows the variables used in the following analysis.

our survey of apprentices	used by Münchmeier
pessimistic view of the future (PESSI)	yes, we used more general statements about the future
interested in politics (INTER)	no, but the variable was used to describe the cluster
trust in institutions (TRUST)	no, but the variable was used to describe the cluster
political alienation (ALIEN)	yes, similar items were used
normlessness (NORMLESS)	no, but anomie/normlessness was used to describe the clusters
positive attitude toward violence (VIOL)	no, relation with conflictual political action was assumed
sympathizing with Green party (GREEN)	no, but the variable was used to describe the cluster
sympathizing with SPD (SPD)	no, but the variable was used to describe the cluster
sympathizing with CDU (CDU)	no, but the variable was used to describe the cluster
sympathizing with CSU (CSU)	no, but the variable was used to describe the cluster
main political task: safety (SAFE)	no, but the variable was used to describe the cluster

Figure 5-1: Variables used to reproduce the typology of Münchmeier

Factor analysis (principal component analysis; FACTOR procedure) and multiple correspondence analysis (HOMALS) were used to build the variables (see table 5-1). HOMALS was applied to test whether the category 'don't know' can be presented on the same dimension as the other categories. If this was the case, a scale value for the 'don't know' category was estimated for further analysis. Otherwise, the category was handled as missing.

dimension (a)	number of items	eigenvalues	reliability Cronbach's alpha	range of the theoretical scale	empirical mean
PESSI	4	2.06 (0.81) (b)	0.68	4-16	11.5
INTER	1	-	-	1-5	2.7
TRUST	2	0.86 (c)	-	2-14	6.8
ALIEN	3	1.96 (0.57) (b)	0.73	3-12	9.5
NORMLESS	4	2.12 (0.73) (b)	0.70	4-16	9.5
VIOL	1	-	-	1-4	2.2
GREEN	1	-	-	-5-5	-1.3
SPD	1	-	-	-5-5	0.2
CDU	1	-	-	-5-5	0.5
CSU	1	-	-	-5-5	0.5
SAFE	3	1.84 (0.69) (b)	0.66	3-21	19.1

(a) Examples of items:

PESSI	Item: The future of our country is uncertain. Response categories: 4=agree totally to 1=disagree totally
INTER	How strong is your interest in politics? Response categories: 5=not at all to 1=very strong
TRUST	How strongly do you trust or distrust in the following institutions? ...government... Response categories: 1=absolutly no trust to 7=very strong trust, 8=can't say, 9=don't know the institutions
ALIEN	Politicians do not care much about what people like me think. Response categories 4=strongly disagree to 1=strongly agree
NORMLESS	Item: It is not important how one wins, it is important to win. Response categories: 4=agree totally to 1=disagree totally
VIOL	Sometimes it is necessary to use violence in order to reach one's aims. Response categories: 4=agree totally to 1=disagree totally, missing=don`t know
GREEN	What do you think about the political parties in Germany? On the GREENs, I have a +5 very positive to –5 a very negative opinion.
SPD	On the SPD, I have a +5 very positive to –5 a very negative opinion.
CDU	On the CDU, I have a +5 very positive to –5 a very negative opinion.
CSU	On the CSU, I have a +5 very positive to –5 a very negative opinion.
SAFE	How important should the following goals be to German politicians? ...to guarantee pension in the future ... Response categories: 1=not important at all to 7=very important

(b) using PCA

(c) results from HOMALS. Only one dimension was extracted.

Table 5-1: Characteristics of the scales used

The variables have different scales (see table 5-1). They are **incommensurable** (see chapter 2.1). Different methods exist to solve this problem. We used standardization of variables.

Technically speaking, we applied DESCRIPTIVE and used the SAVE command:

```
des var=pessi inter trust alien normless viol green spd
     cdu csu safe /save.
```

The SAVE subcommand signals SPSS to compute standardized scores for each variable and to save these scores in new variables. If names are not specified, SPSS uses the prefix 'z'.

Then QUICK CLUSTER was run:

```
quick cluster
   zpessi zinter ztrust zalien znormles zviol zgreen zspd
   zcdu zcsu zsafe
  /missing=pairwise
  /criteria=cluster(4) converge (0.0000) mxiter(200)
  /print initial distance anova.
```

Quick cluster uses the standardized variables ZPESSI, ZINTER and so on. **Pairwise deletion of missing values** was defined in order to include as much cases as possible (see chapter 3.8).

The number of clusters was fixed to four in the CRITERIA command (default: 2), because we supposed that the Münchmeier's KIDS cluster (see chapter 1) - that consist of very young juveniles - is not present in our survey. Figure 5-2 shows the expected characteristics of the clusters. A '+' in a variable means that a high value characterizes the cluster, a '-' stands for a low value, and a '0' for a value about the average. The question mark '?' was used for unclear cases.

	critical, but loyal juveniles	**traditional juveniles**	**conventional juveniles**	**(not yet) integrated juveniles**
PESSI	+	-	-	+
INTER	+	+	-	-
TRUST	0	+	-	-
ALIEN	0	-	+	+
NORMLESS	-	-	0	+
VIOL	0	-	-	+
GREEN	+	-	-	-
SPD	?	+	?	?
CDU	?	+	?	?
CSU	?	+	?	?
SAFE	-	+	+	-

Figure 5-2: Expected cluster centres

The maximum number of iterations (MXITER) was changed, too. SPSS could not find a convergent solution for the default value of ten iterations. If CONVERGE is equal to 0, SPSS continues iteration until cluster centres do not change between two iterations.

Table 5-2 shows the cluster centres.

Final Cluster Centers

	Cluster			
	1	2	3	4
Zscore(PESSI)	,21766	,47277	-,48842	,14215
Zscore(INTER)	,45728	**-,86055**	,11676	,26054
Zscore(TRUST)	-,36848	**-,68924**	**,54906**	-,01174
Zscore(ALIEN)	,31855	**,76615**	**-,56586**	-,04474
Zscore(NORMLESS)	,47941	-,13774	-,32646	,27650
Zscore(VIOL)	**,93171**	-,17434	-,45146	,25334
Zscore(GREEN)	,38134	**-,53822**	**,70122**	-,67138
Zscore(SPD)	,18086	**-,75740**	**,57022**	-,23648
Zscore(CDU)	**-,84280**	**-,92741**	,22732	**,82390**
Zscore(CSU)	**-,73904**	**-,99799**	,19278	**,87236**
Zscore(SAFE)	**-1,52650**	,30414	,27386	,18488

Table 5-2: Cluster centres computed by QUICK CLUSTER

The computed centres differ from our expectations specified in figure 5-2. A traditional cluster of juveniles that has sympathy with all traditional political parties (SPD, CDU, CSU) does not exist. Two traditional clusters emerge instead: one of them (**cluster 3**) sympathizes with the Green party and the SPD. Juveniles of this cluster trust in political institutions, they also do not feel alienated. Their view of the future is optimistic. Political interest is located slightly above the mean. Violence is refused. This cluster may be labelled as **traditional left juveniles**. The other traditional cluster (**cluster 4**) sympathizes with conservative parties. These juveniles look a little more pessimistic into the future. Trust is lower and alienation a little bit higher. This cluster may be named as **traditional conservative juveniles**.

Juveniles of **cluster 1** have a slightly pessimistic view of the future and distrust political institutions. They are interested in politics and prepared to use violence as a way to reach certain aims. The cluster has similarities to **Münchmeier's (not yet) integrated juveniles** as well as to the class of critical, but loyal juveniles. The last cluster (**cluster 2**) shows strong similarities to Münchmeier's **conventional juveniles**. Therefore, this name was adopted for this cluster. Alienated juveniles would be another appropriate name.

The clusters differ in their sizes:

Number of Cases in each Cluster

Cluster	1	87,000
	2	134,000
	3	225,000
	4	174,000
Valid		620,000
Missing		,000

Cluster 3 (traditional left juveniles) is the largest one. Traditional conservative juveniles (Cluster 4) constitute the second largest one. (Not yet) integrated juveniles (cluster 1) are the smallest group. No case was eliminated due to missing cases. In the case of LISTWISE deletion, the number of cases would be reduced from 620 to 436.

We decided to run a second analysis using **five clusters**. Other clusters, like cluster 1 or 3, may contain Münchmeier's cluster of critical, but loyal juveniles that did not exist in the 4-cluster solution. Table 5-3 shows the results.

	Cluster				
	1	**2**	**3**	**4**	**5**
Zscore(PESSI)	,53080	,48571	,17452	-,42875	-,11952
Zscore(INTER)	,17063	**-,77922**	,21403	,16820	,22704
Zscore(TRUST)	-,68860	**-,83380**	-,11868	,54676	,21398
Zscore(ALIEN)	,28576	**,80867**	,30007	-,50350	-,33681
Zscroe(NORMLESS)	,40060	-,15730	,65286	-,50704	,17467
Zscore(VIOL)	,79866	-,25228	1,19604	-,59982	-,00198
Zscore(GREEN)	-,12158	-,69444	,36587	**,84027**	-,59591
Zscore(SPD)	-,14432	-,96063	,29571	**,60830**	-,08296
Zscore(CDU)	-,63657	-,87184	-,37186	,02033	**,95208**
Zscore(CSU)	-,64741	-,87664	-,27690	-,03672	**,95919**
Zscore(SAFE)	-3,17609	,22379	-,12236	,17167	,25048

Table 5-3: Cluster centres of the five cluster solution

The 5-cluster solution contains the two traditional clusters (cluster 4 and 5) and the conventional juveniles (cluster 2). Some characteristics of the (not yet) integrated juveniles exist in cluster 1. Compared to cluster 1 of the 4-cluster solution, some values became more extreme, e.g. the pessimistic view of the future, distrust and - especially - the disagreement to the political aim 'saftey'. The values of other variables decreased slightly.

Most values of the new cluster 3 are near the average, except normlessness and violence. Juveniles of this cluster want to have success, different means – including violence – are allowed. This group of juveniles may be labelled as **success orientated, normless juveniles**.

Table 5-4 shows the relation between the 4-cluster solution and the 5-cluster solution.

Cluster Number of Case * Cluster Number of Case Crosstabulation

			Cluster Number of Case				
			1	2	3	4	Total
Cluster Number of Case	1	Count	28	1		1	30
		% within Cluster Number of Case	32,2%	,7%		,6%	4,8%
	2	Count	4	113		13	130
		% within Cluster Number of Case	4,6%	84,3%		7,5%	21,0%
	3	Count	46	17	20	20	103
		% within Cluster Number of Case	52,9%	12,7%	8,9%	11,5%	16,6%
	4	Count	9	3	171		183
		% within Cluster Number of Case	10,3%	2,2%	76,0%		29,5%
	5	Count			34	140	174
		% within Cluster Number of Case			15,1%	80,5%	28,1%
Total		Count	87	134	225	174	620
		% within Cluster Number of Case	100,0%	100,0%	100,0%	100,0%	100,0%

Table 5-4: Relation between the solution with four clusters and the solution with five clusters.

Summarizing the results, not all clusters of Münchmeier could be reproduced. We were not able to extract a cluster corresponding to critical, but loyal juveniles. Our sample may be one reason for this. Critical, but loyal juveniles have a high educational level. We analysed apprentices. Usually, these juveniles only have a middle educational level.

The other clusters could be reproduced to some extent. In contrast to Münchmeier, two traditional clusters were computed, a traditional left cluster and a traditional right cluster. The further clusters are: conventional juveniles and (not yet) integrated juveniles. Perhaps, five clusters describe the data better than four clusters. The fifth cluster may be labelled as success orientated, normless cluster.

The available results do not answer the question 'do we need four or five clusters?'. Further information is necessary, e.g. measures of stability and statistics for (internal, relative and external) validity, etc. Some of these measures will be discussed in the follwing sections. At first, the computation of different starting partitions will be illustrated.

5.3 Computing Initial Partitions using SPSS

A random starting partition can be generated using the following algorithm:

1. Compute a variable with an uniform random distribution in the interval (0.5 to K+0.5). Round the variable. Label this variable as CLUSTER_.
2. Compute the cluster centres (starting values) using AGGREGATE. Store the centres in a data file.
3. Specify that QUICK CLUSTER reads this input file as starting partition.

Translating these steps into SPSS syntax results in the following commands:

```
compute cluster_=rv.uniform(0.5,4.5).
compute cluster_=rnd(cluster_).
fre var=cluster_.

compute zzpessi=zpessi.
compute zzinter=zinter.
compute zztrust=ztrust.
compute zzalien=zalien.
compute zznorm=znormles.
compute zzviol=zviol.
compute zzgreen=zgreen.
compute zzspd=zspd.
compute zzcdu=zcdu.
compute zzcsu=zcsu.
compute zzsafe=zsafe.

sort cases by cluster_.

aggregate outfile='c:\texte\koeln\spss\rand4.sav'
    /presorted
    /break = cluster_
   /zpessi zinter ztrust zalien znormles zviol
    zgreen zspd zcdu zcsu zsafe=
    mean(zzpessi zzinter zztrust zzalien zznorm zzviol
```

```
         zzgreen zzspd zzcdu zzcsu zzsafe).
execute.

quick cluster
   zpessi zinter ztrust zalien znormles zviol zgreen zspd zcdu zcsu
   zsafe
  /missing=pairwise
  /criteria=cluster(4) mxiter(200)
  /print initial distance anova
  /file = 'c:\texte\koeln\spss\rand4.sav'
  /save cluster(rand4).
```

The first two commands make up step 1 of the algorithm. A random variable is generated, rounded and labelled CLUSTER_. The following FREQ command controls the procedure. (Each category should have approximately the same number of cases.)

The next COMPUTE statements are necessary for the AGGREGATE subcommand. The aggregated variables (= centres) should have identical names like the variables specified for QUICK CLUSTER. The name of the aggregated ZPESSI variable should be ZPESSI, the name of the aggregated ZINTER variable should be ZINTER, and so on. AGGREGATE needs different names for aggregated variables and idividual variables. Therefore, individual variables are doubled and named using the prefix 'ZZ'.

The AGGREGATE procedure computes the starting partition, which is saved in the RAND4.SAV file . Finally, the FILE subcommand specifies that QUICK CLUSTER has to use these starting values. Note: You will get a different starting partition if you re-run the programme, because the random variable generated by SPSS depends on the current time. This effect can be avoided by saving the random variable.

Using a random partition, the patterns of the clusters computed by SPSS starting values are reproduced. Cluster 1 comprises the not yet integrated juveniles, cluster 2 the conventional juveniles, cluster 3 the traditional left juveniles and cluster 4 the traditional right juveniles. However, there are differences in some variables, too. The question 'how much do the two solutions agree?' immediately arises. Measures for comparing classifications and cluster centres of different solutions will be discussed in chapter 5.4 and 5.5.

As already mentioned, random starting values are not the best choice. Better results can be obtained, if a hierarchical method is used for computing the starting configuration:

1. Select a subsample, if the analysed sample is too large.
2. Run CLUSTER and save the membership.
3. Rename the membership variable to CLUSTER_.
4. Compute the cluster centres (starting values) using AGGREGATE. Store the centres in a data file.
5. Reread the original data, transform them (if you have not saved your transformation used to prepare the variables for QUICK CLUSTER) and specify that QUICK CLUSTER reads the saved centres as starting partition.

For our example, the steps result in the following syntax:

```
get file='c:\texte\koeln\spss\km1.sav'.

des var=pessi inter trust alien normless viol
        green spd cdu csu safe /save.

sample 150 from 620.

cluster
   zpessi zinter ztrust zalien znormles zviol zgreen zspd
   zcdu zcsu zsafe
   /measure=seuclid
  /method=ward
  /save =cluster (4)
  /print schedule
  /plot dendrogram.

compute cluster_=clu4_1.
recode cluster_(Sysmis=-99).
freq var=cluster_.

select if (cluster_ > 0).

compute zzpessi=zpessi.
```

```
compute zzinter=zinter.
compute zztrust=ztrust.
compute zzalien=zalien.
compute zznorm=znormles.
compute zzviol=zviol.
compute zzgreen=zgreen.
compute zzspd=zspd.
compute zzcdu=zcdu.
compute zzcsu=zcsu.
compute zzsafe=zsafe.

sort cases by cluster_.

aggregate outfile='c:\texte\koeln\spss\ward4.sav'
    /presorted
    /break = cluster_
    /zpessi zinter ztrust zalien znormles zviol
     zgreen zspd zcdu zcsu zsafe=
     mean(zzpessi zzinter zztrust zzalien zznorm zzviol
     zzgreen zzspd zzcdu zzcsu zzsafe).
execute.

get file='c:\texte\koeln\spss\km1.sav'.

des var=pessi inter trust alien normless viol
        green spd cdu csu safe /save.

quick cluster
   zpessi zinter ztrust zalien znormles zviol
   zgreen zspd zcdu zcsu zsafe
  /missing=pairwise
  /criteria=cluster(4) mxiter(200)
  /print initial distance anova
  /save cluster(ward4)
  /file = 'c:\texte\koeln\spss\ward4.sav'.
```

The results (not reported here) are similar to those computed using randomly selected starting values. The four clusters computed by the SPSS starting procedure are reproduced. Differences in centres occur, too.

Another method to specify starting values is to input values for the centres directly. This method can be applied, if the user has some knowledge about the centres. Such information is available in our example. We know, which cluster has high, low or medium values in the variables. The necessary steps for this approach are:

1. Generate a data file for the centres and enter the theoretically known values.
2. Specify that QUICK CLUSTER reads the starting partition.

