Additional Information

Chapter 4.1 and 4.3

Ward´s method (incremental sum of squares) (pp. 48, 68)

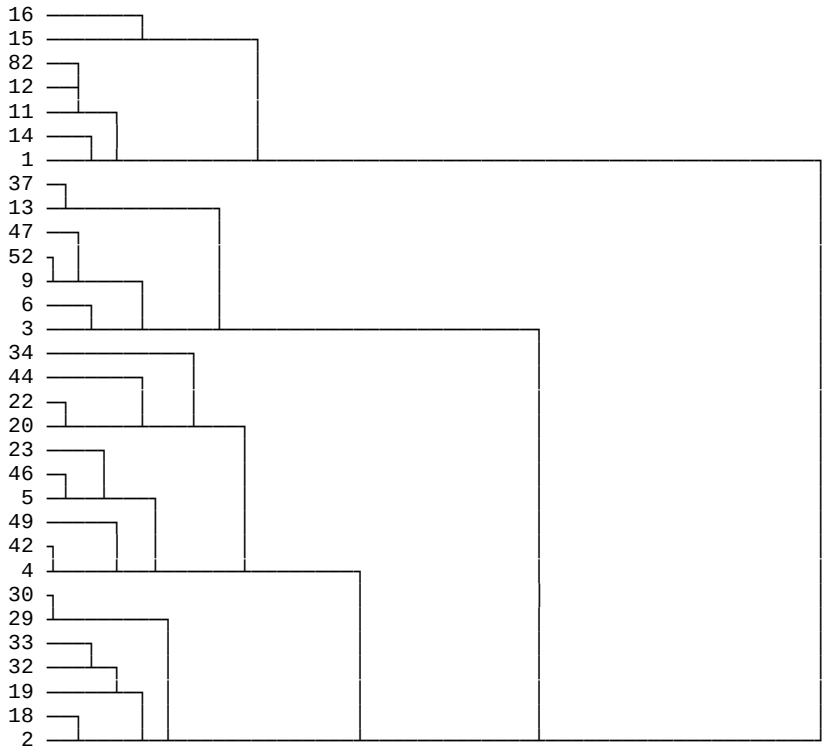
step k: Within sum of squares WSS(k)

Select those two clusters that result in a minimal increase DeltaWSS(k)

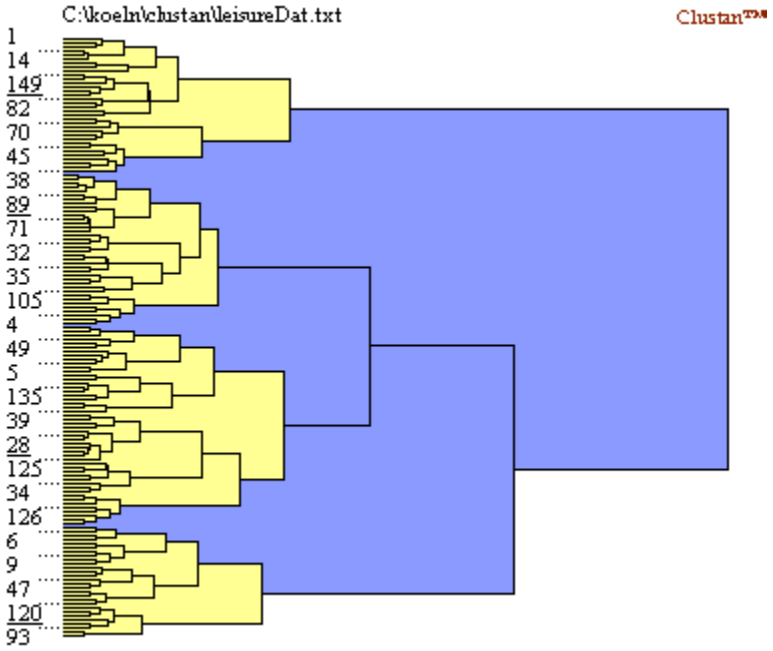
step k+1: Within sum of squares WSS(k+1)

Truncate the dendrogram (ALMO) or using the graphical module of CLUSTAN (p. 76)

Minimum= 8.3, Maximum= 98.6



CLUSTAN



Note:

ALMO and CLUSTAN compute identical results. The results differ from SPSS. Reason: SPSS treats ties in an unpredictable way!