Using QUICK CLUSTER and saving the cluster centres is the best way to generate the data file for the centres.

```
quick cluster
   zpessi zinter ztrust zalien znormles zviol
   zgreen zspd zcdu zcsu zsafe
  /missing=pairwise
  /criteria=cluster(4) mxiter(200)
  /print initial distance anova
  /outfile = 'c:\texte\koeln\spss\theory4.sav'.
```

In the next step, the data file can be edited and the values of the centres can be changed. We used the following coding: +1.00 for high (+), 0 for medium (0) and the question mark (?) and –1.00 for low (–).

Finally, the procedure QUICK CLUSTER was run:

```
quick cluster
   zpessi zinter ztrust zalien znormles zviol
   zgreen zspd zcdu zcsu zsafe
  /missing=pairwise
  /criteria=cluster(4) mxiter(200)
  /print initial distance anova
  /save cluster(theory4)
```

```
/file = 'c:\texte\koeln\spss\theory4.sav'.
```

Again, the four clusters of our first analysis – using SPSS starting values – are reproduced. Differences are observable, too.

5.4 Stability Analysis

The idea of stability analysis is to test, whether modifications of methods or data have a negative effect on the results.

Modification of the method

In contrast to hierarchical methods, the technique and the distance measure (squared Euclidean distances) are fixed for k-means. They cannot be modified. The only variable **method factor** is the starting partition. Therefore, the stability of the results after changing the method can be tested by modifying the starting partitions. The following strategies can be applied:

- Generate different random starting partitions, if random starting values are used.
- Re-order the cases, if SPSS starting values are used.
- Change the starting values, if centres are entered or computed using a hierarchical technique.
- Use different starting procedures (randomly generated starting values, SPSS starting values, etc.).

If a classification is stable, the starting procedure should have no influence.

Stability can be tested in two ways:

- by comparing the cluster centres
- by comparing the classifications (the assignment of cases to the cluster)

The second test produce is frequently used.

Modification of the data

A stability test of the data can analyse two factors: cases (population) and variables. **Population stability** can be tested in the following way:

1. Divide the population in M subpopulations.
2. Run for each population k-means and save the cluster centres.
3. Compare the cluster centres.

M is usually set equal to 2. This method only allows to compare cluster centers. If classifications are to be compared, the strategy has to be modified:

1. Divide the population in M (usually M = 2) subpopulations.
2. Use one subpopulation as reference population. Compute the cluster centres for this reference population.
3. Run two cluster analysis for the other subpopulation: an 'ordinary' k-means analysis and a 'confirmatory' k means analysis with fixed centres. Use the centres of the solution of step 2.
4. Compute the Rand index, its modification or another index to compare the classifications within each subpupolation analysed in step 3.

The **stability** of the results after chaning the variables can be tested by adding randomly distributed variables, e.g. standard normal distributed random variables, if standardized variables are analysed. A factorial design that combines all factors may only sometimes be used (see chapter 4).

Instead of testing stability attempts have been made to change the algorithm in order

- to find a more **robust classification,**
- **to find a best solution** or
- to compute a **partition of partitions**.

A more robust classification can be obtained using the following methods:

- Eliminating outliers in a first stage. They may be assigned to clusters in a second stage.
- Using the city block metric instead of squared Euclidean distances, because the city block metric is less sensitive towards outliers.
- Weighting variables automatically according to their contribution to separate the clusters. By this way, variables with a high proportion of random noise should be eliminated.

These techniques are – for example – implemented in FocalPoint (see chapter 5.10 and 5.11). However, experience with these approaches is too low. Finally, the idea of the third approach is to compute something like an 'average' over a certain number of partitions.

5.5 Comparing Classifications

The methods described in chapter 4.5 can be applied. The Rand index, its modification or other measures can be used to compare different solutions and to test – for example – the stability.

5.6 Comparing Cluster Centres

A first approach to the question 'are the centres of two clusters of two different solutions equal?' is to compute a similarity or dissimilarity index. K-means uses (squared) Euclidean distances. This suggests to use the squared Euclidean distance for this purpose, too. The squared Euclidean distance between two centres is:

$$d^2(i \varepsilon C_I, k \varepsilon C_K) = \sum_j \left(\bar{x}_{ij} - \bar{x}_{kj}\right)^2 .$$

The Euclidean distance is $d(i \varepsilon C_I, k \varepsilon C_K) = \sqrt{d^2(i \varepsilon C_I, k \varepsilon C_K)} = \sqrt{\sum_j \left(\bar{x}_{ij} - \bar{x}_{kj}\right)^2}.$

The (squared) Euclidean distance is equal to zero, if the centres are identical. Both measures are difficult to interpret, because the upper limit is unknown and depends on the data used.

Catell's coefficient of profile similarity r_p (Catell 1949) represents one approach to normalize the distance between two centres. Cattell's coefficient is defined as:

$$r_p(i \varepsilon C_I, k \varepsilon C_K) = \frac{\chi^2_{0,5} - \chi^2}{\chi^2_{0,5} - \chi^2} \quad df = p$$

with

$$\chi^2 = \frac{n_i \cdot n_k}{n_i + n_k} \sum_j \frac{\left(\bar{x}_{ij} - \bar{x}_{kj}\right)^2}{s^2_{j/ik}} \ .$$

$\chi^2_{0,5}$ is the expected value for the null hypothesis 'the centres are equal'. $s^2_{j/ik}$ is the pooled within error variance for variable j of cluster i and k, n_i resp. n_k is the number of cases in cluster i resp. in cluster k.

Note: The generalization of Cattel's coefficient - proposed by Huber - was used in the above formula (Lienert and Raatz 1998: 381).

The null hypothesis 'the (squared) Euclidean distance between the centres is equal to zero' can be analysed with the statistic

$$\chi^2 = \frac{n_i \cdot n_k}{n_i + n_k} \sum_j \frac{\left(\bar{x}_{ij} - \bar{x}_{kj}\right)^2}{s^2_{j/ik}}$$

or the statistic T^2:

$$T^2 = \frac{n_i + n_j - p - 1}{(n_i + n_j - 2)p} \cdot \chi^2$$

χ^2 has a chi-square distribution with p degree of freedoms. T^2 has a F distribution with p and $(n_i + n_j - p - 1)$ degrees of freedom (Fahrmeir and Hamerle 1984: 73). Both test statistics

assume that the variables are uncorrelated within the clusters. If this is not the case, a generalization of T^2 can be used (Fahrmeir and Hamerle 1984: 73).

The coefficients and statistics enable you to analyse stability of centres. The procedure is:

1. Recode cluster membership, so that most similar clusters have the same index. After recoding, cluster 1 in solution 1 is most similar to cluster 1 in solution 2 and most similar to cluster 1 in solution 3 etc. Cluster 2 in solution 1 is most similar to cluster 2 in solution 2 and most similar to cluster 2 in solution 3 etc.
2. Compute for each cluster i in solution 1 the statistic χ^2_{i1j} to cluster i in all other solutions j.
3. Sum up these statistic for cluster i. This statistic $\chi^2_{i1} = \sum_j \chi^2_{i1j}$ can be used to test the hypothesis 'cluster i of solution 1 is stable'. The statistic has p(m-1) degrees of freedom. m is the number of the analysed solutions.
4. Sum up the statistics: $\chi^2_1 = \sum_i \chi^2_{i1}$. The statistic can be used to test the hypothesis 'all clusters of solution 1 are stable'. The statistic has kp(m-1) degrees of freedom. p is the number of variables.

The above example assumes – without loss of generality – solution 1 to be the reference solution, whose stability is to be tested.

SPSS offers neither T^2 nor χ^2. r_p is not available, too. Only the (squared) Euclidean distances can be computed using CLUSTER or PROXIMITIES. The procedure is also not implemented in ALMO and CLUSTAN. It is documented here in order to motivate some reader to implement and test this statistic.

5.7 Explained Variance

A good cluster solution should fit well to the data analysed. The explained variance specifies, how good a partition in K clusters explains the variation in the data. The statistic is defined as

$$ETA_K^2 = 1 - \frac{SS_{within}(K)}{SS_{total}} = 1 - \frac{SS_{within}(K)}{SS_{within}(1)}.$$

$SS_{within}(K)$ is the within cluster sum of squares in the case of K clusters. SS_{total} is the total sum of squares. SS_{total} is equal to $SS_{within}(1)$ - the within cluster sum of squares in the case of 1 cluster. The explained variance criteria is well known from the univariate analysis of variance. In contrast to the analysis of variance, k-means minimizes $SS_{within}(K)$ (resp. maximizes the explained variance). Therefore, the statistical test used in the analysis of variance for the null hypothesis 'the explained variance is equal to zero' cannot be applied for k-means.

Unfortunately, QUICK CLUSTER does not compute ETA_K^2. It only computes within and between cluster variances and an F value for each variable. Th F value specifies, which variable separates the cluster best. In the example of chapter 5.2 (4-cluster solution), the sympathy with the CSU separates the cluster best (see table 5-5). This variable has the highest F-value. The sympathy with the CDU is the second best variable, the importance of safety as a political goal is the third best one, followed by the sympathy with the Greens.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(PESSI)	28,379	3	,859	581	33,051	,000
Zscore(INTER)	42,483	3	,795	606	53,461	,000
Zscore(TRUST)	44,937	3	,775	585	58,007	,000
Zscore(ALIEN)	48,119	3	,746	557	64,483	,000
Zscore(NORMLESS)	19,311	3	,907	592	21,286	,000
Zscore(VIOL)	43,203	3	,783	583	55,188	,000
Zscore(GREEN)	74,265	3	,615	571	120,743	,000
Zscore(SPD)	49,881	3	,740	565	67,366	,000
Zscore(CDU)	94,720	3	,507	570	186,923	,000
Zscore(CSU)	99,068	3	,477	563	207,502	,000
Zscore(SAFE)	77,345	3	,615	595	125,750	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 5-5: Statistics for variables produced by SPSS QUICK CLUSTER

The information SPSS computes for the variables can be used to weight the variables in a next analysis (see chapter 6.2) or to compute ETA_K^2 by hand or by a syntax programme. ETA_K^2 is

$$ETA_K^2 = 1 - \frac{SS_{within}(K)}{SS_{total}} = 1 - \frac{\sum_j MSE(j) \cdot (dfe(j)}{\sum_j MSE(j) \cdot (dfe(j)) + \sum_j MSB(j) \cdot (dfb(j))}$$

MSB(j) is the mean between cluster sum of squares for variable j, MSE(j) is the mean error square of sum for variable j, dfb(j) and dfe(j) are the corresponding degrees of freedom.

A syntax programme can be written using the following method:

1. Copy the statistics of variables in the output into the clipboard.
2. Open a new syntax window.
3. Paste the statistics into the new window.
4. Write the syntax commands.
5. Run the programme.

The syntax for our example is:

```
data list free/name (A20) ssb dfb ssw dfw f sig.
begin data
Zscore(PESSI)   28,379     3     ,859  581   33,051     ,000
Zscore(INTER)   42,483     3     ,795  606   53,461     ,000
Zscore(TRUST)   44,937     3     ,775  585   58,007     ,000
Zscore(ALIEN)   48,119     3     ,746  557   64,483     ,000
Zscore(NORMLESS)19,311     3     ,907  592   21,286     ,000
Zscore(VIOL)    43,203     3     ,783  583   55,188     ,000
Zscore(GREEN)   74,265     3     ,615  571   120,743    ,000
Zscore(SPD)     49,881     3     ,740  565   67,366     ,000
Zscore(CDU)     94,720     3     ,507  570   186,923    ,000
Zscore(CSU)     99,068     3     ,477  563   207,502    ,000
Zscore(SAFE)    77,345     3     ,615  595   125,750    ,000
end data.
execute.
compute ssb=ssb*dfb.
compute ssw=ssw*dfw.
compute sst=ssb+ssw.
fre var=ssb ssw sst/stat=sum.
```

SSB is the between cluster sum of squares, SSW the within cluster sum of squares and SST the total sum of squares. These statistics are computed for each variable. The total sum of squares over all variables is computed using the STAT subcommand in FREQ. The results are:

```
Statistics
                SSB         SSW         SST
N     valid     11          11          11
      missing   0           0           0
Total           1865,13     4535,91     6401,04
```

ETA_K^2 can be computed by dividing SSB with SST or by 1-SSW/SST. ETA_K^2 is equal to 0.291 resp. 29.1%.

Alternatively, ALMO can be used. ALMO produces the following results:

```
Cluster-        Streuungsquadratsummen      F-Wert    ETA**2   PRE
zahl         innerhalb          zwischen
-----------------------------------------------------------------------
  4         4535.866          1865.134       84.432     0.291    KW
```

It is difficult to estimate whether a certain value of ETA_K^2 is high or low, even if a value is 70% or higher. Threshold values do not exist. Simulation studies are one possibility to obtain threshold values. In our example we used the following method to generate a null model:

1. A data set of 620 cases with 11 variables was generated. All variables have a standard normal distribution.
2. A 4-cluster solution was computed.
3. The experiment was repeated 10 times.

We got the following values of ETA_K^2 :

0.156, 0.158, 0.155, 0.161, 0.160. 0.156, 0.162, 0.157, 0.158, 0.163.

All values are considerably smaller than the empirical value of 0.291. Therefore, the conclusion can be drawn that ETA_K^2 departs 'significantly' from the case without any class structure.

5.8 Comparing Different Solutions

Two solutions were computed in chapter 5.2: a partition with four clusters and a partition with five clusters. This chapter discusses measures that allow you to decide, which partition fits better to the data.

These measures are:

- proportional reduction of errors
- F-Max statistic
- Beale´s F statistic

The explained variance specifies, to which extent a solution with K clusters improves the solution with one cluster. The **PRE coefficient** generalizes this idea. It compares the K cluster solution with the previous solution using (K-1) clusters. PRE is defined as:

$$PRE_K^2 = 1 - \frac{SS_{within}(K)}{SS_{within}(K-1)}.$$

The PRE coefficient ignores the fact that more clusters automatically result in a better fit, expressed in a higher explained variance. A solution with K2 > K1 clusters will always have a higher explained variance (except a local minimum was found).

The **F-Max** statistic corrects this 'defect'. It is defined as:

$$F - MAX_K = \frac{SS_{between}(K)/K-1}{SS_{within}(K)/n-K} = \frac{\left(SS_{total} - SS_{within}(K)\right)/K-1}{SS_{within}(K)/n-K}$$

F-Max has no F distribution, because $SS_{within}(K)$ is minimized by k-means (see above).

The following test statistic allows a significance test. It analyses the null hypothesis 'a solution with K1 is not improved by a solution with more clusters' (number of K2 clusters > K1 clusters). The statistic was proposed by Beale (Kendall 1980: 39- 40, Gordon 1999: 63, Everitt 1981: 65) and is defined as

$$F-BEALE_{K_2,K_1}=\left(\frac{SS_{within}(K_1)-SS_{within}(K_2)}{SS_{within}(K_2)}\right)\Bigg/\left(\frac{n-K_1}{n-K_2}\cdot\left(\frac{K_2}{K_1}\right)^{2/p}-1\right).$$

$F-BEALE_{K_2,K_1}$ has an F distribution with $p\cdot(K_2-K_1)$ and $p\cdot(n-K_2)$ degrees of freedom, if the following conditions hold: (a) the variables are independent, (b) the variables have equal scale units, and (c) the clusters are spherical. Beale´s F statistic is a conservative test. It provides only convincing results, if the clusters are well separated.

None of the depicted statistic methods is available in SPSS. Bacher (2001) describes a syntax programme computing ETA, PRE and F-Max. In principle, it is possible to write a syntax programme for the computation of F-Beale, too.

All test statistics are computed in ALMO. The comparison of the solutions with 4 clusters and 5 clusters produces the following results:

```
Cluster-       Streuungsquadratsummen     F-Wert    ETA**2   PRE
zahl        innerhalb         zwischen
-------------------------------------------------------------------
  4        4535.866         1865.134      84.432     0.291    KW
  5        4344.885         2056.115      72.759     0.321   0.042


Bealsche F-Werte:
(Spalte1..4-Clusterloesung, Spalte2..5-Clusterloesung usw.;
 unteres Dreieck = F-Werte;
 oberes Dreieck  = Signifikanzen der F-Werte)

Spalte 1 Spalte 2
   0      57.4424
  1.0199    0
```

The results make us conclude that the solution using 5 clusters does not produce better results than the solution using 4 clusters: F-Max is intended for a maximum of 4 clusters and Beale´s F value is not significant.

5.9 Tests for the number of clusters

The results of the previous chapter do not allow the conclusion that 4 clusters fit best to the data. A solution using 6 clusters can have a significantly better fit to the data, and a solution using 3 or 2 clusters can fit as well as the solution using 4 clusters. In order to test these possibilities, it is useful to calculate the statistics of the previous chapter for different Ks, starting with K=1. K=1 allows to check, whether or not a cluster structure exists.

A systematic variation of K, starting with K=1, can also be used to determine the number of clusters, if K is unknown. K-means is run L times, starting with k=1, k=2, ..., k=K, ..., k=L.

The **'best' solution(s)** can be determined using following rules:

1. **Scree test:** Select the solution with an elbow knick.
2. **(Sharp) decline of ETA or PRE:** Select the solution whose successors have remarkably lower ETA (or PRE) values.
3. **F-Max:** Select the solution with the highest F-Max value.
4. **Statistical significance of Beale´s F:** Select the solution that improves previous solutions significantly and is not improved by succeeding solutions.

Applying these rules to our example results in the following:

- Scree test: It is difficult to identify an elbow. Several weak elbow knicks exist, e.g. for K=2 and K=9. (see figure 5-3)
- Decline of ETA or PRE: It is difficult to identify a sharp decline. Declines of PRE can be observed for K=2, K=4 and K=9. (see table 5-6)
- F-Max: Its maximum value occurs for K=2. In this case F-Max allows no decision. (see table 5-6)
- Statistical significance of Beale's F: The solution using 9 clusters fulfils the criteria described above: The 9-cluster solution produces significantly better results than all previous solutions. The next succeeding solution producing significant results uses 18 clusters, if a significance level of at least 90% is used. Better results are calculated by a

solution using 24 clusters. The solution using 35 clusters produces better results than all previous solutions provided that there are 35 clusters or more. (see table 5-7)

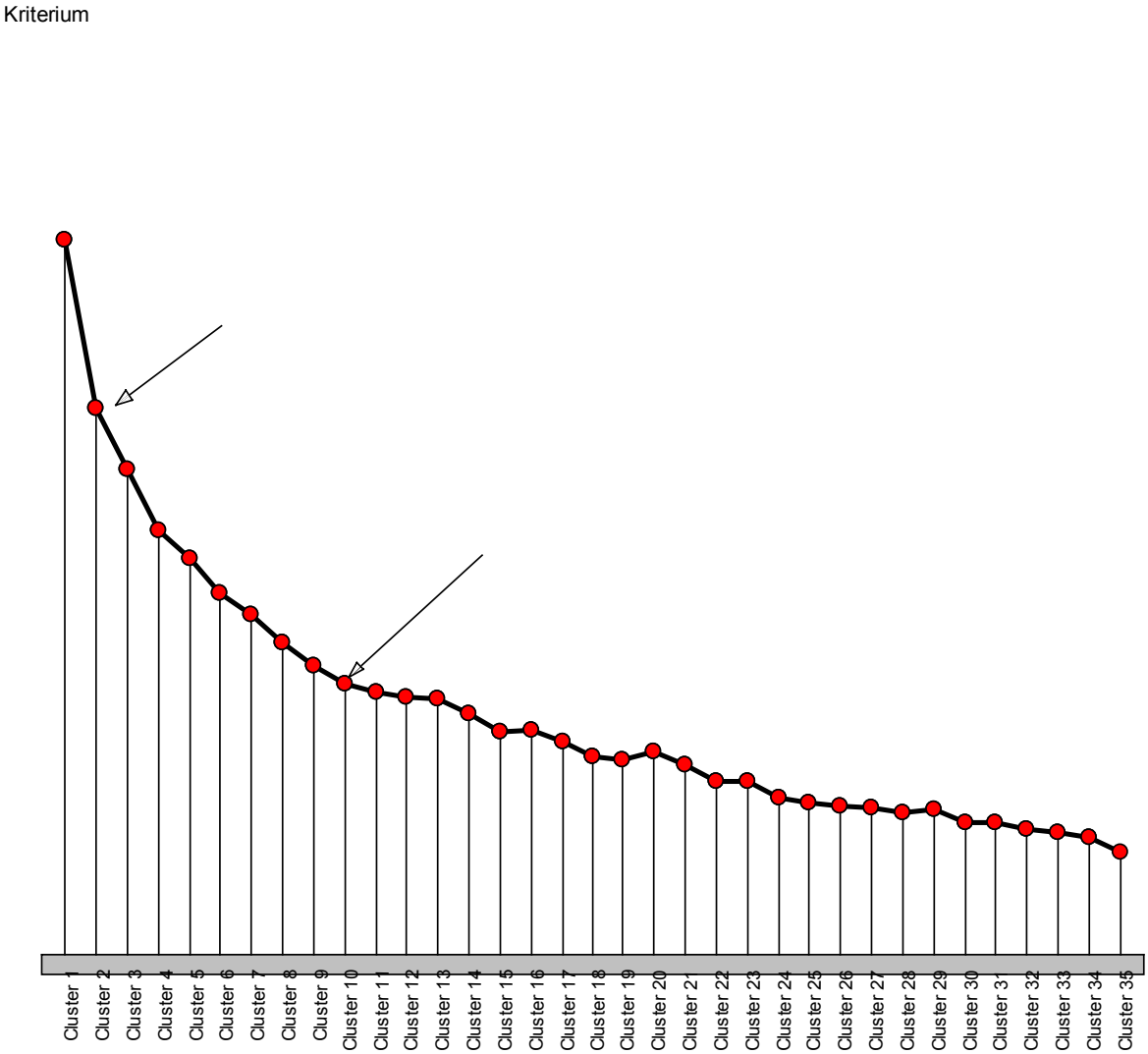


Figure 5-3: Scree diagram for ETA**2 (computed by ALMO)

Cluster-zahl	Streuungsquadratsummen innerhalb	Streuungsquadratsummen zwischen	F-Wert	ETA**2	PRE
1	6401.000	0.000	0.000	0.000	KW
2	5329.283	1071.717	124.280	0.167	0.167**
3	4928.960	1472.040	92.134	0.230	0.075
4	4535.866	1865.134	84.432	0.291	0.080**
5	4344.885	2056.115	72.759	0.321	0.042
6	4107.631	2293.369	68.562	0.358	0.055
7	3979.032	2421.967	62.187	0.378	0.031
8	3784.086	2616.914	60.462	0.409	0.049
9	3618.672	2782.328	58.723	0.435	0.044**
10	3515.423	2885.577	55.634	0.451	0.029
11	3461.111	2939.889	51.729	0.459	0.015
12	3427.248	2973.752	47.959	0.465	0.010
13	3413.216	2987.783	44.278	0.467	0.004
14	3327.060	3073.939	43.069	0.480	0.025
15	3204.881	3196.119	43.096	0.499	0.037
16	3211.295	3189.705	39.996	0.498	-0.002
17	3131.454	3269.546	39.349	0.511	0.025
18	3042.018	3358.982	39.101	0.525	0.029
19	3019.259	3381.741	37.397	0.528	0.007
20	3072.155	3328.845	34.217	0.520	-0.018
21	2987.719	3413.281	34.216	0.533	0.027
22	2885.653	3515.347	34.690	0.549	0.034
23	2876.031	3524.969	33.259	0.551	0.003
24	2771.252	3629.747	33.941	0.567	0.036
25	2748.007	3652.993	32.956	0.571	0.008
26	2726.277	3674.723	32.026	0.574	0.008
27	2703.158	3697.842	31.200	0.578	0.008
28	2667.905	3733.095	30.680	0.583	0.013
29	2692.754	3708.246	29.067	0.579	-0.009
30	2612.716	3788.284	29.499	0.592	0.030
31	2612.585	3788.415	28.470	0.592	0.000
32	2573.420	3827.580	28.212	0.598	0.015
33	2546.247	3854.753	27.771	0.602	0.011
34	2516.650	3884.349	27.408	0.607	0.012
35	2426.216	3974.784	28.188	0.621	0.036

Table 5-6: Test statistics (computed by ALMO)

Bealsche F-Werte:

(Spalte1..1-Clusterloesung, Spalte2..2-Clusterloesung usw.;
unteres Dreieck = F-Werte;
oberes Dreieck = Signifikanzen der F-Werte)

Spalte 1	Spalte 2	Spalte 3	Spalte 4	Spalte 5	Spalte 6	Spalte 7
0	86.7905	86.0168	93.7764	94.2240	97.4969	97.9340
1.4771	0	59.0702	82.0187	81.1315	91.0661	91.7038
1.3270	1.0379	0	89.8127	86.3376	94.7415	94.7944
1.4037	1.2676	1.5641	0	57.4424	84.4903	84.0545
1.3573	1.2113	1.3323	1.0199	0	91.7455	89.0352
1.4085	1.2985	1.4297	1.3030	1.6322	0	62.1914
1.3884	1.2755	1.3707	1.2437	1.3815	1.0737	0
1.4523	1.3644	1.4759	1.4018	1.5674	1.4963	1.9626
1.5059	1.4343	1.5535	1.5056	1.6713	1.6520	1.9845
1.5135	1.4432	1.5525	1.5048	1.6430	1.6117	1.8254
1.4853	1.4074	1.4980	1.4393	1.5428	1.4863	1.6128
1.4472	1.3606	1.4333	1.3639	1.4392	1.3643	1.4371
1.3990	1.3028	1.3579	1.2780	1.3288	1.2389	1.2733
1.4204	1.3294	1.3865	1.3139	1.3671	1.2894	1.3293
1.4795	1.4001	1.4677	1.4089	1.4743	1.4165	1.4745
1.4259	1.3370	1.3895	1.3211	1.3682	1.2982	1.3318
1.4536	1.3698	1.4257	1.3633	1.4139	1.3523	1.3912
1.4940	1.4168	1.4782	1.4230	1.4794	1.4273	1.4735
1.4753	1.3952	1.4506	1.3929	1.4423	1.3873	1.4254
1.3916	1.2992	1.3375	1.2682	1.2994	1.2312	1.2480
1.4325	1.3464	1.3905	1.3278	1.3649	1.3049	1.3291
1.4933	1.4156	1.4684	1.4141	1.4599	1.4099	1.4442
1.4703	1.3895	1.4370	1.3801	1.4204	1.3677	1.3960
1.5392	1.4672	1.5240	1.4754	1.5247	1.4814	1.5197
1.5317	1.4585	1.5125	1.4631	1.5094	1.4654	1.5005
1.5241	1.4498	1.5012	1.4510	1.4946	1.4498	1.4821
1.5190	1.4439	1.4933	1.4426	1.4841	1.4390	1.4690
1.5269	1.4525	1.5017	1.4519	1.4931	1.4491	1.4789
1.4777	1.3979	1.4392	1.3849	1.4186	1.3700	1.3922
1.5308	1.4565	1.5039	1.4546	1.4939	1.4506	1.4785
1.5071	1.4303	1.4736	1.4223	1.4576	1.4124	1.4365
1.5226	1.4472	1.4915	1.4416	1.4778	1.4341	1.4591
1.5271	1.4521	1.4960	1.4465	1.4823	1.4392	1.4639
1.5349	1.4604	1.5044	1.4556	1.4914	1.4490	1.4737
1.6067	1.5387	1.5904	1.5470	1.5895	1.5527	1.5839

Table 5-8: Beale's F-values

Spalte 8	Spalte 9	.....	Spalte17	Spalte18	Spalte19	....
99.3668	99.8093		99.9795	99.9935	99.9930	
97.3196	99.2023		99.8437	99.9558	99.9470	
98.7204	99.6933		99.9319	99.9826	99.9782	
95.9312	99.0557		99.6914	99.9282	99.9027	
97.9696	99.6180		99.8355	99.9653	99.9507	
93.6773	98.9086		99.3389	99.8707	99.8042	
97.2313	99.5749		99.5239	99.9166	99.8662	
0	97.2097		96.3196	99.3273	98.8909	
1.9603	0		81.0760	95.3810	92.8682	
1.7051	1.3978		69.3029	91.9469	87.7336	
1.4394	1.1222		77.9795	95.7141	92.6262	
1.2456	0.9483		91.5195	98.9871	97.8340	
1.0728	0.7900		99.1227	99.9264	99.7978	
1.1635	0.9466		97.1618	99.8044	99.3876	
1.3482	1.1932		38.7294	92.6665	83.7354	
1.1944	1.0301		97.5504	99.8898	99.3646	
1.2708	1.1317		0	99.4526	95.4254	
1.3696	1.2531		2.4247	0	21.4213	
1.3212	1.2063		1.5597	0.6521	0	
1.1309	1.0025		0.5493	KW	KW	
1.2245	1.1119		1.0434	0.5413	0.4850	
1.3533	1.2573		1.5008	1.2288	1.4394	
1.3040	1.2076		1.3232	1.0622	1.1777	
1.4396	1.3572		1.6834	1.5194	1.7163	
1.4209	1.3396		1.6028	1.4451	1.5982	
1.4031	1.3228		1.5372	1.3863	1.5102	
1.3909	1.3119		1.4934	1.3502	1.4552	
1.4029	1.3265		1.5068	1.3754	1.4738	
1.3127	1.2328		1.3101	1.1698	1.2349	
1.4051	1.3322		1.4903	1.3733	1.4562	
1.3621	1.2883		1.3993	1.2816	1.3493	
1.3870	1.3159		1.4409	1.3319	1.4003	
1.3930	1.3234		1.4462	1.3424	1.4078	
1.4042	1.3363		1.4609	1.3623	1.4259	
1.5202	1.4585		1.6573	1.5734	1.6511	

Table 5-7: Beale's F-values (continued)

Spalte23	Spalte24	Spalte25	....	Spalte35
99.9975	99.9996	99.9996		100.0000
99.9781	99.9973	99.9974		100.0000
99.9908	99.9990	99.9990		100.0000
99.9556	99.9961	99.9961		99.9999
99.9769	99.9981	99.9981		100.0000
99.9029	99.9938	99.9935		99.9999
99.9313	99.9961	99.9958		100.0000
99.3977	99.9725	99.9697		99.9998
95.7764	99.7970	99.7711		99.9987
92.4989	99.6438	99.5914		99.9981
95.5995	99.8520	99.8251		99.9992
98.6682	99.9684	99.9614		99.9998
99.8447	99.9966	99.9957		100.0000
99.5190	99.9921	99.9897		99.9999
90.0581	99.8276	99.7693		99.9993
99.2936	99.9927	99.9895		99.9999
95.8704	99.9678	99.9510		99.9998
64.8639	99.5404	99.3043		99.9986
80.3896	99.8965	99.8191		99.9995
99.9816	99.9998	99.9997		100.0000
99.4300	99.9987	99.9967		100.0000
2.3870	99.8178	99.4642		99.9990
0	99.9957	99.9743		99.9998
3.9971	0	48.4974		99.9804
2.4929	0.9252	0		99.9895
1.9826	0.9127	0.9004		99.9949
1.7508	0.9394	0.9475		99.9972
1.7272	1.0951	1.1551		99.9956
1.2690	0.6661	0.5981		99.9999
1.6269	1.1670	1.2194		99.9965
1.4383	1.0108	1.0269		99.9998
1.5049	1.1300	1.1625		99.9995
1.5055	1.1649	1.1986		99.9997
1.5223	1.2108	1.2467		99.9998
1.8272	1.5603	1.6322		0

Table 5-7: Beale's F-values (continued)

Summarizing the results, 9 clusters and possible 2 clusters are more acceptable than 4 clusters from a statistical point of view. Both solutions can be interpreted substantively. The results of the **two cluster solution** show the separation into the two groups of alienated and integrated juveniles. The results of the **nine cluster solution** suggest the following categories:

- **politically non interested juveniles (n=58).** These juveniles are not interested in politics and distrust political institutions and parties.
- **non conventional activists (n=32).** These juveniles have a high political interest. They accept violence and distrust political institutions and parties.
- **integrated juveniles (n=89).** These juveniles are politically interested, trust political institutions and parties. Weak preferences for the Greens and the SPD exist.
- **'outliers' (n=19):** These juveniles refuse safety as a political goal.
- **conventional juveniles (n=123):** These juveniles are not interested in politics. They neither sympathize with nor refuse political parties and institutions.
- **pessimists (n=76):** These juveniles feel alienated, distrust political institutions and have a pessimistic view of the future.
- **individualists (n=82):** These juveniles correspond to the average with two exceptions: They want to be successful and do not care about the means to reach this goal.
- **traditional 'left' (n=46):** These juveniles sympathize with SPD and Greens.
- **traditional 'conservative' (n=95):** These juveniles sympathize with CDU and CSU.

Even if the solutions using 2 and 9 clusters can be substantively interpreted and produce formally better results than the solution using 4 clusters, the 4-cluster solution can be selected for further analysis for the following reasons:

- The fit to data of the two cluster solution (ETA**2 = 17 %) can be interpreted as too poor.
- Nine clusters might be regarded as too much, because the small cluster sizes complicate further analysis.
- The 4-cluster solution can meet other formal criteria better, like stability for example
- The 4-cluster solution can have an higher validity.

5.10 Alternative Software

QUICK CLUSTER has been improved during the last years. In the middle of the 80s SPSS-X, for example, only provided the starting procedure. SPSS Windows 10.8 allows different treatments of missing values, different strategies for updating cluster centres and test statistics for variables. Nonetheless, the situation is unsatisfactory: Test statistics for the clusters are not available. Variables cannot be weighted implicitly. The influence of the order of the cases on the computed configuration is not tested, etc.

ALMO and CLUSTAN eliminate some of these disadvantages (see figure 5-4).

	ALMO	**CLUSTAN (K-Means and FocalPoint)**
different starting procedures	yes, four procedures	yes, six procedures
different starting procedures available in one analysis	no	yes, but only in FocalPoint
different distance measures	no	yes, four different distance measures
definition and elimination of outliers	no, but outliers are reported	yes
definition and elimination of intermediates	no, but intermediates are reported	yes
transformation of variables	yes, standardization	yes, standardization and transformation to [0,1]
explicit weighting of variables	yes	yes
implicit or automatic weighting of variables	yes	yes, but only in FocalPoint
different optimization criteria	yes, generalized variance criteria	yes, exact re-assignment test
two stage procedure	not implemented	yes (FocalPoint)
computation of exemplars	yes	yes, may be used in iteration in FocalPoint
statistics for describing clusters	yes, ETA^2	no
statistics for comparing clusters	yes, ETA^2, F-Max and Beale´s F value	no
statistics for comparing classifications	not yet integrated	no
computation of statistics for variables	yes	yes, but only in FocalPoint
test of reproducibility	no	yes, but only in FocalPoint

Figure 5-4: Comparison of ALMO and CLUSTAN

CLUSTAN provides two procedures: K-means and a modification of k-means, called FocalPoint (Wishart 2000). Both methods allow six different starting procedures:

1. Random clusters: The cases are randomly assigned to K clusters (see chapter 5.3) .
2. Seed points: Starting values are specified by the user (see chapter 5.3).
3. Tree partition: Starting values are computed using a hierarchical method (see chapter 5.3).
4. Dense cliques: The largest K cliques are selected as starting values. 'Cliques are tight, densely populated clusters of cases, which are all very similar within each cluster.' (Wishart 2000: 21)
5. Cluster exemplars: The most typical case of each cluster – obtained from a tree – will be used.
6. Contingency tables: The user specifies an initial classification.