Chapter 4.4

Rand index for the example

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

Results of BAVERAGE by COMPLETE

BAVERAGE	COMPLETE 1	2	3
1	2	0	0
2	0	2	
3	0	2	2

Rand index = 0.714; adjusted Rand index = 0.444 (using the syntax in the appendix)

or ALMO!

```
************************************************************************
Rand-Indizes zur Stabilitaetspruefung der Clusterloesungen

          Phi-Koeffizient       Weighted-Average
            Phi-Koeffizient       Complete-Linkage
                                                  Cluster=  4   1.000
                                                  Cluster=  3   0.714
------------------------------------------------------------------------
          Phi-Koeffizient       Weighted-Average
            Phi-Koeffizient         Single-Linkage
                                                  Cluster=  4   1.000
                                                  Cluster=  3   0.714
------------------------------------------------------------------------
          Phi-Koeffizient       Complete-Linkage
            Phi-Koeffizient         Single-Linkage
                                                  Cluster=  4   1.000
                                                  Cluster=  3   0.714
------------------------------------------------------------------------


Aggregierte Randindizes fuer (Un)Aehnlichkeitsmasse:
----------------------------------------------------

                            (Un)Aehnlichkeitsmass   Rand-Index
                                  Phi-Koeffizient        0.857


Aggregierte Randindizes fuer Modelle:
-------------------------------------

                                           Modell   Rand-Index
                                 Weighted-Average        0.857
                                 Complete-Linkage        0.857
                                   Single-Linkage        0.857

Aggregierte Randindizes fuer Clusterzahl:
-----------------------------------------

                                       Clusterzahl   Rand-Index
                                  Clusterzahl=  4        1.000
                                  Clusterzahl=  3        0.714
```

Additional Information – 3rd day

Missing values (p. 97)

	hierarchical methods	k-means
SPSS	CLUSTER listwise	QUICK CLUSTER listwise and pairwise
CLUSTAN	pairwise	pairwise
ALMO	P=36; pairwise and listwise	P=37; pairwise and listwise

for more details see p. 35ff

Treatment of Ties (p. 51)

SPSS description: last pair, programme: ????

CLUSTAN and ALMO: either the last or the first pair

Example (see syntax vgl.sps):

	ALMO+CLUSTAN		SPSS	
NCLUS	ACLUS1	ACLUS2	SCLUS1	SCLUS2
142,00	63,00	117,00	28,00	117,00
140,00	28,00	114,00	10,00	100,00
139,00	10,00	100,00	37,00	93,00
138,00	37,00	93,00	34,00	78,00
137,00	34,00	78,00	32,00	75,00
136,00	32,00	75,00	11,00	54,00
135,00	11,00	54,00	38,00	91,00
134,00	38,00	91,00	38,00	89,00
133,00	118,00	127,00	72,00	127,00
etc.				

DLEVEL - Difference of Agglomeration Levels (p. 71f)

The computation of DLEVEL is only useful for WARD's method in SPSS, because SPSS doesn't print the increase, instead it prints the within sum of squares.

Technique to determine the number of cluster	**WARD**	**all other methods**
"sharp" increase in the dendrogram	compute DLEVEL	use the original LEVEL
scree diagram	plot DLEVEL	plot LEVEL
mojena I	standardize DLEVEL	standardize LEVEL

Mojena I in SPSS, CLUSTAN and ALMO

SPSS:

stage/step stand levels = t1 = Mojena I

.

k

k+1 stand. level --> above the threshold --> number of cluster at stage k

.

CLUSTAN and ALMO:

Cluster t1 = Mojena I

.

k

k+1 t1 --> above the threshold --> number of cluster = k+1

Therefore:

stand. level at step k+1 = t1 test statistic for number of clusters at stage k !!!!

Example:

```
data list free/stage cluster1 cluster2 coeff stage1 stage2 next.
begin data.
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
end data.

compute ncluster=5-stage.
desc var=coeff/save.

list var=ncluster zcoeff.

NCLUSTER       ZCOEFF
    4,00      -,85553
    3,00      -,47529
    2,00      -,09506
    1,00      1,42588
```

CLUSTAN:

Best Cut Significance Test - Upper tail

Proposed Partition	Realised Deviate	t-Statistic
2 clusters	1,43	2,85

Note:

This "defect" is corrected in mojena1.sps and mojena2.sps and in Errata3.

You can either correct this "defect" or use the method learnt yesterday:

1. copy the schedule to the clipboard
2. paste the schedule into the syntax LEVEL1.sps
3. delete the COMPUTE lines for PLEVEL and DLEVEL, if you are not using WARD
4. Correct the number of cases
5. standardize the variable LEVEL, standardized levels = t1
6. switch to the data matrix
7. read the column ZLEVEL (other methods) resp. DZLEVEL (WARD)
8. search an "significant" increases

Application (p. 5f)

hierarchical techniques	compute a tree representation
	compute a classification

sample size small (SPSS), moderate (ALMO), large (CLUSTAN)

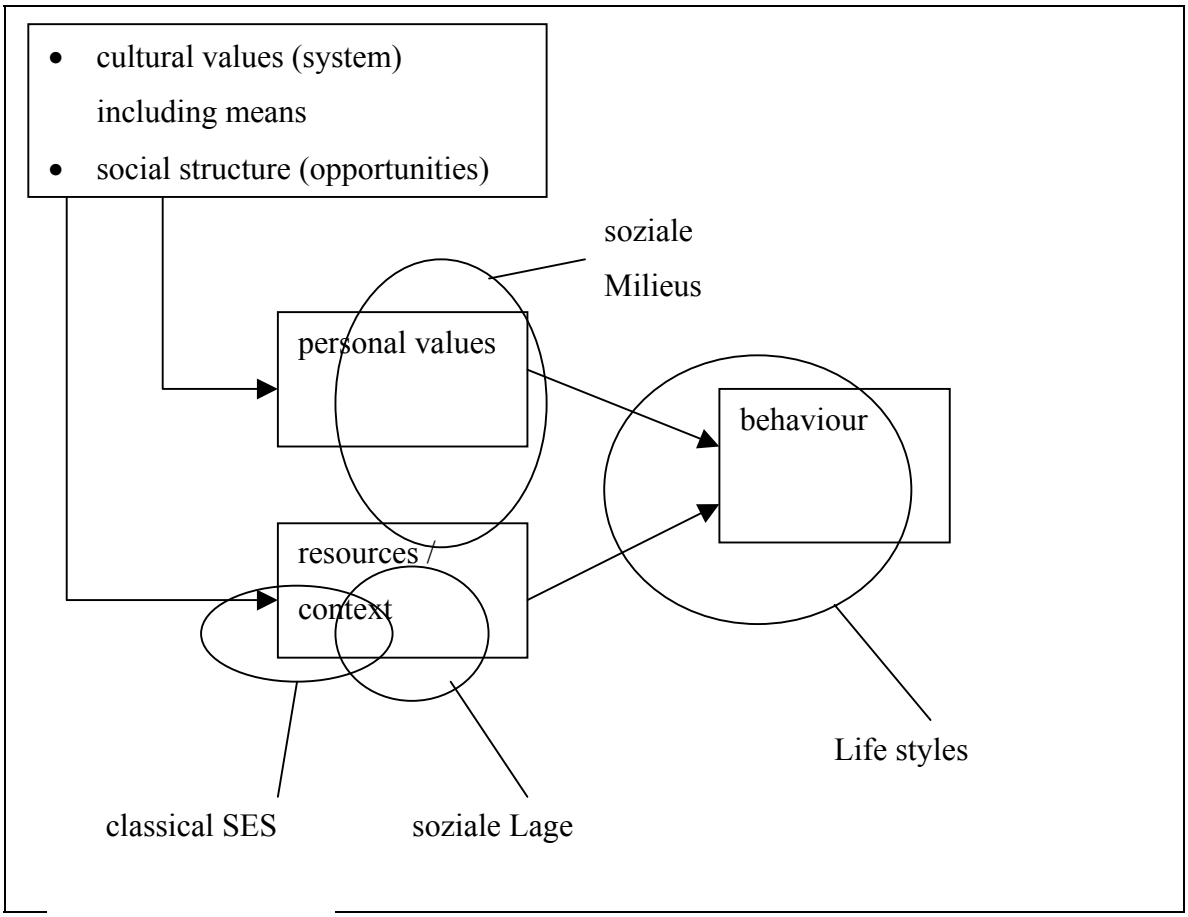
Typical application:
large data set (n=600, 1000, 40000) --> cluster cases --> interest: detect types

e.g.:
life style research: many cases and many variables

variables = behaviour or "near" behaviour = AOI – variables (Blackwell et al. 2001: Consumer Behavior. Orlando, p. 219ff)