Applying k-means, the user of CLUSTAN can define only one of the six starting procedures in one analysis. The procedures can be combined in FocalPoint (for more details see next chapter). In addition, for each starting procedure a certain number of calculations using random numbers can be specified in FocalPoint. This enables you to test reproducibility.

K-means in CLUSTAN allows four different distance measures:

1. Euclidean Sum of Squares
2. Euclidean distances
3. City Block
4. Pearson distance

K-means and FocaPoint use an exact re-allocation criteria (called **'exact assignment test'**, Wishart 2001). A case i from cluster p is assigned to a new cluster q, if:

$$ESS_p + ESS_q > ESS_{p-i} + ESS_{q+i}.$$

ESS_p is the error sum of squares of cluster p, ESS_q the error sum of squares of cluster q.

ESS_{p-i} is the remaining error sum of squares, if case i is moved towards cluster q.

ESS_{q+i} is the error sum of squares, if case i is added to cluster q.

Variables and cases can be weighted. In addition, it is possible to eliminate outliers and intermediates. Outliers are cases with a distance to the nearest cluster larger than a treshold defined by the user. (In FocalPoint a certain percentage can be defined, too). Intermediates are cases that are close to two clusters. If the relation

$100 \cdot d_{ik} / d_{ik*} > P$ holds,

a case is defined as intermediate. d_{ik} is the distance of case i to the closest centre, d_{ik} * is the distance of case i to the second closest centre. If the relation is larger than a certain percentage, e.g. 90%, case i is defined as an intermediate. Outliers and intermediates can be eliminated in k-means.

Exemplars are computed to describe the clusters. The exemplar of a cluster is the most typical case of a cluster. In K-Means and FocalPoint exemplars are the cases with the smallest distance to their centres.

Appendix A5-1 reports the results of CLUSTAN applying k-means. The specifications were: Compute a random starting partition and use Euclidean Sum of Squares. Outliers and intermediates were not defined. All cases were included in the analysis.

ALMO provides fewer distance measures than CLUSTAN. It does not allow the elimination of outliers or intermediates during one analysis. An exact relocation test is not implemented. The advantages of ALMO are:

- More test statistics are integrated.
- It is possible to compute different solutions and to compare them.
- SPSS results can be reproduced and test statistics can be added.

The output of ALMO has already been used in the previous sections.

In order to compare CLUSTAN, ALMO and SPSS we run ALMO and SPSS using the results of CLUSTAN. The results are reported in appendix A5-2 and A5-3. The results of the programmes differ in some aspects.

Identical results were obtained for:

- Cluster centres.
- The distances between clusters. CLUSTAN returns mean squared Euclidean distances. They must be multiplied by the number of variables to obtain the ALMO distances.
- The nearest distance of a case to a cluster (You must use the specification OPTION9=3;). The distances must be transformed in the following way:

$$d_{case}(SPSS) = n_{valid\_variables\_in\_case} \cdot d_{case}(CLUSTAN)$$

$$d_{case}(ALNO) = (n_{valid\_variables\_in\_case} \cdot d_{case}(CLUSTAN))^2 \cdot \frac{n_{variables}}{n_{valid\_variables\_in\_case}}$$

Note: SPSS ignores missing values. Ceteris paribus, a case with more missing values always has a smaller distance. This may result in wrong conclusions.

5.11 New Developments

Different modifications of the K-mean method have been proposed. Only some of them can be mentioned.

- **Modification of the assignment rule of cases:** K-means assigns cases deterministically to one cluster. Fuzzy clustering techniques and probabilistic clustering techniques weaken this assumption and allow a case to belong to more than one cluster. Probabilistic clustering techniques will be described in chapter 7. Fuzzy techniques are described in Gordon (1999: 111-114), Jain and Dubes (1988: 132-133) or Kaufman and Rousseeuw (1990: 171). The motivation: Data are usually fuzzy. Therefore, a fuzzy clustering model or a probabilistic clustering technique is more realistic and produces 'better' results.
- **Computation of a secondary partition or consensus partition from a set of partitions:** Different specifications can produce different results. (Example: The results returned by CLUSTAN using the 4-cluster solution with a random starting partition differs from the results returned by SPSS using the 4-cluster solution). Consensus techniques can be used

to get a less sensitive partition. They compute some kind of 'average' or 'median' over a set of first partitions. Consensus methods are discussed in Gordon (1999: 142-146)

- **Measures for influencing the classification of cases and variables:** Different statistics and strategies to identify cases or variables, which influence results, have been proposed. An example describing such a measure can be found in Cheng and Milligan (1996).
- **Two stage procedures:** These techniques have been developed for data mining. They are designed for large data sets. Variables, cases and the number of clusters are to be selected automatically by the computer programme. SPSS provides a two stage procedure in CLEMENTINE (SPSS Inc. 2000) for data mining. In the first step cases are reduced to a manageable number by k-means. The clusters should be compact. Therefore, the number of clusters is higher than in usual k-means analysis. The number may be 200, 300 or higher. A new data matrix is computed containing only the centres. This data matrix is used in the second step (stage). Cases are clustered by hierarchical cluster analysis .

FocalPoint (Wishart 2000) also is a two stage procedure. It allows the elimination of outliers and intermediates in the first stage. In the second stage the eliminated cases are included. The variables may be re-weighted according to the performance in the first step. However, there is one serious disadvantage in FocalPoint: missing variables cannot be handled. Appendix A5-4 shows the output of FocalPoint. Cases with missing values were eliminated for analysis. This reduces the number of cases to 436. The cluster levels were set to 2 - 15. For each level 10 partitions were computed using random starting values. The percentage of outliers was set to 0 %. All intermediates were included. The best solution was determined using 15 clusters. The criterion (Error Sum of Squares) is very low.

References

Bacher, J., 1996: Clusteranalyse [Cluster analysis]. Opladen. [available only in German]

Bacher, J., 2001: Teststatistiken zur Bestimmung der Clusterzahl für QUICK CLUSTER. ZA-Information, 48, 71-97 [available only in German].

Cattell, R. B., 1949: r_p and other Coefficients of Pattern Similarity. Psychometrika, Vol. 14, 279-298.

Cheng, R., Milligan, G.W., 1996: Measuring the influence of individual data points in a cluster analysis. Journal of classification, Vol. 13, 315-335.

Everitt, B., 1981: Cluster analysis. Second edition. New York.

Fahrmeir, L., Hamerle, A., 1984: Mehrdimensionale Zufallsvariablen und Verteilungen. In: L. Fahrmeir and A. Hamerle (Hg.): Multivariate statistische Verfahren. Berlin-New York.

Gordon, A. D., 1999: Classification. 2nd edition. London-New York.

Hajanal, I., Loosveldt, G., 2000: The Effects of Initial Values and the Covariance Structure on the Recovery of some Clustering Methods. In: H.A.L Kiers et al. (Eds.): Data Analysis, Classification, and Related Methods. Berlin-Heidelberg and others, 53-58.

Jain, A. K., Dubes, R. C., 1988: Algorithms for Clustering Data. Englewood Cliffs (New Jersey).

Kaufman, L., Rousseeuw, P. J., 1990: Finding Groups in Data. An Introduction to Cluster Analysis. New York-Chichester-Brisbane-Toronto-Singapore.

Kendall, M., 1980: Multivariate Analysis. 2nd Edition. London.

Lienert, G. A., Raatz, U., 1998: Testaufbau und Testanalyse. Weinheim [available only in German]

Milligan, G. W., 1980: An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. Psychometrika, Vol. 45, 325-342.

SPSS Inc., 2000: The SPSS TwoStep cluster component. White paper – technical report.

Wishart, D., 2000: FocalPoint Clustering. Edinburgh.

Appendix A5-1: CLUSTAN Output

```
k-Means Cluster Results


Cluster Model Summary

Cluster   Size     Distance      ESS
  1        220      0,188       129,143
  2        175      0,216       107,918
  3        146      0,408       105,947
  4        79       0,547       89,060

Outliers   0

Model      620     0,000  432,067


Cluster Centres

Cluster

1          -0,388   0,107 0,538 -0,512      -0,329      -0,400      0,807
           0,626
           0,073    0,045 0,187
2          -0,086   0,313 0,179 -0,277      0,196 0,168 -0,627      -0,118
           0,905    0,907 0,225
3          0,499    -0,827      -0,731      0,799 -0,083      -0,149      -
0,583      -0,803
           -0,790   -0,804      0,318
4          0,328    0,512 -0,586      0,449 0,614 1,013 0,184 -0,028
           -0,799   -0,701      -1,591
Model      0,000    0,000 0,000 0,000 0,000 0,000 0,000 0,000
           0,000    0,000 0,000


Exemplars for 4 clusters

Cluster   Distance      Exemplar
   1      0,060         263
   2      0,070         249
   3      0,143         125
   4      0,055         67


Distances between clusters

Cluster   1  2     3     4
1          0,000    0,451 0,961 1,004
2          0,451    0,000 0,917 1,061
3          0,961    0,917 0,000 0,786
4          1,004    1,061 0,786 0,000
```

```
t-tests on variables

Cluster

1         -0,388   0,107 0,538 -0,512      -0,329      -0,400      0,807
          0,626
          0,073    0,045 0,187
2         -0,086   0,313 0,179 -0,277      0,196 0,168 -0,627      -0,118
          0,905    0,907 0,225
3         0,499    -0,827      -0,731      0,799 -0,083      -0,149      -
0,583     -0,803
          -0,790   -0,804      0,318
4         0,328    0,512 -0,586      0,449 0,614 1,013 0,184 -0,028
          -0,799   -0,701      -1,591


k-Means Case Results for Euclidean Sum of Squares


Case      Cluster  Nearest
1         2  0,315
2         2  0,416
3         1  0,451
4         2  0,401
5         1  1,224
6         3  0,430
7         2  0,284
8         1  0,714
9         1  0,307
```

Appendix A5-2: ALMO Output

```
Ergebisse aus ALMO
------------------

Modellspezifikation:
Verfahren                   Minimaldistanzverfahren fuer
                            Varianzkriterium
Startwerte                  ueber Tabelle eingegeben
Cluster                     von 4 bis 4
Gewichtung der Distanzen    keine
KEIN_WERT-Schwelle          0.99
Alpha_Niveau                0.05
Mindestfallzahl               5
------------------------------------------------------------------------

Fuer Analyse ausgewaehlte Variable:

Klassifikationsvariablen:
V12    ZPESSI                       quantitativ
V13    ZINTER                       quantitativ
V14    ZTRUST                       quantitativ
V15    ZALIEN                       quantitativ
V16    ZNORMLES                     quantitativ
V17    ZVIOL                        quantitativ
V18    ZGREEN                       quantitativ
V19    ZSPD                         quantitativ
```

```
V20    ZCDU                              quantitativ
V21    ZCSU                              quantitativ
V22    ZSAFE                             quantitativ

Deskriptionsvariablen:

Klassifikationsvariablen nach Dummy-Aufloesung =   11

Deskriptionsvariablen nach Dummy-Aufloesung    =    0

Gewichte der Variablen in der Analyse:
V12    ZPESSI                             Gewicht=    1.00
V13    ZINTER                             Gewicht=    1.00
V14    ZTRUST                             Gewicht=    1.00
V15    ZALIEN                             Gewicht=    1.00
V16    ZNORMLES                           Gewicht=    1.00
V17    ZVIOL                              Gewicht=    1.00
V18    ZGREEN                             Gewicht=    1.00
V19    ZSPD                               Gewicht=    1.00
V20    ZCDU                               Gewicht=    1.00
V21    ZCSU                               Gewicht=    1.00
V22    ZSAFE                              Gewicht=    1.00

------------------------------------------------------------------------
lineare Restriktionen = nein
------------------------------------------------------------------------
Es wurden 620 Datensaetze eingelesen,
davon werden 620 Datensaetze analysiert.
------------------------------------------------------------------------

Masszahlen fuer Klassifikationsvariablen

Variable      n=        MA       SA   z-Wert   Name
------------------------------------------------------------------
  V12        585      0.00     1.00     0.00   ZPESSI
  V13        610      0.00     1.00     0.00   ZINTER
  V14        589      0.00     1.00     0.00   ZTRUST
  V15        561      0.00     1.00     0.00   ZALIEN
  V16        596      0.00     1.00     0.00   ZNORMLES
  V17        587      0.00     1.00     0.00   ZVIOL
  V18        575      0.00     1.00     0.00   ZGREEN
  V19        569      0.00     1.00     0.00   ZSPD
  V20        574      0.00     1.00     0.00   ZCDU
  V21        567      0.00     1.00     0.00   ZCSU
  V22        599      0.00     1.00     0.00   ZSAFE

Gesamtstreuungsquadratsumme =   6401.000
Gesamtfallzahl              =       6412
Freiheitsgrade              =       6401
========================================================================

H-0-Kriterium (1-Clustermodell) = 6805.13
Fallzahl = 620
```

Ergebnisse der Iteration

Cluster-zahl	Itera-tionen	Kriterium	prozentuelle Verbesserung gegenueber H-0
4	3	4788.404	29.635

Bei Modellen 1 bis 6: Kriterium = Wert des Varianzkriteriums
Bei Modell 7: : Kriterium = Wert der Log-Likelihood-Funktion
Bei Modell 8: : Kriterium = Ueberlapp. + Nichtklass.

Cluster-zahl	Streuungsquadratsummen innerhalb	zwischen	F-Wert	ETA**2	PRE
4	4527.399	1873.601	84.974	0.293	KW

Beachte: Die Interpretation dieser Testgroessen ist fuer
die Modelle 3 bis 6 nur sinnvoll, wenn gleiche Skalen-
einheiten der Variablen vorliegen. Dies gilt auch fuer
die Bealschen F-Werte. Setze OPTION35=1;

==========

Die 4-Clusterloesung wird weiter untersucht
==========

Clustergroessen:

C1	220	(35.484 %)
C2	175	(28.226 %)
C3	146	(23.548 %)
C4	79	(12.742 %)

KW-Faelle (ungewichtet)= 0

==========
Zellenmittelwerte der Klassifikationsvariablen
(Mittelwerte bei quantitativen / ordinalen Variablen)
(Anteilswerte bei nominalen Variablen)

Variable	C1	C2	C3	C4
V12	-0.39	-0.09	0.50	0.33
V13	0.11	0.31	-0.83	0.51
V14	0.54	0.18	-0.73	-0.59
V15	-0.51	-0.28	0.80	0.45
V16	-0.33	0.20	-0.08	0.61
V17	-0.40	0.17	-0.15	1.01
V18	0.81	-0.63	-0.58	0.18
V19	0.63	-0.12	-0.80	-0.03
V20	0.07	0.91	-0.79	-0.80
V21	0.04	0.91	-0.80	-0.70
V22	0.19	0.22	0.32	-1.59

Standardabweichungen:

Variable	C1	C2	C3	C4
V12	0.86	0.96	0.94	1.04
V13	0.85	0.80	0.86	1.14
V14	0.77	0.80	0.90	1.03
V15	0.88	0.90	0.65	0.90
V16	0.92	0.92	1.01	0.95
V17	0.80	0.95	0.95	0.89
V18	0.67	0.65	0.79	1.06
V19	0.55	0.93	0.93	1.08
V20	0.72	0.48	0.87	0.93
V21	0.74	0.46	0.89	0.94
V22	0.62	0.57	0.49	1.62

Besetzungszahlen:

Variable	C1	C2	C3	C4
V12	203	169	137	76
V13	220	173	141	76
V14	210	168	134	77
V15	190	161	136	74
V16	209	169	141	77
V17	206	164	142	75
V18	206	160	136	73
V19	205	159	134	71
V20	206	163	134	71
V21	200	163	133	71
V22	211	170	141	77

Z-Werte:

Variable	C1	C2	C3	C4
V12	-6.40	-1.15	6.18	2.74
V13	1.87	5.12	-11.34	3.88
V14	10.10	2.88	-9.32	-4.98
V15	-8.01	-3.88	14.24	4.27
V16	-5.13	2.76	-0.97	5.66
V17	-7.12	2.26	-1.86	9.82
V18	17.13	-12.18	-8.54	1.48
V19	16.22	-1.59	-9.97	-0.22
V20	1.44	24.04	-10.42	-7.17
V21	0.85	25.04	-10.43	-6.23
V22	4.39	5.11	7.66	-8.55

Signifikanz der z-Werte:

Variable	C1	C2	C3	C4
V12	100.00	75.01	100.00	99.23
V13	93.80	100.00	100.00	99.98
V14	100.00	99.55	100.00	100.00
V15	100.00	99.99	100.00	100.00
V16	100.00	99.35	66.82	100.00
V17	100.00	97.52	93.46	100.00
V18	100.00	100.00	100.00	85.54
V19	100.00	88.61	100.00	17.11
V20	85.00	100.00	100.00	100.00
V21	60.37	100.00	100.00	100.00
V22	100.00	100.00	100.00	100.00

Statistiken der Cluster:

Cluster	n=	Streuung innerhalb	Homogenitaet innerhalb
C1	220	0.593	0.407
C2	175	0.622	0.378
C3	146	0.735	0.265
C4	79	1.151	-0.151

Entfernungen der Clusterzentren bzw. Repraesentanten zueinander:
quadrierte (gewichtete) euklidische Distanzen
bei Rep.verfahren gewaehltes Distanzmass

Spalte 1	Spalte 2	Spalte 3	Spalte 4
0	4.9627	10.5662	11.0432
4.9627	0	10.0861	11.6662
10.5662	10.0861	0	8.6464
11.0432	11.6662	8.6464	0

Paarweise Clusterdifferenzen fuer Cluster=1 (n= 220)

Klassifikationsvariablen:

Variable	C2	C3	C4
V12	<	<	<
V13	=	>	<
V14	>	>	>
V15	=	<	<
V16	<	=	<
V17	<	<	<
V18	>	>	>
V19	>	>	>
V20	<	>	>
V21	<	>	>
V22	=	=	>

Paarweise Clusterdifferenzen fuer Cluster=2 (n= 175)

not reported

```
==========================================================================

Gesamtstatistiken fuer Klassifikationsvariablen:
            F-Wert  Signifikanz   ETA**2
                    (1-p)*100              Name
----------------------------------------------------------
  V12      28.112    100.000      0.127
  V13      58.001    100.000      0.223
  V14      75.520    100.000      0.279
  V15      76.834    100.000      0.293
  V16      21.771    100.000      0.099
  V17      48.842    100.000      0.201
  V18     142.463    100.000      0.428
  V19      79.796    100.000      0.298
  V20     161.992    100.000      0.460
  V21     154.049    100.000      0.451
  V22     119.827    100.000      0.377

Variablengruppen innerhalb der Cluster:

Beziehung der Variablen im Cluster   1
====================================


Verschmelzungsprotokoll:

 Schritt   1  alpha=  0.882  V12=
                           + V17=
 Schritt   2  alpha=  0.696  V20=
                           + V21=
 Schritt   3  alpha=  0.533  V13=
                           + V20=
 Schritt   4  alpha=  0.452  V12=
                           + V16=
 Schritt   5  alpha=  0.181  V14=
                           + V19=
 Schritt   6  alpha=  0.139  V13=
                           + V22=
 Schritt   7  alpha=  0.089  V12=
                           + V15=
 Schritt   8  alpha=  0.001  V14=
                           + V18=

Variablengruppen bei einem Fehlerniveau von alpha=  0.050
(Alpha-Niveau nach Bonferroni-Korrektur = 0.000227273)

Variablengruppe   1
         V12=
         V15=
         V16=
         V17=
   (Mittelwert=   -0.43; Varianz=    0.78)

Variablengruppe   2
         V13=
         V20=
         V21=
         V22=
   (Mittelwert=    0.14; Varianz=    0.51)

Variablengruppe   3
         V14=
```

```
         V18=
         V19=
   (Mittelwert=    0.69; Varianz=    0.47)


========================================================================

gepoolte Korrelationsmatrix W:

not reported

------------------------------------------------------------------------

Analyse der Klassifikationsobjekte:
===================================

durchschnittl. Distanzen fuer H-0-Modell = 10.976

Cluster   C1

stand.Distanz  Faelle (in   %)
      < 0.25        1 (  0.45 %)
0.25  < 0.50       25 ( 11.36 %)
0.50  < 1.00      174 ( 79.09 %)
1.00  < 1.25       17 (  7.73 %)
     >= 1.25        3 (  1.36 %)

etc.

Cluster    Repraesentanten     Ausreisser      Ueberlappungen
-------------------------------------------------------------
  C1          1 (  0.455%)      3 (  1.364%)     75 ( 34.091%)
  C2          0 (  0.000%)      1 (  0.571%)     65 ( 37.143%)
  C3          0 (  0.000%)      2 (  1.370%)     47 ( 32.192%)
  C4          1 (  1.266%)     16 ( 20.253%)     19 ( 24.051%)

Objekt   Charakeristik
-----------------------------------------------------------------
      3    liegt im Ueberlappungsb. von C1  und C3
      4    liegt im Ueberlappungsb. von C2  und C1
      7    liegt im Ueberlappungsb. von C2  und C1
     11    Ausreisser in Cluster        C1
     12    liegt im Ueberlappungsb. von C3  und C1  und C2
     13    liegt im Ueberlappungsb. von C3  und C2  und C4
     25    liegt im Ueberlappungsb. von C1  und C2
     27    liegt im Ueberlappungsb. von C1  und C2
     28    liegt im Ueberlappungsb. von C1  und C2  und C3
     30    liegt im Ueberlappungsb. von C4  und C3
     31    liegt im Ueberlappungsb. von C2  und C1
etc.
```

```
Clusterzugehoerigkeit der Objekte(Datensaetze):
( -1 = wegen Kein_Wert eliminiert)
```

Objekt	Clustererzu-gehoerigkeit	quadrierte Distanz zum Clusterzentrum	standardisierte. mittlere Entfern.
1	2	3.44	0.56
2	2	4.55	0.64
3	1	4.94	0.67
4	2	4.38	0.63
5	1	13.39	1.10
6	3	4.70	0.65
7	2	3.10	0.53
8	1	7.82	0.84
9	1	3.36	0.55
10	1	3.68	0.58

etc.

Appendix A5-3: SPSS Specification and Output

SPSS Specification

```
quick cluster
   zpessi zinter ztrust zalien znormles zviol zgreen zspd zcdu zcsu zsafe
  /missing=pairwise
  /criteria=cluster(4) mxiter(200)
  /method=classify
  /print initial distance anova cluster distance
   /file = 'c:\texte\koeln\spss\clustan.sav'.
```

SPSS Output (German version - Sorry!)

Anfängliche Clusterzentren

	Cluster			
	1	2	3	4
Z-Wert(PESSI)	-,38800	-,08600	,49900	,32800
Z-Wert(INTER)	,10700	,31300	-,82700	,51200
Z-Wert(TRUST)	,53800	,17900	-,73100	-,58600
Z-Wert(ALIEN)	-,51200	-,27700	,79900	,44900
Z-Wert(NORMLESS)	-,32900	,19600	-,08300	,61400
Z-Wert(VIOL)	-,40000	,16800	-,14900	1,01300
Z-Wert(GREEN)	,80700	-,62700	-,58300	,18400
Z-Wert(SPD)	,62600	-,11800	-,80300	-,02800
Z-Wert(CDU)	,07300	,90500	-,79000	-,79900
Z-Wert(CSU)	,04500	,90700	-,80400	-,70100
Z-Wert(SAFE)	,18700	,22500	,31800	-1,59100

¢
Aus Unterbefehl FILE eingeben

Cluster-Zugehörigkeit

Fallnummer	Cluster	Distanz
1	2	1,769
2	2	2,133
3	1	2,120
4	2	2,094
5	1	3,659
6	3	1,728
7	2	1,761
8	1	2,666
9	1	1,834
etc.		

Clusterzentren der endgültigen Lösung

	Cluster			
	1	2	3	4
Z-Wert(PESSI)	-,38795	-,08564	,49870	,32771
Z-Wert(INTER)	,10714	,31313	-,82715	,51167
Z-Wert(TRUST)	,53767	,17949	-,73090	-,58603
Z-Wert(ALIEN)	-,51208	-,27704	,79881	,44947
Z-Wert(NORMLESS)	-,32866	,19621	-,08348	,61430
Z-Wert(VIOL)	-,40015	,16817	-,14867	1,01283
Z-Wert(GREEN)	,80684	-,62694	-,58321	,18379
Z-Wert(SPD)	,62611	-,11788	-,80303	-,02822
Z-Wert(CDU)	,07293	,90542	-,79002	-,79920
Z-Wert(CSU)	,04451	,90654	-,80370	-,70107
Z-Wert(SAFE)	,18681	,22488	,31830	-1,59127

¢

Distanz zwischen Clusterzentren der endgültigen Lösung

Cluster	1	2	3	4
1		2,228	3,251	3,323
2	2,228		3,176	3,416
3	3,251	3,176		2,940
4	3,323	3,416	2,940	

¢

ANOVA

	Cluster		Fehler		F	Sig.
	Mittel der Quadrate	df	Mittel der Quadrate	df		
Z-Wert(PESSI)	24,675	3	,878	581	28,112	,000
Z-Wert(INTER)	45,285	3	,781	606	58,001	,000
Z-Wert(TRUST)	54,716	3	,725	585	75,520	,000
Z-Wert(ALIEN)	54,637	3	,711	557	76,834	,000
Z-Wert(NORMLESS)	19,707	3	,905	592	21,771	,000
Z-Wert(VIOL)	39,233	3	,803	583	48,842	,000
Z-Wert(GREEN)	81,906	3	,575	571	142,463	,000
Z-Wert(SPD)	56,346	3	,706	565	79,796	,000
Z-Wert(CDU)	87,901	3	,543	570	161,992	,000
Z-Wert(CSU)	85,053	3	,552	563	154,049	,000
Z-Wert(SAFE)	75,074	3	,627	595	119,827	,000

¢

Die F-Tests sollten nur für beschreibende Zwecke verwendet werden, da die Cluster so gewählt wurden, daß die Differenzen zwischen Fällen in unterschiedlichen Clustern maximiert werden. Dabei werden die beobachteten Signifikanzniveaus nicht korrigiert und können daher nicht als Tests für die Hypothese der Gleichheit der Clustermittelwerte interpretiert werden.

Anzahl der Fälle in jedem Cluster

Cluster	1	220,000
	2	175,000
	3	146,000
	4	79,000
Gültig		620,000
Fehlend		,000

¢

Appendix A5-4: FocalPoint Output

Computed Solutions.

```
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 8 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 6 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 7 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 7 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 9 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 8 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 7 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 7 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 9 [First Stage]
Criterion = 363,2165 Number of clusters = 2 Total moves = 208 Cases
classified = 436 Iterations = 7 [First Stage]
Criterion = 332,4150 Number of clusters = 3 Total moves = 279 Cases
classified = 436 Iterations = 7 [First Stage]
Criterion = 332,2396 Number of clusters = 3 Total moves = 284 Cases
classified = 436 Iterations = 7 [First Stage]
Criterion = 332,2395 Number of clusters = 3 Total moves = 284 Cases
classified = 436 Iterations = 14 [First Stage]
Criterion = 332,2396 Number of clusters = 3 Total moves = 284 Cases
classified = 436 Iterations = 11 [First Stage]
Criterion = 332,5432 Number of clusters = 3 Total moves = 282 Cases
classified = 436 Iterations = 7 [First Stage]

etc.
```

```
Criterion = 208,4905 Number of clusters = 15 Total moves = 388 Cases
classified = 436 Iterations = 11 [First Stage]
Criterion = 207,8950 Number of clusters = 15 Total moves = 392 Cases
classified = 436 Iterations = 9 [First Stage]
Criterion = 209,3486 Number of clusters = 15 Total moves = 390 Cases
classified = 436 Iterations = 9 [First Stage]
Criterion = 208,9965 Number of clusters = 15 Total moves = 389 Cases
classified = 436 Iterations = 11 [First Stage]
Criterion = 209,7964 Number of clusters = 15 Total moves = 395 Cases
classified = 436 Iterations = 11 [First Stage]
Criterion = 211,7050 Number of clusters = 15 Total moves = 398 Cases
classified = 436 Iterations = 11 [First Stage]
Criterion = 211,6531 Number of clusters = 15 Total moves = 395 Cases
classified = 436 Iterations = 14 [First Stage]
```

Top Solution

```
Criterion = 207,89497375 Sum of distances = 207,89497375
```

New Weights

Variable	First Wt.	Second Wt.
	0,7806	0,0000
	1,0318	0,0000
	0,7503	0,0000
	0,8651	0,0000
	0,7574	0,0000
	0,8716	0,0000
	0,9679	0,0000
	1,0470	0,0000
	1,4586	0,0000
	1,4503	0,0000
	1,0195	0,0000

Means after revising weights

Cluster 1 -1,0702 0,7619 0,7929 -0,3145 0,8607
0,8759 -0,8120 -0,3716 0,8235 0,7574 -0,3905
Cluster 2 0,4678 1,5830 -0,8460 0,5703 0,6600
0,7568 -0,3805 -0,6714 -1,3220 -1,1534 -0,5857
Cluster 3 -0,1975 -0,6511 -0,2215 0,1474 -0,9589
-0,5494 -0,2202 -0,2435 -0,2253 -0,2624 0,2794
Cluster 4 -0,5277 0,6610 0,5971 -0,4218 -0,3193
-0,2411 1,1422 1,0694 -1,3710 -1,5292 -0,0677
Cluster 5 -0,1180 -0,3559 0,5986 -0,0292 0,8141
-0,4494 0,3949 0,7069 0,7015 0,7226 0,2719
Cluster 6 -1,6822 0,1101 0,5132 -1,0704 -0,6504
-0,2387 -0,2095 0,3160 0,5339 0,6161 -0,0055
Cluster 7 0,0707 0,2996 -0,1309 -0,3098 0,0275
1,0550 0,0799 0,5097 0,2755 0,3729 0,0200
Cluster 8 0,5903 -0,8166 -1,3052 0,9392 0,2505
-0,2471 0,7281 0,1492 -0,0829 -0,1474 -0,3742
Cluster 9 0,5028 -1,2469 -0,6624 0,9089 -0,1651
-0,2052 -1,0430 -1,4345 -1,6911 -1,6538 0,2242
Cluster 10 -0,6177 1,0865 0,8683 -1,6434 -0,4168
-0,5399 1,0353 0,8056 0,4311 0,2782 0,2151
Cluster 11 0,7765 -0,8768 0,0056 0,9294 0,9523
1,3462 0,0421 0,1380 -0,6509 -0,8227 0,1066
Cluster 12 0,2780 0,1850 0,2850 0,2116 -1,0035
-0,7864 0,9102 0,6350 0,3458 0,3088 0,3686
Cluster 13 0,2227 0,6126 0,0056 -0,2070 -0,0570
0,2511 0,4209 0,0932 -0,2483 -0,2037 -4,4925
Cluster 14 0,0686 0,7059 0,0601 -0,6016 -0,3592
-0,7072 -0,7545 -0,6358 1,0401 0,9979 0,3209
Cluster 15 0,6148 -0,2208 -1,0787 0,6311 0,2839
0,4975 -1,0537 -1,3470 0,5656 0,5955 0,1758

Cluster Membership:

Cluster 1: 35 63 65 79 96 104 109 163 207 217 224 240 261 273 290
295 302 306 307 310 376 393 395 396 427 428

Cluster 2: 21 66 70 105 114 117 136 172 200 246 259 277 281 283 298 314 319 327 387 397 420 424 434 436
Cluster 3: 6 23 43 45 55 74 84 85 107 118 145 151 155 158 182 201 203 216 221 223 239 255 265 326 345 352 354 356 359 361 363 403 415 416 430
Cluster 4: 7 11 15 48 50 53 54 56 61 77 91 157 193 195 235 236 237 245 264 269 289 311 321 357 362 369 385 423
Cluster 5: 2 22 33 37 46 87 108 119 120 128 130 131 139 143 148 170 189 191 205 206 222 253 254 272 274 293 342 346 350 351 371 381 392 400 426 431 433
Cluster 6: 20 29 64 112 153 165 168 177 178 186 198 214 242 250 252 275 296 299 313 328 329 334 341 344 372 390 408 413 414 419
Cluster 7: 1 4 12 28 32 34 60 71 99 102 106 133 162 167 192 210 212 219 230 256 266 270 276 286 292 294 300 304 308 316 320 322 323 330 335 336 343 353 360 365 388 389 391 394 410 417 435
Cluster 8: 16 18 31 39 58 62 78 103 122 129 164 220 251 280 285 301 349 374 379 383 406 411 425 429
Cluster 9: 24 38 49 51 81 95 116 140 160 174 188 190 197 234 247 249 282 288 312 340 373 377 380 384 398 399 401 402 404 421
Cluster 10: 3 26 30 52 57 82 93 150 171 180 196 218 225 231 248 258 262 278 333 337 338 364 412 432
Cluster 11: 8 41 68 80 101 110 121 123 124 126 127 134 156 208 279 284 303 305 331 348 368 386
Cluster 12: 5 13 19 27 36 69 92 138 141 146 149 159 161 176 179 183 184 187 194 202 204 211 213 226 227 232 243 257 263 267 271 297 309 325 355 382 405 409
Cluster 13: 40 72 73 113 166 185 291 318 366 407 422
Cluster 14: 10 14 17 25 47 76 83 89 90 97 98 111 115 137 142 144 147 152 154 173 181 215 228 233 238 241 260 268 332 339
Cluster 15: 9 42 44 59 67 75 86 88 94 100 125 132 135 169 175 199 209 229 244 287 315 317 324 347 358 367 370 375 378 418

Chapter 6:
Special Issues

6.1 Mixed Measurement Levels

Different methods have been developed to handle mixed measurement levels. Gower's proposal is quoted very frequently (Everitt 1981: 16-17; Gordon 1999: 21-22; Wishart 2001). The idea of **Gower's (dis)similarity coefficient** is simple: Compute for each variable the appropriate (dis)similarity coefficient (that corresponds to the measurement level of each variable) and compute a weighted average as a general (dis)similarity coefficient. If d_{ikj} resp. s_{ikj} is the dissimilarity resp. similarity between case i and k in variable j and w_{ikj} is the weight for variable j, Gower's (dis)similarity coefficient is defined as

$$d_{ik} = \frac{\sum_k w_{ikj} \cdot d_{ikj}}{\sum_k w_{ikj}} \; resp. \; s_{ik} = \frac{\sum_k w_{ikj} \cdot s_{ikj}}{\sum_k w_{ikj}}.$$

Gower's (dis)similarity coefficient also allows pairwise deletion of missing values . If case i or j have a missing values in variable k, the weight w_{ijk} is set equal to zero.