Sociology (Social Stratification/Social Inequality): "Soziale Milieus" are used. They are very frequently discussed in the framework of life styles. Most prominent: SINUS (milieus), developed by Flaig (?) and Ueltzhöffer, used e.g. by Vester (social stratification and political orientations) and Heitmeyer (social inequality and right extremism). See p. 16f.

Differences?



Strategies:

- variables: many variables no problem, but nonetheless a reduction is useful (p. 27f).
- cases: see below

Strategies for cases	**disadvantage**	**advantage**
aggregate	homogeneity assumed within cluster	influence of errors is reduced
sample	small clusters may not be present	use everything you have learnt until now (determination of the number of clusters, number of clusters may be underestimated
other software	to become familiar with the software	all cases can be clustered

However: hierarchical solution (e.g. WARD) is not necessarily an optimal partition.

Assumption: No access to other software!

no idea about the number of clusters

draw a sample --> CLUSTER --> determine the number of cluster (=k1, k2, ..)

--> QUICK CLUSTER with k1, k1+1 and k1+2 (and k2, k2+1, k2+2)

--> interpretation ---> stability ---> validation

some idea about the number of clusters (e.g. by qualitative research, theoretical consideration etc.)

QUICK CLUSTER with k --> interpretation ---> stability ---> validation

analyse other k's

Additional Information – 4th day

Thresholds for Missing Values in Pairwise?

QUICK CLUSTER	1 valid value per case is sufficient
CLUSTAN	1 valid value per case is sufficient
ALMO	threshold can be defined

Example (see threshold.sps)

```
data list free/x1 to x10.
begin data.
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 0
1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 0 0 0
1 1 1 0 0 0 0 0 0 0
2 2 2 2 2 2 2 2 2 2
end data.
recode x1 to x10 (0=sysmis).
execute.

quick cluster x1 to x2
  /missing = pairwise
  /criteria = cluster (2)
  /print cluster initial.
```

SPSS			CLUSTAN		
Case Number	**Cluster**	**Distance**	**Case**	**Cluster**	**Nearest**
1	1	,000	1	1	0,167
2	1	,000	2	1	0,130
3	1	,000	3	1	0,033
4	,	,	4	1	Missing
5	1	,000	5	1	0,042
6	1	,000	6	1	0,056
7	2	,000	7	1	0,688

ALMO (KW_SCHWELLE = 0.70;) (see threshold.alm)

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

 Objekt   Clustererzu-   quadrierte Distanz    standardisierte.
          gehoerigkeit   zum Clusterzentrum    mittlere Entfern.
-------------------------------------------------------------------
      1           1              0.00            0.00
      2           1              0.00            0.00
      3          -1             -1.00           -1.00
      4          -1             -1.00           -1.00
      5          -1             -1.00           -1.00
      6           1              0.00            0.00
      7           2              0.00            0.00
```

Weighting variables in CLUSTAN?

Problem: You want to eliminate one variable from the analysis.

Strategy:

- Delete variable in SPSS and transform the data again
- Give variable a weight of 0 in CLUSTAN

Example

Data Matrix

1 1 1 1

1 1 1 5

2 2 2 5

Squared Euclidean Distances with X4

Case	1	2	3
1	0,0000		
2	4,0000	0,0000	
3	4,7500	0,7500	0,0000

Squared Euclidean Distances with X4

Case	1	2	3
1	0,0000		
2	0,0000	0,0000	
3	1,0000	1,0000	0,0000

Standardized Values or t-value?

Background: Clustan reports the stand. value (=deviate) and a t value for Mojena I.

"(...) Select group level corresponding to the first stage j, j=1, ..., N-2 satisfying

$$\alpha_{j+1} > \overline{\alpha} + k \cdot s_{\alpha},$$

where α_{j+1} represents the value of the criterion in stage j+1; k is the standard deviate; $\overline{\alpha}$ and s_{α} are, respectively, the mean and unbiased standard deviation of the α distribution.
(...)
If no value for α satisfies the above inequality, then the decision maker must choose: ... (b= the stage j for which the stage j+1 yields to largest standard deviate (...)
(...)
5. In terms of the predicted number of clusters, values of k in the range 2.75 to 3.50 gave the best results. (...)".

Mojena (1977: 359-361)

Consequences:

(1.) The decision rule is:

$$\frac{\alpha_{j+1} - \overline{\alpha}}{s_{\alpha}} > k$$

k should be in the range between 2.75 to 3.50. The expression on the left side is the standardized level. Therefore, the standardized level or the deviate (=CLUSTAN) should be larger than 2.75 to 3.50.

(2.) The recomputation of test statistic for the truncated tree is problematic. Truncate the dendrogramme after you have analysed Mojena I.

Validation of Clusters

aspect	criteria	test statistic in k-means (a)
intern	high homogeneity	variance within WSS or WSS/TSS, WSS(k) and/or some derived measure based on WSS or WSS(k)
	good separation	variance between BSS or BSS/TSS, distances between clusters
	good model fit	explained variance
relative	better than the null model	explained variance, Beales' F value, (interpretability)
	better than other models	PRE, F-Max, Beales' F value, (interpretability)
external	confirmation of hypothesis that include cluster membership as one variable and a further variable, not used to classify the cases	crosstabulation, regression, anova, logistic regression, ...

(a) You can also use some test statistics of hierarchical methods

Example:

H5: Not integrated juveniles have an higher proportion of males. They more likely to be immigrants.

Crosstab

			Cluster Number of Case				
			not integrated	alienated j.	trad. left	trad right	Total
gender	male	Count	62	44	80	106	292
		% within Cluster Number of Case	72,9%	33,3%	36,0%	62,0%	47,9%
	female	Count	23	88	142	65	318
		% within Cluster Number of Case	27,1%	66,7%	64,0%	38,0%	52,1%
Total		Count	85	132	222	171	610
		% within Cluster Number of Case	100,0%	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	58,705[a]	3	,000
Likelihood Ratio	59,899	3	,000
Linear-by-Linear Association	,003	1	,954
N of Valid Cases	610		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 40,69.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,310	,000
	Cramer's V	,310	,000
N of Valid Cases		610	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Crosstab

			Cluster Number of Case				
			not integrated	alienated j.	trad. left	trad right	Total
national	german	Count	66	113	192	158	529
		% within Cluster Number of Case	78,6%	86,3%	86,9%	93,5%	87,4%
	other	Count	18	18	29	11	76
		% within Cluster Number of Case	21,4%	13,7%	13,1%	6,5%	12,6%
Total		Count	84	131	221	169	605
		% within Cluster Number of Case	100,0%	100,0%	100,0%	100,0%	100,0%

Explained Variance in CLUSTAN

Example

X1	X2	SEUCLID(X,C)/m
2	3	1 / 2 = 0.5
2	1	1 / 2 = 0.5
-2	-4	(1 + 1) / 2 = 1.0
-4	–2	(1 +1) / 2 = 1.0

	X1	X2
C1	2	2
C2	-3	-3

WSS(mean) = 3

TSS = (2 – (-0.5))**2 + ... + (3- (-0.5))**2 = 56

TSS(mean) = 28

explained variance = 1 – WSS/TSS = 1 – WSS(mean)/TSS(mean) = 1 – 3 /28 = 0.893.

CLUSTAN-Results

Cluster	Size	Distance	ESS
1	2	6,250	5,000
2	2	6,250	8,243
Model	4	0,000	2,000

k-Means Case Results for Euclidean Sum of Squares

Case	Cluster	Nearest
1	1	1,000
2	1	1,000
3	2	2,000
4	2	2,000

k-Means Case Results for Euclidean Distance

Cluster	Size	Distance	ESS
1	2	6,250	6,000
2	2	6,250	9,657

Outliers 0

Model	4	0,000	2,414

Case	Cluster	Nearest
1	1	0,500
2	1	0,500
3	2	0,707
4	2	0,707

Homogeneity index in ALMO

k-Means Case Results for Euclidean Distance