The **city block metric** is very often recommended as dissimilarity measure for Gower's coefficient. The city block metric can be computed for all measurement levels. The maximum values of the city block metric depend on the measurement levels (see table 6-1). The weights are defined as the inverse value of the maximum distance. This guarantees that the maximum distance in each variable is 1.0.

measurement level	**maximum value of city block metric**	**weight**
binary variables	1 (if binary variables are treated as quantitative)	1/1
nominal variables	2 (if nominal variables are split into their dummies)	1/ 2
ordinal variables	range	1 / range
continuous variables	range	1 / range

Table 6-1: Weights for Gower's dissimilarity index using the city block metric

Instead of city block metric, squared Euclidean distance can be used. The weights are shown in table 6-2.

measurement level	**maximum value of city block metric**	**weight**
binary variables	1 (if binary variables are treated as quantitative)	1/1
nominal variables	2 (if nominal variables are split into their dummies)	1/ 2
ordinal variables	$range^2$	1 / $range^2$
quantitative variables	$range^2$	1 / $range^2$

Table 6-2: Weights for Gower's dissimilarity coefficient using squared Euclidean distances

Instead of weighting distances the variables can be transformed (weighted) (see chapter 2) by

$x'_{ij} = x_{ij}$ for binary variables

$x'_{ij(p)} = \begin{cases} 1/\sqrt{2} = 0.7071 \; if \; x_{ij} = p \\ 0 \; otherwhise \end{cases}$ for nominal variables.

($x'_{ij(p)}$ are the dummies of variable j)

$x'_{ij} = x_{ij} / r$ for ordinal and quantitative variables. (r is the range).

Instead of $x'_{ij} = x_{ij} / r$, the transformation $x'_{ij} = (x_{ij} - \min(x_j)) / r$ can be used. The resulting variables have values between 0 and 1. Interpretation is easier.

Generally, theoretical or the empirical values for the range can be used. Usually, the empirical range is selected. The transformation by 1/r can cause problems. Consider the following example: Two ordinal variables x1 and x2 are given:

x_1 = item with the response categories

1 = stronlgy agree	0.167	new range = 1.0
2 = agree	0.333	
3 = agree a bit	0.500	
4 = indifferent	0.667	
5 = disagree a bit	0.833	
6 = disagree	1.000	
7 = strongly disagree	1.167	

x_2 = item with the response categories

1 = agree	0.500	new range = 1.0
2 = indifferent	1.000	
3 = disagree	1.500	

As a consequence of the transformation by 1/r the difference between 'strongly agree' and 'strongly disagree' becomes equal to the difference between 'agree' and 'disagree'. Theoretical standardization (see chapter 3) can overcome this 'defect'.

Instead of the theoretical standardization, empirical standardization (z-transformation) can be applied.

$$x'_{ij} = (x_{ij} - \bar{x}_j) / s_x .$$

(A value greater than zero indicates a positive deviation from the average, a value less than zero a negative deviation.)

The (empirical) z-transformation raises the question, whether or not the binary variables and the dummies of the nominal variables should also be standardized. Little attention has been paid to this question in literature. In my opinion, all variables (binary variables, dummies, ordinal variables and quantitative variables) should be standardized. The dummies have to be multiplied by 0.707 after standardization.

Concerning squared Euclidean distances, z-transformation guarantees that the average of the distances in each variable is equal 1.

Summarizing the discussion, the following procedures are possible in the case of mixed measurement levels:

1. Standardizing all variables and then multiplying dummies by 0.707 (if squared Euclidean distances are used) or by 0.5 (if city block metric is used).
2. Weighting all variables - except the dummies – using '1 / standard deviation' and weighting the dummies using '0.707 / standard deviation' or '0.5 / standard deviation'.
3. Weighting the distances of the dummies using '0.5 / variance' and all other variables using '1 / variance', if squared Euclidean distances are used. Using city block metric, the weights for the distances are '0.5 / standard deviation' or '1 / standard deviation'.

The first two strategies lead to identical results:

$$d_{ik}(strategy_1) = d_{ik}(strategy_2)$$

The distances for the third strategy differ from the distances for the first two strategies by a scalar:

$d_{ik}(strategy_1) = d_{ik}(strategy_2) = \alpha \cdot d_{ik}(strategy_3)$, if Gower's formula is used for the third case.

The scalar has no influence on the results.

SPSS does not provide any strategy as option in CLUSTER or QUICK CLUSTER. Therefore, it is necessary to write a syntax program for strategy 1 or strategy 2. (Strategy 3 would require a programming of the whole cluster algorithm!). Generally, strategy 1 will be preferred to strategy 2, because the interpretation of the values of the variables is easier.

The syntax will be discussed for the following data matrix:

SEX	STUDY	GRADE	INCOME
1,00	1,00	1,00	3000,00
2,00	2,00	2,00	4000,00
1,00	3,00	1,00	2000,00
2,00	1,00	3,00	2500,00
2,00	1,00	2,00	3000,00
2,00	3,00	2,00	3000,00

The data file consists of the following variables:

sex	binary variable: 1 = female, 2 = male
branch of study	nominal variable with three categories: 1 = BWL, 2 = VWL and 3 = SOWI
grade	ordinal variable with five categories: 1 = excellent, 2 = good, 3 = satisfactory, 4 = poor, 5 = failed
income	quantitative variable in Euro

The data were read and printed using following syntax:

```
data list free/SEX  STUDY  GRADE  INCOME.
begin data.
1     1     1,0      3000
2     2     2,0      4000
1     3     1,0      2000
2     1     3,0      2500
2     1     2,0      3000
2     3     2,0      3000
end data.

list variables=sex study grade income.
```

Next, the variable SEX was recoded to 1/0 (1 = male) and three dummies were generated for the variable BRANCH OF STUDY.

```
compute male=sex.
recode male (1=0) (2=1).

compute bwl=study.
recode bwl (1=1) (2,3=0).
compute vwl=study.
recode vwl (2=1) (1,3=0).
compute sowi=study.
recode sowi (3=1) (1,2=0).

list variables=male,bwl,vwl,sowi,grade,income.
```

The new data matrix is:

MALE	BWL	VWL	SOWI	GRADE	INCOME
,00	1,00	,00	,00	1,00	3000,00
1,00	,00	1,00	,00	2,00	4000,00
,00	,00	,00	1,00	1,00	2000,00
1,00	1,00	,00	,00	3,00	2500,00
1,00	1,00	,00	,00	2,00	3000,00
1,00	,00	,00	1,00	2,00	3000,00

```
Number of cases read:  6    Number of cases listed:  6
```

The variables were standardized using DESCRIPTIVE in the next step.

```
DES VAR=MALE BWL VWL SOWI GRADE INCOME/SAVE.
```

Before CLUSTER was run, the dummies were multiplied by 0.707.

```
compute zbwl=zbwl*0.707.
compute zvwl=zvwl*0.707.
compute zsowi=zsowi*0.707.
```

```
cluster zmale,zbwl,zvwl,zsowi,zgrade,zincome
     /measure=seuclid
     /method=complete_linkage
    /print=distance schedule
    /plot=dendrogram.
```

The results are shown in figure 6-1. Three clusters can be distinguished. Cluster 1 contains persons 5, 6 and 4; cluster 2 person 2 and cluster 2 person 1 and 3.

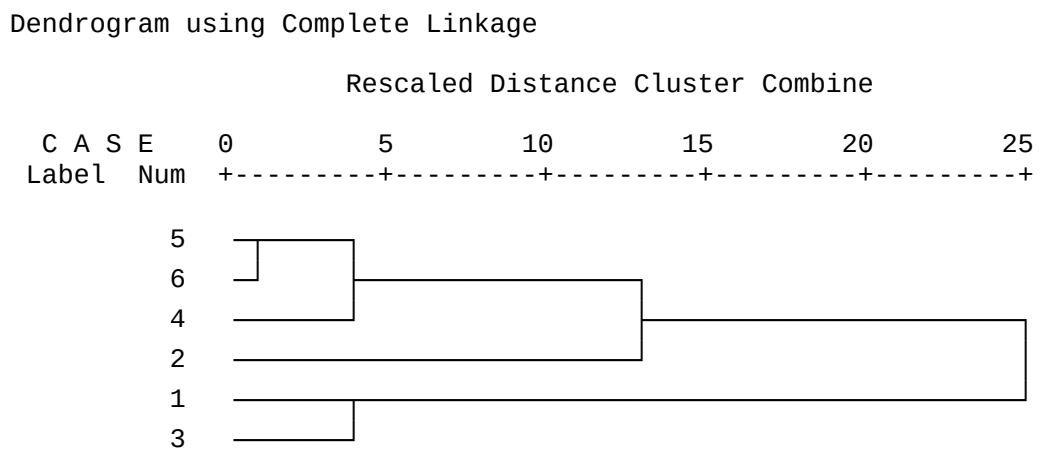


Figure 6-1: Results of a cluster analysis using mixed variables

For more details see Wishart (2001) or Bacher (1996: 173-191). Bacher only discusses the application of city-block metric. In this case the dummies must be multiplied by 0.5.

6.2 Weighting Variables

The selection of variables is crucial. Irrelevant variables can bias the result (see chapter 4.15). Therefore, different attempts were undertaken to develop methods selecting the best variables. FocalPoint allows you to weight the variables according to their F-value or t-value in the first stage. ALMO enables you to weight the variables by their pooled within error sum of squares. The pooled error sums are computed after each iteration. The consequence: Variables with a higher explained variance have more weight and more influence on the result.

In order to check, if irrelevant variables can be detected, we added four random variables to the data file of apprentices. Both kinds of analysis (no weighting, weighting) detected the irrelevant variables. The weighting of variables results in smaller differences between some model variables and irrelevant variables.

If we add six random variables, standard analysis still detects the six random variables as irrelevant. Two of the random variables are significant ((1-p)*100 > 95%), but their F values are considerably smaller than all the values of the original variables. For dynamic weighting, this is not the case. Two original variables become insignificant, whereas two random variables become significant (see table 6-3).

This small experiment shows that automatic weighting procedures should be used cautiously. For a further discussion of this problem see Gnanadesikan, Kettenring and Tsao (1995).

	F-Wert	(1-p)*100	ETA**2	F-Wert	(1-p)*100	ETA**2	
V12	33.051	100.000	0.146	32.283	100.000	0.143	ZPESSI
V13	53.461	100.000	0.209	57.251	100.000	0.221	ZINTER
V14	58.007	100.000	0.229	66.457	100.000	0.254	ZTRUST
V15	64.483	100.000	0.258	76.009	100.000	0.290	ZALIEN
V16	21.286	100.000	0.097	23.771	100.000	0.108	ZNORMLES
V17	55.188	100.000	0.221	56.585	100.000	0.226	ZVIOL
V18	120.743	100.000	0.388	112.090	100.000	0.371	ZGREEN
V19	67.366	100.000	0.263	67.015	100.000	0.262	ZSPD
V20	186.923	100.000	0.496	172.256	100.000	0.476	ZCDU
V21	207.502	100.000	0.525	186.532	100.000	0.498	ZCSU
V22	125.750	100.000	0.388	106.126	100.000	0.349	ZSAFE
				0.492	30.747	0.002	r1
				9.368	99.996	0.044	r2
				3.218	97.796	0.015	r3
				1.164	67.755	0.006	r4
				2.110	90.368	0.010	r5
				1.763	84.867	0.009	r6

(a) no weighting, no random variables
ETA**2 = 29.5%;
Improvement of H0 = 0.295

(b) no weighting, six random variables
ETA**2 = 18.9%;
Improvement of H0 = 0.189

V12	18.065	100.000	0.085	5.647	99.886	0.028	ZPESSI
V13	7.542	99.983	0.036	4.390	99.509	0.021	ZINTER
V14	45.747	100.000	0.190	8.699	99.993	0.043	ZTRUST
V15	49.469	100.000	0.210	19.410	100.000	0.095	ZALIEN
V16	15.163	100.000	0.071	0.949	58.183	0.005	ZNORMLES
V17	27.134	100.000	0.123	1.153	67.325	0.006	ZVIOL
V18	22.541	100.000	0.106	13.796	100.000	0.068	ZGREEN
V19	15.309	100.000	0.075	11.704	99.999	0.059	ZSPD
V20	1040.665	100.000	0.846	792.679	100.000	0.807	ZCDU
V21	1643.441	100.000	0.898	1261.786	100.000	0.871	ZCSU
V22	7.986	99.988	0.039	164.088	100.000	0.453	ZSAFE
				0.752	47.555	0.004	r1
				3.823	99.000	0.018	r2
				3.517	98.514	0.017	r3
				1.761	84.810	0.009	r4
				0.344	20.410	0.002	r5
				0.565	35.757	0.003	r6

(c) weighting with the pooled within variance, no random variables
ETA**2 = 12.8%;
Improvement of H0 = 0.156

(d) weighting with the pooled within variance, six random variables
ETA**2 = 14.3%;
Improvement of H0 = 0.403

Table 6-3: Consequences of automatically weighted variables

6.3 Finding Statistical Twins

Different problems, like the imputation of missing values, data fusion, statistical re-identification etc. (see chapter 1) require the finding of statistical twins. Statistical twins can be found using cluster analysis. The problem can be described in the following way: Two data sets (data set I and data set II) are given. A statistical twin should be found in data set II for each case in data set I. The statistical twin can be used for the following problems:

- **Estimating missing values (data imputation)**. In this case the two data sets are generated from one data set. Data set I contains the cases with missing values, data set II the cases without missing values.
- **Estimating missing variables (data fusion).** Some variables are not collected in data set I. They are available in data set II. The two data sets should be combined the way that data set I contains also the variables, that are only observed in data set II.
- **Finding a control group (optimal matching)**. A control group for data set I should be generated from data set II, e.g. registered data.
- **Computing the re-identification risk.** A research group in a university has collected data (data set I). The group asks the statistical office for further comparative statistical data (=data set II). The statistical office calculates the identification risk. If it is lower than a certain threshold, the statistical office can give away the data.

Example (Brand/Bender/Bacher 2001):
Data set I contains 273 cases, data set II 19355 cases. 14 variables are common and should be used to find statistical twins. The variables are:

- occupation (327 categories, treated as interval scaled)
- birth cohort (dichotomous)
- daily earnings (interval scaled)
- sex (dichotomous)
- land of the Federal Republic of Germany (nominal scaled: 11)
- nationality (dichotomous)
- schooling/training (nominal: 3)
- occupational status (nominal: 4)
- part time (dichotomous)

- interruption in the working life (dichotomous)
- number of months in the working life (count variable maximum: 24 months)
- duration of the last employment (count variable maximum: 24 months)
- marital status (dichotomous)
- number of children (count variable)

The general procedure to find statistical twins using cluster analysis consists of the following steps:

1. Transform variables, if necessary. In our example transformation is necessary because mixed data should be used and quantitative variables have different scales (see chapter 2)
2. Run a cluster analysis of data set I. Contrary to the usual application of the clustering procedures, as many clusters as possible should be computed. Under ideal conditions, the number of clusters should be equal to the number of cases.
3. Save the cluster centres.
4. Read data set II. Transform the variables, if necessary. Use the weights of step 1.
5. Assign the cases from data set II to the clusters obtained for data set I.
6. Select the nearest case of data set II as statistical twin.
7. Build a new data set. Add or match the files depending on the problem you analyse.

If the programme does not find statistical twins for all cases, step 2 to 7 can be repeated for those cases without a twin. The statistical twins in data set II must be eliminated in advance.

The appendix shows the syntax for our example. Nominal variables are transformed to dummies. All variables are weighted using the inverse value of their standard deviation. In addition, the dummies are multiplied by 0.707 (see chapter 6.1). The cluster analysis of step 2 reports 232 clusters. 41 cases were eliminated due to missing values. Each cluster contains one case. Step 4 finds for 216 clusters one statistical twin. Three of the statistical twins have distances greater than 6.0 to their 'real' twin. The user has to decide, whether or not to use these twins. Generally, 53 percent of twins have a distance smaller than 0.5.

For further analysis, only the 216 cases in data set I with a statistical twin were selected.

Notes:

- The procedure can be repeated for the cases eliminated due to missing values or due to missing twins. In the first case, those (or some of those) variables with missing values are ignored. Steps 2 to 7 are repeated for the cases with missing values. In data set II the statistical twins have to be eliminated in advance to avoid a case beeing used more than once as a twin. The procedure for the second situation is similar. Steps 2 and 7 are repeated for the cases without statistical twins. Again, it is necessary to delete the statistical twins in data set II in advance.
- The variables may be weighted. It can be – for example – assumed that sex, day of birth and occupation are more stable. Hence they should be weighted (multiplied) using the factor 10. The corresponding SPSS commands are:

```
compute geb71=10*geb71/0.50.
compute weibl=10*weibl/0.49.
compute beruf=10*beruf/216.74.
```

- The procedure is very restrictive. A statistical twin satisfies three conditions: (1.) The statistical twin is assigned to the same cluster as its 'real' twin. (2.) The statistical twin is the nearest neighbour to its 'real' twin. (3.) A statistical twin is used only once. These restrictions can result in the exclusion of 'better' statistical twins. Some cases of data set II can have smaller distances to their 'real' twin, but are ignored, because they are assigned to another cluster.
- The procedure assumes that each cluster consists of one case. If this is not the case, the assignment rules have to be changed.

Summarizing, it is possible to find statistical twins using SPSS. However, this can result in a complex syntax programme.

6.4 Analysing Large Data Sets using SPSS

CLEMENTINE (see chapter 5.11, SPSS 2000) provides a two-stage clustering procedure for large data sets. This strategy can also be used in SPSS to cluster large data sets. The steps are:

1. Compute K compact clusters using QUICK CLUSTER. Set K equal to 200 or higher. The variance within the clusters should be very small. Save the centres for further analysis. Instead of centres examplers can be used.
2. Run a hierarchical cluster analysis using the centres as cases. Ward's method or any other method can be used. It is also possible to use other distance measures.

References

Bacher, J., 1996: Clusteranalyse [Cluster analysis]. Opladen. [available only in German]

Brand, R., Bender, S., Bacher, J., 2001: Re-Identifying Register Data by Survey Data: An Empirical Study. forthcoming.

Everitt, B., 1981: Cluster analysis. Second edition. New York.

Gnanadesikan, R., Kettenring, J.R., Tsao, S.L., 1995: Weighting and Selection of Variables in Cluster Analysis. Journal of Classification, Vol. 12, 113-136.

Gordon, A. D., 1999: Classification. 2nd edition. London-New York.

SPSS Inc., 2000: The SPSS TwoStep cluster component. White paper – technical report.

Wishart, D., 2001: Gower's Similarity Coefficient.
http://www.clustan.com/gower_similarity.html

Appendix A6-1: Syntax for Finding Statistical Twins

```
* read data set I.

data list file='c:\texte\bender\mpi.dat' free/
  id gebjahr sex monent beruf bula deutsch schul ausb
  nstib teil luecke gesdur lastdur famst kidz sample.
execute.

*new variables are generated for the further analysis.

recode beruf (-1=sysmis).

compute geb71=gebjahr.
recode geb71 (64=0) (71=1).

compute eink=monent.

compute inl=deutsch.

compute weibl=sex.
recode weibl (1=0) (2=1).

compute abi=schul.
recode abi (1=0) (2=1) (9=sysmis).

*eduction is split into dummies.
compute ausb1=ausb.
recode ausb1 (1=1) (2,3,4=0) (9=sysmis).
compute ausb2=ausb.
recode ausb2 (2=1) (1,3,4=0) (9=sysmis).
compute ausb3=ausb.
recode ausb3 (3=1) (1,2,4=0) (9=sysmis).
compute ausb4=ausb.
recode ausb4 (4=1) (1,2,3=0) (9=sysmis).

*occupational status is split into dummies.
compute beruf0=nstib.
recode beruf0 (0=1) (9=sysmis) (else=0).
compute beruf1=nstib.
```

```
recode beruf1 (1=1) (9=sysmis) (else=0).
compute beruf2=nstib.
recode beruf2 (2=1) (9=sysmis) (else=0).
compute beruf3=nstib.
recode beruf3 (3=1) (9=sysmis) (else=0).
compute beruf4=nstib.
recode beruf4 (4=1) (9=sysmis) (else=0).
compute beruf5=nstib.
recode beruf5 (5=1) (9=sysmis) (else=0).

*federal country is split into dummies.
compute land1=bula.
recode land1 (1=1) (99=sysmis) (else=0).
compute land2=bula.
recode land2 (2=1) (99=sysmis) (else=0).
compute land3=bula.
recode land3 (3=1) (99=sysmis) (else=0).
compute land4=bula.
recode land4 (4=1) (99=sysmis) (else=0).
compute land5=bula.
recode land5 (5=1) (99=sysmis) (else=0).
compute land6=bula.
recode land6 (6=1) (99=sysmis) (else=0).
compute land7=bula.
recode land7 (7=1) (99=sysmis) (else=0).
compute land8=bula.
recode land8 (8=1) (99=sysmis) (else=0).
compute land9=bula.
recode land9 (9=1) (99=sysmis) (else=0).
compute land10=bula.
recode land10 (10=1) (99=sysmis) (else=0).
compute land11=bula.
recode land11 (11=1) (99=sysmis) (else=0).

compute teilz=teil.
compute lluecke=luecke.
compute ggdur=gesdur.
compute verh=famst.
compute kinder=kidz.

desc var=geb71 eink weibl inl abi ausb1 ausb2 ausb3 ausb4
          land1 to land11
```

```
         beruf0 beruf1 beruf2 beruf3 beruf4 beruf5
         teilz lluecke ggdur verh kinder beruf.

* mark the programme and execute it.
* The results of desc are necessary for
* weighting variables.
* Each variable is weighted by '1/standard deviation'
* Dummies of nominal variables are multiplied by 0.707,
* to garantuee commensurability.

compute geb71=geb71/0.50.
compute eink=eink/1711.69.
compute weibl=weibl/0.49.
compute inl=inl/0.15.
compute abi=abi/0.48.
compute ausb2=0.707*ausb2/0.37.
compute ausb3=0.707*ausb3/0.23.
compute ausb4=0.707*ausb4/0.31.
compute land1=0.707*land1/0.25.
compute land2=0.707*land2/0.12.
compute land3=0.707*land3/0.39.
compute land4=0.707*land4/0.13.
compute land5=0.707*land5/0.43.
compute land6=0.707*land6/0.25.
compute land7=0.707*land7/0.12.
compute land8=0.707*land8/0.40.
compute land9=0.707*land9/0.38.
compute land10=0.707*land10/0.12.
compute land11=0.707*land11/0.10.
compute beruf1=0.707*beruf1/0.26.
compute beruf2=0.707*beruf2/0.42.
compute beruf3=0.707*beruf3/0.11.
compute beruf4=0.707*beruf4/0.47.
compute teilz=teilz/0.30.
compute lluecke=lluecke/0.48.
compute ggdur=ggdur/5.33.
compute verh=verh/0.50.
compute kinder=kinder/0.90.
compute beruf=beruf/216.74.


* k-means. The maximal number of clusters
```

```
* is determined. The cluster centres are saved
* for the next step.

QUICK CLUSTER
         geb71 eink weibl inl ausb2 ausb3 ausb4
         land1 to land11
         beruf1 beruf2 beruf3 beruf4
         teilz lluecke ggdur verh kinder beruf
  /MISSING=listwise
  /CRITERIA= CLUSTER(232) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (clus) DISTANCE (dclus)
  /PRINT ANOVA
  /OUTFILE='C:\texte\bender\cmpi1.sav'.

recode clus (sysmis=-1).
select if (clus > 0).
sort cases by clus.
compute set1=1.
fre var=set1.
save outfile='c:\texte\bender\mpiclu.sav'.
execute.


*read data set II.

data list file='c:\texte\bender\iab.dat' free/
  id gebjahr sex monent beruf bula deutsch schul ausb
  nstib teil luecke gesdur lastdur famst kidz sample.
execute.

*generate new variables.
*see above.

recode beruf (-1=sysmis).

compute geb71=gebjahr.
recode geb71 (64=0) (71=1).

compute eink=monent.

compute inl=deutsch.
```

```
compute weibl=sex.
recode weibl (1=0) (2=1).

compute abi=schul.
recode abi (1=0) (2=1) (9=sysmis).

compute ausb1=ausb.
recode ausb1 (1=1) (2,3,4=0) (9=sysmis).
compute ausb2=ausb.
recode ausb2 (2=1) (1,3,4=0) (9=sysmis).
compute ausb3=ausb.
recode ausb3 (3=1) (1,2,4=0) (9=sysmis).
compute ausb4=ausb.
recode ausb4 (4=1) (1,2,3=0) (9=sysmis).

compute beruf0=nstib.
recode beruf0 (0=1) (9=sysmis) (else=0).
compute beruf1=nstib.
recode beruf1 (1=1) (9=sysmis) (else=0).
compute beruf2=nstib.
recode beruf2 (2=1) (9=sysmis) (else=0).
compute beruf3=nstib.
recode beruf3 (3=1) (9=sysmis) (else=0).
compute beruf4=nstib.
recode beruf4 (4=1) (9=sysmis) (else=0).
compute beruf5=nstib.
recode beruf5 (5=1) (9=sysmis) (else=0).

compute land1=bula.
recode land1 (1=1) (99=sysmis) (else=0).
compute land2=bula.
recode land2 (2=1) (99=sysmis) (else=0).
compute land3=bula.
recode land3 (3=1) (99=sysmis) (else=0).
compute land4=bula.
recode land4 (4=1) (99=sysmis) (else=0).
compute land5=bula.
recode land5 (5=1) (99=sysmis) (else=0).
compute land6=bula.
recode land6 (6=1) (99=sysmis) (else=0).
compute land7=bula.
```

```
recode land7 (7=1) (99=sysmis) (else=0).
compute land8=bula.
recode land8 (8=1) (99=sysmis) (else=0).
compute land9=bula.
recode land9 (9=1) (99=sysmis) (else=0).
compute land10=bula.
recode land10 (10=1) (99=sysmis) (else=0).
compute land11=bula.
recode land11 (11=1) (99=sysmis) (else=0).


compute teilz=teil.
compute lluecke=luecke.
compute ggdur=gesdur.
compute verh=famst.
compute kinder=kidz.

* weight the variables.
* Note: the weights of the first analysis must be
* used.

compute geb71=geb71/0.50.
compute eink=eink/1711.69.
compute weibl=weibl/0.49.
compute inl=inl/0.15.
compute abi=abi/0.48.
compute ausb2=0.707*ausb2/0.37.
compute ausb3=0.707*ausb3/0.23.
compute ausb4=0.707*ausb4/0.31.
compute land1=0.707*land1/0.25.
compute land2=0.707*land2/0.12.
compute land3=0.707*land3/0.39.
compute land4=0.707*land4/0.13.
compute land5=0.707*land5/0.43.
compute land6=0.707*land6/0.25.
compute land7=0.707*land7/0.12.
compute land8=0.707*land8/0.40.
compute land9=0.707*land9/0.38.
compute land10=0.707*land10/0.12.
compute land11=0.707*land11/0.10.
compute beruf1=0.707*beruf1/0.26.
compute beruf2=0.707*beruf2/0.42.
```

```
compute beruf3=0.707*beruf3/0.11.
compute beruf4=0.707*beruf4/0.47.
compute teilz=teilz/0.30.
compute lluecke=lluecke/0.48.
compute ggdur=ggdur/5.33.
compute verh=verh/0.50.
compute kinder=kinder/0.90.
compute beruf=beruf/216.74.

*assign cases.

QUICK CLUSTER
  geb71 eink weibl inl ausb2 ausb3 ausb4
         land1 to land11
         beruf1 beruf2 beruf3 beruf4
         teilz lluecke ggdur verh kinder beruf
  /MISSING=listwise
  /CRITERIA= CLUSTER(232)
  /METHOD=classify
  /SAVE CLUSTER (clus) DISTANCE (dclus)
  /PRINT ANOVA
  /FILE='C:\texte\bender\cmpi1.sav'.

*select the nearest case.

recode clus(sysmis=-1).
select if (clus>0).
sort cases by clus dclus.

compute cclus=lag(clus).
recode cclus(sysmis=0).
if (clus ne cclus) nonelim=1.
execute.

select if (nonelim=1).
compute clus2=clus.
compute set2=1.
fre var=dclus.
save outfile='c:\texte\bender\iabclu.sav'.
execute.
```

```
*match files.

get file='c:\texte\bender\mpiclu.sav'.

match files file=*
   /file='c:\texte\bender\iabclu.sav'
   /by clus/map.
execute.

recode clus2(sysmis=-1),

compute fehlend=1.
if (clus2 eq clus) fehlend=0.

fre var=fehlend.
temp.
select if (fehlend=1).
list variables id clus clus2.
execute.


select if (fehlend=0).
execute.

*add files for further analysis.

add files
    /file=*
   /file='c:\texte\bender\iabclu.sav'
   /map.
execute.


recode set1(sysmis=2).
fre var=set1.

crosstabs tabels=sex by set1/cells=count column/stat=chisq.
```

Chapter 7:
Probabilistic Clustering

Note: This chapter is based on Bacher (2000) and Bacher (1996: 353-408).

7.1 A Probabilistic Clustering Model for Variables of Mixed Type

K-means clustering models are widely used for partitioning large data bases to homogeneous clusters (Jain and Dubes 1988: 90). However, k-means clustering has certain disadvantages:

- The variables must be commensurable (Fox 1982, chapter 2). This implies interval or ratio variables with equal scale units.
- Each pattern (case) is assigned deterministically to one and only one cluster. This may result in biased estimators of the cluster means if the clusters overlap.
- There is no accepted statistical basis, even though a lot of approaches are now available (Bock 1989, Bryant 1991, Jahnke 1988, Pollard 1981, 1982).

This chapter develops a probabilistic clustering model for variables of mixed type (therefor labeled as general probabilistic clustering model) that overcomes the problems of k-means clustering:

- Variables with different measurement levels — nominal, ordinal and/or interval or ratio (=quantitative variables) — and different scale units can be analyzed without any transformation of the variables.
- Each pattern is assigned probabilistically to the clusters.
- Finally, the model has a statistical basis, the maximum likelihood approach.

The Model

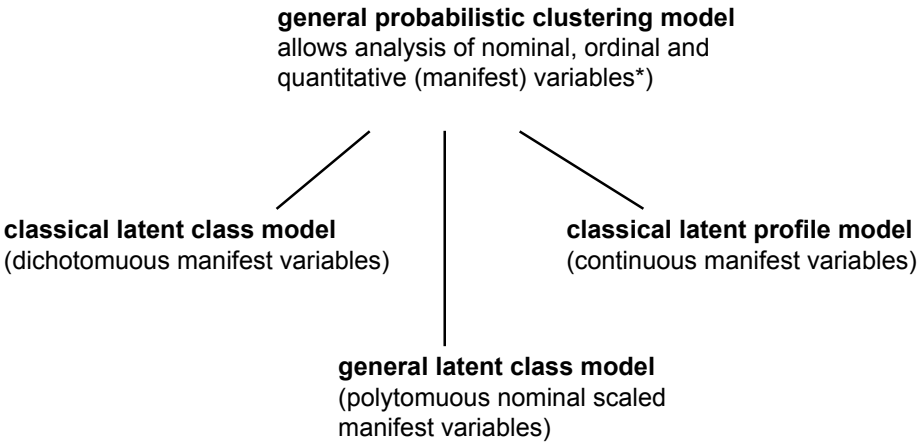
The main idea of model is to use probabilities $\pi(k/g)$ instead of distances d_{gk}^2, as is the case in k-means clustering. $\pi(k/g)$ is the probability that pattern g belongs to cluster k, whereas d_{gk}^2 is the squared Euclidean distance between pattern g and the center of cluster k. By Bayes' theorem, $\pi(k/g)$ may be expressed as follows

$$\pi(k/g) = \frac{\pi(k) \cdot \pi(g/k)}{\sum_{k=1}^{K} \pi(k) \cdot \pi(g/k)}$$

where $\pi(k)$ is the probability that a random selected pattern belongs to cluster k (k = 1, 2, ..., K) and $\pi(g/k)$ is the probability of observing the pattern g given cluster k. The model assumes that the $\pi(g/k)$'s may be expressed as functions of unknown parameters of the clusters k:

$$\pi(g/k) = \pi(x_{g1}/\theta_{1k}) \cdot \pi(x_{g2}/\theta_{2k}) \cdot \ldots \cdot \pi(x_{gm}/\theta_{mk}) = \prod_{i=1}^{m} \pi(x_{gi}/\theta_{ik})$$

where θ_{ik} are the unknown parameters of cluster k for the variables i (i = 1, 2, ..., m), x_{gi} is the value of pattern g in variable i, and $\pi(x_{gi}/\theta_{ik})$ is the probability of observing x_{gi} given the parameters θ_{ik} of cluster k for variable i. The model is based on the assumption that the probability of observing x_{gi} is independent of the probabilities of observing x_{gi*} within each cluster k for all $i^* \neq i$. This assumption is well known from latent class analysis as the axiom of local independence (for instance, Fielding 1977: 128). In fact, our probabilistic clustering model contains all known models of latent class analysis (see figure 7-1). But it also allows analysis of data with variables of mixed type. The manifest variables may be nominal, ordinal *and* quantitative. For each variable i, $\pi(x_{gi}/\theta_{ik})$ is calculated. $\pi(x_{gi}/\theta_{ik})$ is a probability and can only vary between zero and one. As a consequence, the problem of incommensurability does not occur.



*) Quantitative variables are interval or ratio variables. They may be discrete or continuous. If they are discrete with only few categories, it is better to treat them as ordinal variables.

Figure 7-1: Latent class models as submodels of a general probabilistic clustering model

The functional form of the $\pi(x_{gi}/\theta_{ik})$´s and the parameters θ_{ik} depend on the measurement levels of the variables i. If variable i is nominal, $\pi(x_{gi}/\theta_{ik})$ is

$$\pi(x_{gi} / \theta_{ik}) = \pi_{i(j)k} \quad \text{for } x_{gi} = j \text{ (pattern g has category j in variable i)}$$

where pattern g has the value j in variable i and $\pi_{i(j)k}$ is the probability of observing category j in variable i given cluster k.

If variable i is ordinal, our model uses the latent class model for ordinal variables of Rost (1985). $\pi(x_{gi}/\theta_{ik})$ is

$$\pi(x_{gi} / \theta_{ik}) = \binom{m_i - 1}{x_{gi}} \cdot \pi_{ik}^{x_{gi}} \cdot (1-\pi_{ik})^{m_i - x_{gi}},$$

where x_{gi} is the score of the pattern g in the variable i, m_i is the number of categories of variable i, and π_{ik} is the probability of a higher score in the variable i in cluster k. For the model, the scores of the ordinal variable must begin with zero. The actual interpretation of π_{ik} depend on the coding of the ordinal variable. If we analyse for example educational levels, where 0 indicates low education and 2 high education, π_{ik} is the probability of having higher education. If the ordinal variable i is an attitude scale, where higher number expresses disagreement, π_{ik} is the probability of higher disagreement.
For quantitative variables, our model uses the following function for the $\pi(x_{gi}/\theta_{ik})$'s:

$$\pi(x_{gi} / \theta_{ik}) = \varphi\left(\frac{x_{gi} - \mu_{ik}}{\sigma_{ik}}\right)$$

where $\varphi(...)$ is the density function of the standard normal distribution, x_{gi} is the value of pattern i in the quantitative variable i, μ_{ik} is the mean of cluster k, and σ_{ik} is the standard deviation in the quantitative variable i. Contrary to the classical latent profile model, we assume a normal distribution of the patterns for each quantitative variable in each cluster k. This is necessary for the maximum likelihood estimation (see next section). However, it

should be noted that our model is more restrictive in this aspect than the classical latent profile model.