```
fprintf(dateikanal,"\n\nStatistiken der Cluster:\n\n");

fprintf(dateikanal,"\n%s\n%s\n%s\n",
"Cluster   n=     Streuung   Homogenitaet   ",
"                innerhalb      innerhalb   ",
"-----------------------------------------");

tt=0e0;
for (i=1;i<=ian;i++) tt+=cmean[i][nvar+2];

for (i=1;i<=ian;i++) begin
   iout(dateikanal,i,7,3,1);    /* Cluster.i*/
   iout(dateikanal,(int16)cmean[i][0],8,2,0); /* Fallzahl */
   ssi=0e0;
   if (cmean[i][nvar+2] > ex_eps) ssi=cmean[i][nvar+1]/cmean[i][nvar+2];
   fout(dateikanal,ssi,10,3,0);       /* Str.innerh */
   if (cmean[i][nvar+2] > ex_eps && sst > ex_eps)   /* Homo.index */
      index=1.0-ssi/sst;
   else index=h_kw;
   fout(dateikanal,index,10,3,4);
   fprintf(dateikanal,"\n");
endfor
```

Additional Information – 5th day

Minkowski Metric

dissimilarity measures

- distance measures (triangular inequality = metric)
  - Minkowski Metric
    - Euclidean
    - City
    - others
  - others
- derived measures from distances (can be metric, too)
- derived from correlation measures (e.g. d=(1-r)**(1/2) = metric)
- others (esp. for binary data)

similarity measures

- correlation measures
  - Pearson r
  - others
- derived from correlation coefficient
- derived from distance measures
- others (esp. for binary data)

Example:

1 1 1 1 1	A
1 1 1 1 5	B
2 2 2 2 5	C

SEUCLID(A,B) = 16	SEUCLID(B,C) = 4
EUCLID(A,B) = 4	EUCLID(B,C) = 2
CITY(A,B) = 4	CITY(B,C) = 4
MINK(A,B;p=1/2;q=1) = 2	MIN(B,C; p=1/2;q=1) = 4

Minkowski Metric p=q implemented in CLUSTER and ALMO, not in CLUSTAN

Validation Criteria for Hierarchical Techniques

aspect	criteria	test statistic in hierarchical techniques
intern	high homogeneity	d-within g=(d-between/d-within), script p. 87 ff
	good separation	"Average" of dissimilarities between cluster (d-between)
	good model fit	Gamma or/and cophenetic corr WARD: explained variance
relative	better than the null model (solution with 1 cluster)	WARD: explained variance **(1-l(k)/l(n-1)),** F-max, bootstrapping
	better than other models	PRE (1-lev(k)/lev(k+1)),
external	confirmation of hypothesis that include cluster membership as one variable and a further variable, not used to classify the cases	Crosstab, chi, regression, t-test, analysis of variance, discrimination analysis

Do not use external variables (passive variables, criteria variables) to cluster cases!

Validation in ALMO

can be done within the cluster analysis module. You can distinguish between

active variables: A_Quantitative_V = ...; A_Ordinale_V = ...; and A_Nominale_V = ...;
passive variables: U_Quantitative_V = ...; U_Ordinale_V = ...; and U_Nominale_V = ...;

--> Example: kmvali.alm

Selection of Appropriate Techniques

--> to find a tree representation

---> hierarchical techniques

tree for variables = correlation coeff. and BAVERAGE, or complete/single

tree for objects = distance coeff. and WARD or BAVERAGE

other coeff. and BAVERAGE

---> to find a classification

---> hierarchical techniques

class. of variables = correlation coeff. and BAVERAGE

class. of objects = distance coeff. and WARD or BAVERAGE

other coeff. and BAVERAGE

---> (large sample size k-means or another Software)

---> to find characteristics of clusters, types, ..

---> small: WARD, (median, centroid)

---> medium and large: K-Means

Minimal Standard for Publication

- You can use SPSS
- method, dissimilarity or similarity measures (if not WARD, median, centroid, K-means), starting values (in the case of k-means)
- ties (hierarchical methods) and stability to ties (report RAND and adjusted RAND)
- influence of the order for k-means, if you used SPSS-Starting values (report RAND and adjusted RAND)
- methods to select the appropriate solution (theoretical reasons, formal reasons, ...)
- explained variance (for k-means and WARD)

You will find additional files on (end of next week)

http://www.wiso-soziologie.uni-erlangen.de/koeln

informations will be in english

bacher@wiso.uni-erlangen.de

knut@wenzig.de