Table 7-1 summarizes the parameters of the general probabilistic model. For K clusters, the number of parameters r to be estimated is

$$
\begin{aligned}
r &= (K-1) + K \cdot \sum_{i=1}^{I_{nom}} (J_i - 1) \cdot + I_{ord} \cdot K + 2 \cdot K \cdot I_{quant} \\
&= (K-1) + K \cdot \left(\sum_{i=1}^{I_{nom}} (J_i - 1) + I_{ord} + 2 \cdot I_{quant} \right)
\end{aligned} \tag{6}
$$

where I_{nom} is the number of nominal variables, J_i is the number of categories for the nominal variable i, I_{ord} is the number of ordinal variables, and I_{quant} is the number of quantitative variables.

Parameters θ_{ik}	Interpretation	Restrictions
for the clusters:		
$\pi(k)$	probability that a random selected pattern belongs to cluster k	$\sum_{k=1}^{K} \pi(k) = 1$
for nominal variables:		
$\pi_{i(j)k}$	probability of observing category j for variable i in cluster k	$\sum_{j=1}^{J_i} \pi_{i(j)k} = 1$ for all i's and k's
for ordinal variables		
π_{ik}	probability of observing a higher or lower value in variable i in cluster k	
for quantitative variables		
μ_{ik}	mean of variable i in cluster k	
σ_{ik}	standard deviation of variable i in cluster k	

Table 7-1: The parameters θ_{ik} of the general probabilistic clustering model

Estimation of the Model

The parameters can be estimated by the EM algorithm (see Bacher 2000).

Model Fit and Model Selection

In practice, the number K of clusters is unknown. To determine the number of clusters, the model is estimated for different K's, where K should start with one. After estimation, the likelihood ratio test can be used for model selection. For two cluster solutions with K and K* clusters (K* < K), the LR test statistic is defined as

$$LR(K-K^*) = -2 \cdot (l_{K^*} - l_K)$$

where l_{K^*} is the log-likelihood value for the model with K* clusters and l_K is the log-likelihood value for the model with K clusters. LR(K-K*) is asymptotically distributed as chi-square with r - r* degrees of freedom, where r is the number of parameters for the model with K parameters and r* is the number of parameters for the model with K* parameters.
The LR(K-K*) statistics can be used to select the most appropriate number of clusters K within a set of cluster solutions {1, 2, ..., Kmax} by the following rule: Choose the cluster solution K that satisfies the following conditions: (1) All LR(K**-K*) are significant, if K* < K and K** ≥ K. (2) All LR(K**-K*) are not significant if K** > K* and K* ≥ K. This rule was originally proposed by Kendall (1980: 41-42) for k-means clustering. In practice, the rule may lead to no definitive solution and should not be applied automatically.

Instead of the LR test, the adjusted LR test of Wolfe (for instance, Everitt 1981: 66) can be used. The adjusted LR test statistic is defined as:

$$LR^{adj}(K-K^*) = -2 \cdot \tfrac{1}{G} \cdot \left(G - 1 - r - \frac{K^*}{2} \right) \cdot \left(l_{K^*} - l_K \right)$$

where K* < K, r is equal $\sum_{i=1}^{I_{nom}} (J_i - 1) + I_{ord} + I_{quant}$

In addition, the following goodness-of-fit indices may be computed

$$GFI_0(K) = 1 - \frac{l_K}{l_1}$$

$$GFI_1(K) = 1 - \frac{l_K}{l_{K-1}}$$

The $GFI_0(K)$ index measures the amount of deviance of the one-cluster model explained by the model with K clusters; the $GFI_1(K)$ the proportional improvement of deviance explained by the model with K clusters in contrast to the model with K-1 clusters. Both indices are heuristic measures. Both of them should be between zero and one, although $GFI_1(K)$ might become negative if the algorithm fails to converge.

Modification of the model

Similarly to the structural equation models, certain parameters may be fixed or constrained. Fixed parameters are parameters that have been assigned given values. These values are not changed during the iteration. Contrary to fixed parameters, constrained parameters are unknown but equal to one or more other parameters. In a model - for example - the mean μ_{ik} may be set equal to the mean μ_{i*k}. The main advantage of imposing restrictions on the parameters is to facilitate interpretation. To judge a constrained model, the test statistics and measures of the previous section can be applied.

An Example with Artificial Data

In order to demonstrate the reliability of the general probabilistic model, artificial data with a known cluster structure will be analyzed in this section. The artificial data set consists of two clusters and three independent variables. In each variable cluster 1 has a normal distribution with a mean of +1.5 and a standard deviation of 1.0; cluster 2 a normal distribution with a mean of -1.5. Each cluster consists of 500 patterns. Two of the three variables have been categorized for the analysis. A category value of 0 was assigned if the value of the variable was less than -1; if the value was greater than or equal to -1 and less than +1, a category value of 1 was assigned; and if the value was greater than or equal to +1, a category value of 2 was assigned. One of the two variables was analyzed as a nominal variable, the other one as an ordinal variable.

For comparison, a second artificial data set was generated by the same logic. Contrary to the first artificial data, both clusters have means of 0 in the variables. Whereas the first artificial data set represents a two-cluster structure, the second artificial data set represents no cluster structure. For the first artificial data set, we expect the following results: (1) All LR(K**-K*) values should be significant for the one-cluster solution, if K** = 1 and K* ≥ 1, whereas all LR(K**-K*) values should not be significant, if K** > K* and K* ≥ 1. (2) GFI_0(K=1) assumes a relatively high value, and GFI_1(K=1), whereas GFI_0(K) should not increase for solutions with K > 2, and GFI_1(K) should have values near zero for K > 2. (3) The estimated parameters should not differ too much from the true parameters.

Table 7-2 shows the number of iterations, the values of the log-likelihood functions, and the values of the goodness-of-fit measures. The log-likelihood functions do not decrease continuously with the number of clusters. The log-likelihood value of the 5-cluster solutions is larger than the functions value for the 4-cluster solutions. However, the EM algorithm needs more than 500 iterations to find any solutions for 5 clusters. As mentioned earlier, a large number of iterations very often indicates that the model does not fit for the analyzed number of clusters. This will also be the case for the artificial data set with no cluster structure.

K	number of iterations	log(L(K))	GFI0(K) in %	GFI1(K) in %
1	3	-4075.082	0.000	-
2	24	-3514.699	13.751	13.751
3	128	-3511.324	13.834	0.096
4	146	-3508.804	13.896	0.072
5	528	-3509.433	13.881	-0.018
6	441	-3506.367	13.956	0.087

Table 7-2: Number of iterations, log-likelihood functions and goodness of-fit measures for the artificial data set with two clusters

The goodness-of-fit measures show a sharp increase between the one-cluster and the two-cluster solution. The increase is about 14%. Note that GFI_0(K=1) always equals GFI_1(K=1). For the further solutions, GFI_0(K) does not increase and GFI_1(K) is nearly 0. This indicates a two-cluster solution. The LR(K-K*) test statistics lead to the same result (see table 7-3): All LR((K**=1)-K*) are significant for K* = 2, 3, .., 6 and all LR(K**-K*) are not significant if K** > K* and K* = 2, 3, 4 or 5.

LR(K-K*) test statistics
(lower triangle = test statistics,
upper triangle = significance level 100·(1-p),
where p is the error level)

	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6
1	0	100.0000	100.0000	100.0000	100.0000	100.0000
2	1120.7668	0	65.4822	53.5898	8.7839	13.9032
3	1127.5162	6.7494	0	45.9718	1.3617	6.6188
4	1132.5562	11.7894	5.0401	0	-(a)	3.8482
5	1131.2984	10.5316	3.7822	-(a)	0	59.0204
6	1137.4299	16.6631	9.9137	4.8736	6.1315	0

adjusted LR(K-K*) test statistics of Wolfe (lower triangle = test
statistics, upper triangle = significance level 100·(1-p),
where p is the error level)

	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6
1	0	100.0000	100.0000	100.0000	100.0000	100.0000
2	1114.6026	0	43.0370	23.6999	0.8269	1.1865
3	1121.3148	6.7089	0	24.1891	0.1107	0.5419
4	1126.3272	11.7187	5.0073	0	-(a)	0.4065
5	1125.0763	10.4684	3.7577	-(a)	0	36.0685
6	1131.1740	16.5631	9.8493	4.8395	6.0855	0

(a) This value cannot be computed because the log-likelihood value increases between the four-cluster and the five-cluster solution

Table 7-3: LR test statistics for the first artificial data set with two clusters

Finally, the estimated parameters do not differ too much form the true parameters.

Contrary to the first artificial data set with a two-cluster structure, no sharp increase for the goodness-of-fit measures can be found for the second artificial data set with no cluster structure. The measures are all smaller than 1% (see table 7-4). For each cluster solution except for the one-cluster solution, a large number of iterations are needed. The LR(K-K*) values show no significance (see table 7-5). This indicates that no cluster structure underlies the data.

K	number of iterations	log(L(K))	GFI0(K) in %	GFI1(K) in %
1	3	-2626.448	0.000	-
2	225	-2626.175	0.010	0.010
3	936	-2625.189	0.048	0.038
4	1001	-2624.535	0.073	0.025
5	756	-2620.473	0.227	0.155
6	1001	-2619.593	0.261	0.034

Table 7-4: Number of iterations, log-likelihood functions and goodness of-fit measures for the second artificial data set with no cluster structure

```
LR(K-K*) test statistics (lower triangle = test statistics,
upper triangle = significance level 100·(1-p),
where p is the error level)

      K = 1    K = 2    K = 3    K = 4    K = 5    K = 6
1     0       0.4037   0.2478   0.0366   2.0221   0.5355
2    0.5458    0       7.8273   0.7476  12.4606   3.7578
3    2.5174   1.9716    0       3.0206  33.3343  11.5092
4    3.8262   3.2804   1.3088    0      77.0532  37.2940
5   11.9490  11.4032   9.4316   8.1228    0       6.0577
6   13.7105  13.1647  11.1931   9.8843   1.7614    0

adjusted LR(K-K*) test statistics of Wolfe (lower triangle = test
statistics, upper triangle = significance level 100·(1-p),
where p is the error level)

      K = 1    K = 2    K = 3    K = 4    K = 5    K = 6
1     0       0.0522   0.0163   0.0014   0.0735   0.0107
2    0.5428    0       1.8770   0.0551   1.4228   0.1645
3    2.5036   1.9598    0       0.5440  10.3488   1.2504
4    3.8052   3.2607   1.3003    0      57.1675  12.4756
5   11.8833  11.3348   9.3703   8.0660    0       1.3261
6   13.6351  13.0857  11.1203   9.8151   1.7482    0
```

Table 7-5: LR test statistics for the second artificial data set with no cluster structure

An Empirical Example

The goodness-of-fit indices and the LR test statistics perform quite well for the artificial data sets. In practice, the situation is very often more complex and the determination of the number of clusters can be difficult. Different cluster solutions may be appropriate. This is the case for the example of table 7-6. The results are based on a survey of 1960 children in Austria. The aim of the analysis was to find out if the children can be assigned to different social classes. The occupations of the parents, their educational levels and the per capita household income were used as classification variables. The occupation was treated as nominal variable, the educational level as ordinal variable and the per capita household income as quantitative variable. Table 7-6 shows the number of iterations, the log-likelihood values and the goodness-of-fit indices. All $GFI_0(K)$ values are smaller than 6%. Only a small amount of the basis deviance is explained by the cluster solutions. This indicates that an underlying class structure with very distinct and well separated social classes does not exist.

K	number of iterations	log(L(K))	GFI0(K) in %	GFI1(K) in %
1	3	-26798.661	-	-
2	45	-26005.419	2.960	2.960
3	243	-25905.318	3.334	0.385
4	91	-25576.964	4.559	1.268
5	237	-25549.555	4.661	0.107
6	289	-25497.352	4.856	0.204
7	361	-25486.569	4.896	0.042
8	415	-25471.113	4.954	0.061
9	405	-25459.052	4.999	0.047
10	501	-25441.478	5.064	0.069
11	501	-25433.492	5.094	0.031
12	492	-25332.899	5.470	0.396

Table 7-6: Number of iterations, log-likelihood functions and goodness of-fit measures for an empirical data set

For the two-cluster and four cluster solutions, the $GFI_1(K)$ index is greater than 1%. These two solutions also require much fewer iterations than the other models (except, of course, the one-cluster solution). Hence, a two-class or four-class structure may be used for further analysis. These solutions are shown in table 7-7. The two-cluster solution shows two classes, a lower class and an upper class. The lower class has an average per capita income of about 7700 ATS, lower education, and mainly recruits from skilled workers, unskilled worker and white-collar workers in a middle position. The upper class has a higher per capita income and a higher education level. White-collar workers in an high or middle positions and professionals build this class. Clusters 2 to 4 of the four-cluster solution can be labeled as "middle class", "lower class" and "upper class". Cluster 1 shows an inconsistency. This cluster is characterized by a middle level of education, but a very low per capita income. It recruits from quite different occupational levels.

Finally the LR test statistics does not lead to a definitive decision on the number of clusters. It suggests, that 12 or more clusters may underlie the data set because the 12-cluster solution is significant to all previous solutions with fewer clusters.

	two-cluster		+ four-cluster solution			
k	1	2	+ 1	2	3	4
p(k/g)	0.64	0.36	+ 0.13	0.29	0.30	0.28
occupation of parents			+			
white-collar worker			+			
in a high position	0.05	0.42	+ 0.16	0.11	0.02	0.46
skilled worker	0.31	0.03	+ 0.14	0.32	0.30	0.02
professionals	0.06	0.19	+ 0.12	0.07	0.07	0.20
unskilled worker	0.18	0.01	+ 0.10	0.10	0.25	0.01
white-collar worker			+			
in a middle position	0.21	0.31	+ 0.20	0.35	0.14	0.29
farmer	0.10	0.00	+ 0.20	0.00	0.12	0.00
white-collar worker			+			
in a low position	0.08	0.02	+ 0.08	0.05	0.09	0.02
educational level of parents			+			
p(2/k)	0.31	0.80	+ 0.40	0.41	0.28	0.85
per capita household income			+			
mu(3/k)	7654.36	13671.84	+ 1245.63	11776.63	7102.03	15027.41
sigma(3/k)	4346.61	7351.55	+ 282.85	3854.00	2088.46	7060.39

Table 7-7: Estimated parameters for the two-cluster and the four-cluster solution of the example of table 7-6

Summary and Discussion

This paper has proposed a general probabilistic model. The model can be used for variables with mixed types and contains the well known models of latent class analysis as submodels. It helps to overcome some problems of k-means clustering. The variables can be of mixed type, the underlying structure can overlap, and the model selection is based on a statistical basis, the maximum likelihood method.

However, the determination of the number of clusters still remains a problem if real data are analyzed. The proposed goodness-of-fit measures and the LR test strategy give some guide to selecting appropriate cluster solutions, but in most cases more than one solution will be appropriate. We recommend using all these solutions in further analyses and selecting the most appropriate one according to the results of these analyses (Bacher 1996: 17-24).

7.2 Computer Programmes

The model described above is available in ALMO (Bacher 1999).

A more sophisticated model is included in LatentGold (Vermunt and Magidson 2000). For nominal and ordinal variables a multinomial distribution is assumed with further restrictions on ordinal variables. For continuous (quantitative) variables a multivariate normal distribution is assumed. LatentGold contains three models:

1. LC cluster models for mixed variables: An unknown nominal (= latent) variable (the latent classes) describes the variation in a set of p dependent variables with mixed measurement levels.
2. LC factor models for mixed variables: One or more dichotomous (or ordinal) latent variable(s) describes the variation in a set of p dependent variables with mixed measurement levels.
3. LC regression models for mixed variables: One or more regressor(s) describes the latent classes and the indicators used to estimate the latent classes.

LatentGold computes different test statistics that allow you to select the best model.

References

Bacher, J., 1996: Clusteranalyse. [Cluster analysis, in German]. München.

Bacher, J., 1999: P36 und P37. Clusteranalyse. ALMO STATISTIK SYSTEM. Linz.

Bacher, J., 2000: A Probabilistic Clustering Model for Variables of Mixed Type. Quality and Quantity, Vol. 34, 223-235.

Bock, H. H., 1989: Probabilistic Aspects in Cluster Analysis. In O. Optiz (Ed.), Conceptual and Numerical Analysis of Data. Berlin-Heidelberg-New York, 12-44.

Bryant, P. G., 1991: Large-Sample Results for Optimization-Based Clustering Methods. Journal of Classification, 8, 31-44.

Everitt, B., 1980: Cluster Analysis. Second Edition. New York.

Fielding, A., 1977: Latent Structure Models. In: C. A. O'Muircheartaigh & C. Payne (Ed.). Exploring Data Structures. London - New York - Sydney - Toronto, 125-158.

Fox, J., 1982: Selective Aspects of Measuring Resemblance for Taxonomy. In H. C. Hudson (Ed), Classifying Social Data. New Applications of Analytic Methods for Social Science Research. San Francisco-Washington-London, 127-151.

Jahnke, H., 1988: Clusteranalyse als Verfahren der schließenden Statistik. [Cluster Analysis as a Method of Inference Statistics, in German], Göttingen,1988.

Jain, A. K., Dubes, R. C., 1988: Algorithms for Clustering Data. Englewood Cliffs (New Jersey).

Kendall, M., 1980: Multivariate Analysis. 2nd Edition.

Pollard, D., 1981: Strong Consistency of K-Means Clustering. Annals of Statistics, 1981, 9, 135-140.

Pollard, D., 1982: A Central Limit Theorem for K-Means Clustering. Annals of Probability, 1982, 10, 919-926.

Rost, J., 1985: A Latent Class Model for Rating Data. Psychometrika, 1985, 50, 37-39.

Vermunt, J. K., Magidson, J., 2000: Latent Gold. User's Guide. Belmont.